THE FUTURE OF INFORMATION SCIENCES

FF press

6th International Conference
The Future of Information Sciences,
INFuture2017: Integrating ICT in Society
Zagreb, 8-10 November 2017

Organizers
Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
The Miroslav Krleža Institute of Lexicography, Zagreb, Croatia

Under auspices of
Ministry of Science and Education of the Republic of Croatia
Central State Office for the Development of the Digital Society, Government of the Republic of Croatia

Publisher
Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia

Printed by
Web 2 tisak d.o.o.

Impression
150 copies

All papers were reviewed by at least two reviewers. INFuture relies on the double-blind peer review process in which the identity of both reviewers and authors as well as their institutions are respectfully concealed from both parties.

ISSN 1847-8220

THE FUTURE OF INFORMATION SCIENCES

INFUTURE2017

INTEGRATING ICT IN SOCIETY

Edited by

Iana Atanassova, Wajdi Zaghouani, Bruno Kragić,
Kuldar Aas, Hrvoje Stančić, Sanja Seljan

Zagreb, November 2017

CONTENTS

Preface

This is the sixth publication in the series of biennial international conferences, *The Future of Information Sciences (INFuture)* organised by the Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb. This year it is co-organised with the Miroslav Krleža Institute of Lexicography. The title of the conference is *INFuture2017: Integrating ICT in Society*. The conference explores the influence the information and communication sciences have on the society as a whole.

The *INFuture2017* conference consists of 32 papers from 61 authors from nine countries – Bosnia and Herzegovina, Croatia, Estonia, France, Ireland, Slovakia, Sweden, Switzerland, and Qatar. Through eleven chapters this publication follows the topics of the conference.

Keynote papers address the issues of the future of digitised society, language technologies, social media, information retrieval, semantic annotation as well as the future of encyclopaedia in the digital era. The chapter on *Human-Computer Interaction and Language Technologies* elaborate the issues of machine translation, localisation in the video games and a comparative analysis of the storage devices appearing in the fantasy movies with the existing ones. The chapter on *e-Science* analyses topics of education of digital linguists in the context of EU business and life-cycle management of research data in Switzerland. The following chapter *Digitisation, Records Management and Digital Preservation* starts with a proposal of the blockchain-based solution for long-term preservation of the validity of digital signatures and proceeds to focus on the preservation of spatial information, digital archival information system, and digital records in the context of e-services in Sweden. *Personal Digital Information Management* is the chapter with only one paper on the creation of contemporary personal archives which clearly shows the novelty of the topic. The following chapter *e-Encyclopaedia* is the largest. Firstly it focuses on the epistemological value of contemporary encyclopaedic projects, encyclopaedic knowledge as a semantic resource, Slovak e-Belinea project, online encyclopaedias for children and young adults, and the development of thematic portal Znameniti.hr and then switches the focus to the online dictionaries for sign and Asian languages, and a newly developed web dictionary Mrežnik. This is followed by the *Her.IT.age* chapter where the service of historical and cultural heritage and tourism promotion is presented followed by the investigation of the cultural heritage memory economy. The chapter *Governmental and Business Sector Informatics* concentrates on the internet security threats and cyber-attacks to critical infrastructure. The topic on e-health was firstly initiated during the INFuture2015 conference. This year the chapter *e-Health Applications and Solutions* investigates how the digital technology can be used to gain more information about the painful syn-

dromes in the lower back. The chapter *Community Informatics and Service-learning* presents the project developed during the Europe engage project student tour and investigates the influence of ad blockers on the online advertising industry. The last chapter, *Integration of ICT in Education*, explores the teachers' experiences with ICT in education, creation of game-based learning material and finally shows a cryptography-based learning application.

We believe that the results of the research presented in this publication will help you better understand the overarching influence of the ICT to the society. We also hope the research presented here will help you expend your own research, motivate you to cooperate and internationalise your activities, formulate new projects, and achieve high quality results.

See you INFuture!

Editors

KEYNOTE PAPERS

e-Estonia
What will the future of a digitised society be?

Kuldar Aas
Digital Archives of the National Archives of Estonia
Nooruse 3, Tartu, Estonia
kuldar.aas@ra.ee

Summary

When Estonia started building its information society at the end of the 1990s, there was no digital data being collected about our citizens. The general population did not have the internet, or even devices with which to use it. Two decades later, the country has one of the most advanced e-governments in the world, where more than 90% of public information is managed in digital form and most public services, ranging from pet registration, to medical prescriptions, and property ownership, are delivered in a "digital by default" manner. However, with such widespread digitisation of a society comes the responsibility to ensure that the vast amount of digital information being gathered on a daily basis continues to be available for decades to come in a secure, trusted and sustainable manner.
This paper provides an overview of the origins and core aspects of the e-government approach in Estonia, and the effect it has had on the lives of ordinary people. Based on this introduction the paper discusses future challenges in regard to the long-term availability, trust and security of an extensive e-government.

Key words: e-government, long-term availability, Estonia, no-legacy, once-only

Introduction

Named 'the most advanced digital society in the world' by Wired[1], Estonia benefits from an advanced e-government ecosystem which is efficient, saves time and money and provides its citizens with hundreds of seamless and easy to use electronic services covering most of their needs, ranging from registering their pets to dispensing medical prescriptions.
The success story began in the late 1990s. Estonia had only recently gained its independence from the Soviet Union, and the young country's economy was barely able to manage itself. Furthermore, Estonia is one of the least densely populated countries in Europe with just about 1.35 million people living in

[1] http://www.wired.co.uk/article/estonia-e-resident, last visited 16.10.2017

roughly the same land area as Switzerland or the Netherlands. As the 1990s saw also the emergence of the internet boom, the nation's government made a conscious decision to prioritise the digitisation of its processes. Following these political decisions, the early 2000s saw the creation of Estonia's national interoperability requirements, and the according X-Road infrastructure[2], and the establishment of the national digital identification and e-signature framework. These two core components facilitated the development of a huge ecosystem of interoperable institutional information systems over the next few years. To date there are about 500 information systems connected to the X-Road, collectively offering businesses, citizens and the government more than 5,000 digital services (including around 850 web services for citizens and businesses from the central access portal eesti.ee). In total, more than 99% of public services are currently being offered online, most of which in a 'digital-by-default' manner.
To name just some additional facts, Estonia was the first country in the world to introduce nationwide digital elections (2005), 98% of companies in Estonia are established, 99% of banking transactions are carried out and 95% of tax declarations are filed online.
Of course, such widespread digitisation and interoperability requires that careful attention is paid to digital literacy and security. In parallel to developing the digital infrastructure Estonia has concentrated on widespread education initiatives for all age groups, and has developed some of the most rigorous security mechanisms in its public-sector operations. For example, Estonia was one of the first countries to introduce hash chain based security mechanisms on some of the most sensitive data (2008). In addition, the country is currently implementing a nationwide 'no-legacy' policy, meaning that no IT component in active use should be older than 13 years, helping to ensure that all security mechanisms are always up to date.

e-Government sustainability issues

As mentioned above, much of the extensive digital infrastructure in Estonia was developed about 15–20 years ago. This means that, by implementing the 'no-legacy' policy, most of the core components have gone through two or three migrations into newer generations of technology, and the first issues in regard to long-term sustainability are becoming apparent.
Most significantly, the leading way of thinking in IT development in the early 2000's was the so-called 'technology first' approach. This means that information systems were developed with a focus on the effective implementation of available technological components, rendering questions around information lifecycle (including retention, destruction, long-term accessibility and usability of data) of secondary importance or giving them no thought at all. In practice, this has lead in a number of cases to the development of information systems

[2] https://e-estonia.com/solutions/interoperability-services/x-road/ (last visited 16.10. 2017)

where much of the data was only usable with the help of complex programmed logic, specific to a given technological platform and/or version. This approach was reinforced by the perception of 'storage is cheap', with the effect that it was common to develop systems which allowed for the constant addition and creation of new information, and had almost no reasonable means to export or delete data in a controlled manner.

These assumptions and approaches were certainly reasonable and justified in their day, as they allowed for the quick development of new state-of-the-art services and there were hardly any best-practices or scientific research available to prove the contrary. However, it has become clear that, particularly as regards the embedding and accumulation of data into specific software components, this approach presents significant issues. First, and most obviously, is the difficulty of migration. If data is highly dependent on its underlying software platform, the migration of the data into newer platforms can become extremely difficult and costly, as there is the need to re-programme much of the original logic. It has also become clear that data structures cannot be relied upon to be constant over time but, rather, their semantics and syntax evolve along with the evolution of thinking around service provision, data collection and reuse, and with the fresh opportunities offered by innovative new software platforms. The nett effect is that in the long run, any act of technology refreshment includes the need to take into account that the system includes data originating from different eras in terms of data semantics and structures, resulting in a situation where an ever larger amount or resources must be spent on data migration. In the worst cases, this means that available funding will only cover the data migration exercise and will not stretch to the research and development of more effective service models. Looking decades ahead, we can predict with some certainty that, through numerous 'semantic translations' so much of the original context of some older data will be lost, that the data itself will become virtually unusable.

Road towards a long-term e-Government

Given the grim outlook voiced in the previous section one might question whether it is reasonable to undertake digital transformation at all? The unequivocal answer in the Estonian case is, yes. The main reason for this claim is that the general public has become so accustomed to the availability of seamless services delivered straight to their computers or mobile phones. There is also no denying, that despite the issues raised above, in general the digitisation of public processes has greatly benefitted the transparency and efficiency of the government, with some reports claiming an annual save of 800 working years thanks to the use of digital[3].

However, we cannot ignore the problems which we are facing presently, and we must find appropriate and cost-effective solutions.

[3] https://e-estonia.com/solutions/e-governance/ (last visited 16.10.2017)

One of the most important aspects of the journey towards long-term e-governance is a change in how we think about technology and its evolution. In short, we need to embrace change and see it as natural, rather than something to be resisted. We also need to understand that change is not only something happening right now, but is something that will continue to happen in the future. In terms of e-government longevity, we have to take into account that, even if we have already identified the problems associated with previous technology generations, we need to understand that there are potentially many other issues which we do not yet know about embedded into the platforms we implement today. The implication of this is that anybody who is serious about managing their processes digitally in long-term needs to make sure that service development and delivery is not only managed by IT, but must also include an understanding and discussion around data and service lifecycles. This realisation was 'made official' in Estonia in 2017, when a new government regulation[4] was issued to make agencies responsible for assigning a 'chief information governance officer' (CIGO) next to the more regular data protection officers and CIOs. Furthermore, the regulation describes a set of measures which public authorities can use in order to closely monitor and proactively manage their service portfolio.
Another, more technical, aspect is the change in the way how IT is thinking about system development. A growing number of projects take a 'data-first' approach, meaning that the core focus in system development is data standardisation and interoperability, and that any given technological component is just a temporary solution in order to deliver services, keep information secure and well managed in the current point in time. Though many of the practical details of implementing such an approach remain to be discussed and tested to date, first examples of 'data-first' system implementations seem promising.
Finally, it is worth mentioning that the e-government issues outlined above have also changed how public sector digital archives are perceived and have to be implemented. While archives are, and will continue to be, the guardians of most important aspects of documented cultural heritage there is also a growing importance in digital archives as the knowledge centres for digital longevity issues. For example, the digital archives of the National Archives of Estonia has for the last ten years been actively involved formal groups working with principles for information governance, IT design and interoperability frameworks. Additionally, national archives has the mandate to review any newly proposed system development or update in the public sector in order to make sure that individual information systems are also considering and implementing methods for data retention, destruction and/or export.
In more practical terms digital archives are increasingly taking on the role of not only providing tools and specifications for information of archival value, but

[4] Principles for Managing Services and Governing Information, https://www.riigiteataja.ee/en/eli/507072017004/consolide

also for keeping long-term valuable information within agencies. For example, since 2012 the National Archives of Estonia has the legal mandate to offer agencies a 'digital deposit service', which allows public agencies to transfer their legacy data to the safe long-term storage of the national archives, therefore reducing the amount of data kept in live information systems, and ultimately significantly reducing the cost of technological migrations following the 'no-legacy' principle.

Conclusion

The establishment of a digital ecosystem can provide great benefits to the efficiency and transparency of a state, but at the same time presents also many problems in regard to the cost and complexity of keeping the ecosystem available and evolving through extended periods of time. In the case of Estonia, where extensive digital transformation began in early 2000s, the first of these issues are starting to be recognised.

Though there is still a lot of ground to cover it is worth mentioning that the country is going through a transition where systems and digital services are not only delivered for the 'now' but aspects of information lifecycle management are also taken into account. Based on the Estonian experience, any long-term sustainable e-government has to make sure that overall system design and development takes place in close cooperation between IT and information governance specialists, with significant input from digital archiving specialists.

Language Technologies for Social Media

Wajdi Zaghouani
Carnegie Mellon University Qatar
Education City, Doha, Qatar
wajdiz@cmu.edu

Summary

We are witnessing an increased interest from stakeholders to collect and analyze in real time the large-volume of information from social media streams using all kinds of applications ranging from information extraction tools to social media analytics and decision support systems. Social media text is generally noisy, short and linguistically rich as witnessed by the high-frequency rate of code-switching and colloquial expressions used. In this paper, I present an overview of the language technologies within the context of social media, and will discuss the data collection and annotation of social media content. Afterwards, several text processing tools and techniques used before building social media applications will be presented. Finally, some social media applications and their evaluation benchmarks are explored.

Key words: Language Technologies, Social Media, Text Processing, Corpus Annotation

Introduction

On-line social networking has revolutionized the way we communicate. Recent research on social media has revealed the impact of social media on the lives of millions of people. Language technologies could help process social media data using the most recent techniques and algorithms to reveal insightful information from the multilingual big data available online (Pouliquen et al. 2006; Zaghouani 2014). We present an overview of the Natural language processing within the context of Social media. First, we will discuss the data collection and annotation of social media content. Afterwards, we will explain the main challenges faced during the text pre-processing of social media text. Finally, we will explore some tools and applications related to social media and their evaluation benchmarks.

Social Media Data Collection and Annotation

In order to build natural language processing tools and systems, training data is needed (Jeblee et al. 2014; Zaghouani et al. 2015). Social media popularity is increasing and a large amount of public user-generated content is becoming available for collection. However, the collection and the annotation of social

media textual data need to be carefully considered for each task before starting the collection effort. Moreover, the data should also be annotated in a consistent way (Maamouri et al. 2010; Zaghouani et al. 2014; Zaghouani et al. 2015).
Social media data collection depends on the planned task and its applications. For instance, social media textual data could be collected in multiple forms such as image descriptions, videos, posts and metadata as explained in (Ford and Voegtlin, 2003). Furthermore, social media data is often full of spam that should be detected and removed from the dataset.
In order to collect social media data, there exist application programming interface (API) used to integrate with other applications (Obeid et al. 2013). However, some restrictions are possible, for instance, the Twitter API has a limitation per user, per the number of the Tweets to be collected and per the application. This will lead to a limited number of requests. Those interested in getting a larger volume of data may opt for paid access.
The annotation of social media content is a challenging task and clear guidelines should be provided to the human annotators (Zaghouani et al. 2016d). In general, a minimum of two annotators is needed for a given task and the guidelines should clearly explain what and how to annotate (Zaghouani et al. 2016e). In order to access the quality of the annotation, inter-annotator measures are frequently performed to check the agreement rate between the annotators (Zaghouani and Dukes 2014). In case of disagreements, the issue is resolved by taking the majority vote of the annotators and for this reason, it is advised to have an odd number of annotators (Bouamor et al. 2015; Zaghouani et al. 2012). To improve the agreement score, the annotators are encouraged to discuss any disagreement until they reach an agreement. The inter-annotator agreement is measured using the kappa statistical measure used to compensate the agreement obtained for possible agreements due to chance (Artstein and Poesio, 2008, Carletta, 1996).
The annotation tasks can also be performed in a semi-automatic way through intelligent interfaces between the annotations and the users as in the case of GATE (General Architecture for Text Engineering) and TwitIE, a related social media tool used for corpus annotation (Bontcheva et al., 2013).

Social Media Text Processing Tools and Techniques

Social media text is full of useful information, however, it is usually informal and written in a naturally occurring way such as the abbreviations in SMS phone messages. The occurrence of non-standard words and misspelled text poses a big challenge for natural language processing (Pouliquen et al. 2005). In order to build language technologies applications for social media, the collected text should go through various normalization steps (Zaghouani et al. 2016b). Text normalization is especially needed to reduce the linguistic noise from the data (Diab et al. 2018; Zaghouani and Awad 2016). Furthermore, the normalization will reduce the linguistic ambiguity in a language such as Arabic (Haw-

wari et al. 2013; Draffan et al. 2015; Zaghouani et al. 2016c). During the normalization process, the orthographic errors are identified and later on they could be corrected using a dictionary of correctly spelled terms. The dictionary generally allows the detection of out-of-vocabulary entries and unknown words.
Natural language processing tools are essential in language technologies projects especially those involving data annotation (Maamouri et al. 2012; Obeid et al. 2016). We identified several tools frequently used for social media text processing and tools specifically developed for social media.

- **The Stanford CoreNLP:** this is a suite of Natural Language Processing tools for the English language. It supports tokenization, parsing, part-of-speech tagging, and named entity recognition. The Stanford POS tagger was trained for social media text by Derczynski et al. (2013a)
- **Open NLP:** this suite supports various functions from tokenization to sentence segmentation, part-of-speech tagging, named entity extraction, chunking, parsing, and co-reference resolution. The OpenNLP chunker by Ritter et al. (2011) was trained specifically also for social media text.
- **FreeLing:** a set of tools for a variety of languages including English. It can do the text tokenization, sentence splitting, morphological analysis, phonetic encoding, named entity recognition, POS tagging, parsing and co-reference resolution.
- **NLTK**: this is a well-known suite of text processing libraries written in Python for classification, tokenization, stemming, POS tagging, parsing, and semantic reasoning task.
- **GATE**: another well-known toolkit that includes various language processing components such as parsers, morphological analyzer, Part-of-speech tagging. It also contains information retrieval tools, information extraction components for various languages among others. Gate has been adapted to social media text processing through the TwitIE module (Derczynski et al., 2013b). This module supports the tokenization of Twitter texts and also the POS tagging and the named entities recognition.
- **NLPTools:** this is a Natural Language Processing library dedicated to text classification, tokenizing, stemming and clustering.
- **TweetNLP:** this part-of-speech tagger was developed at Carnegie Mellon University and was built especially for social media texts (Owoputi et al., 2013). It was created with manually labeled POS annotated tweets. A dedicated Web-based annotation tool was used in this project
- **TweeboParser**: A dependency parser was built using the Twitter annotated Treebank for 929 tweets (Kong et al., 2014).
- **The University of Washington (UW)** Twitter NLP Tools: this is a suite of tools created by (Ritter et al., 2011) and includes a POS tagger and an annotated Twitter data.

Since social media messages are available in multiple languages and in some cases, there is a situation of code-switching, for example, some users may write in Arabic and write part of the text in English. In order to detect the language of social media text, several language detection systems were built.
In order to build these language identification tools for social media, existing tools need to be re-trained and the performances are generally lower due to nature of the social media messages. For instance, language identification systems can achieve around 98% of precision in detecting languages while it will decrease to 90% for data for example as explained in Derczynski et al. (2013a).
Lui and Baldwin (2014) tested several language identification systems on Twitter data and obtained an F-score of 89% with the best system. Twitter dataset became the standard testing and training data used for this kind of tasks. Testing is usually done using existing tools after various text normalization and cleaning steps such as removing hashtags, emoticons, mentions, re-tweets etc..
For language identification task, the methods used relied mostly on the text of the message, but in some cases such as in Carter et al. (2013), they used metadata information, a unique approach in social media. They found that several features can help in identifying the language such as the language profile of the user, the hyperlink content, the language profile of the other users mentioned in the given post, the language of the original post and the language profile of the given tag. They tested their method and it improved by 5% over the baseline.
We identified the following language identification tools:

- **LangDetect**: this is a Bayes classifier and it is based on character n-grams without feature selection and a set of normalization heuristics.
- **Whatlang:** this tool is based on a vector-space model with per-feature weighting over character n-grams (Brown, 2013).
- **Langid.py**: this tool is adapted for more than 90 languages and uses a feature set selection process from various sources (Lui and Baldwin, 2012).
- **LDIG**: this is a Java language identification tool done specifically for Twitter messages. It was trained on 47 languages. It uses a document representation based on data structures.

In some cases, social media posts are written in a dialectal variety and dedicated dialectal identification tools are required in this case. We cite the case of the various Arabic varieties used in social media in 22 Arabic countries. We noticed that dialectal Arabic is usually mixed with standard Arabic in social media messages. In recent years, more attention was given to building applications for Arabic dialectal identification (Zaghouani et al. 2016a).
This task attempts to find dialect variety used in a set of texts that use the same character set in a known language and since dialects within the same language are sometimes very similar, this task is more difficult than language identification. the various machine learning techniques and methods used for language

identification were adapted for dialect identification as well. Once a dialect is identified, it will be mapped to standard Arabic for further processing using the MSA tools as there is a lack of dialect dedicated NLP tools.
We located several projects related to dialectal Arabic, we cite in particular the efforts of (Habash, 2010) and Diab et al. (2010) within the context of the CO-LABA project, a large-scale project to create resources and processing tools for Dialectal Arabic blogs. The project focused mostly on four Arabic dialects: Moroccan, Iraqi, Egyptian and Levantine.

Social Media Language Technologies Application

In this section, we present a selection of some social media related applications based on language technologies. These applications based on social analytic could give useful insights on social media user behavior to small businesses, industry, financial institutions among other institutions.

Health applications

In healthcare, many patients tend to write information about their health and possible treatments and the side effects of medication. They also share their experiences with other social media users. all this data can be useful for health care professionals, for example collecting data about depression could be useful in detecting possible mental health issues. Ali et al. (2013) built a collection of texts from on-line medical groups related to hearing devices and sorted them into positive, neutral or negative. This sentiment annotation is useful for example when we are interested in filtering only messages with a specific opinion. When dealing with health-related data, we need to take into consideration the user privacy and a de-identification process should be performed to remove sensitive personal information from this data.

Financial Applications

In the financial domain, social media analysis can be useful for example in studying the relation between the economic indicators and the financial news and the role of rumors in the stock exchange market fluctuations. Moreover, social media can be used to do surveys and studies and we can cite the example of Twitter data the revealed the public mood of a given population for market research. Sul et al. (2014) collected data from Twitter messages related to companies in the S&P 500 and they analyzed the cumulative emotional valence by comparing the average daily stock market income. Their results revealed that the cumulative emotional valence (negative or positive) of tweets about a specific company was related to a given company stock income. In another application, Bollen et al. (2010) did some analysis on the content of Tweets on a daily basis using mood tracking applications. They measured the negative versus positive mood using six dimensions (sure, calm, alert, vital, happy and kind).

Disaster Relief Applications
Social media messages can be used to monitor and detect signs of an emergency situation in a timely manner for stakeholders in crisis management. For example, a sudden change in trending topics in social media can be a sign of a possible emergency and should be tracked such as early indications of fire, earthquake or Flooding. Also, social media can be used a tool to send updated information about the evolution of the crisis. Language technologies can play a vital role here by the automatic analysis and monitoring of such messages which could help the government agency to quickly react. We cite the work of Yin et al. (2012) who monitored monitor Twitter streams to detect emergency situations by creating an automatic system to enhance situation awareness.

Security and Defence Applications
The massive amount of user-generated social media messages could be vital for safety and security, but it is hard for Humans to manually scroll through these messages in order to detect possible security threats. For example, Terrorists may post their messages and could use social media to spread their views. Security-related social media applications can be applied to find these patterns of suspicious behavior and investigate the suspects profiles such as the work of Mohay et al., (2003) who built an intrusion detection application.

Media Monitoring Applications
Monitoring the online media could be a helpful application for business intelligible and also for computational journalism as it helps interested parties to quickly detect important information that is difficult to get in a traditional way. For instance, these tools can quickly track millions of articles and broadcast media and report in a summary the most meaningful information. For example, Nagarajan et al. (2009) extracted from Twitter the observations on spatial temporal-thematic analysis to real-world events. They used Twitris, a Semantic Web application. Another related application is TwitInfo which can track events on Twitter and collect and visualize in a concise way the events according to the user preference.

Evaluation
In order to evaluate the performance of social media tools and applications, several standard benchmarks were created (Chiao et al. 2006; Temnikova et al. 2016; Rozovskaya et al. 2015; Mohit et al. 2014; Atwell et al. 2010). In recent years, we witnessed a surge in social media related evaluation campaigns such as the annual SemEval campaign,[1] the annual CLEF labs and workshops[2] and

[1] http://alt.qcri.org/semeval2017/index.php?id=tasks
[2] http://www.clef-initiative.eu/

the various iteration of TREC campaign.[3] In EMNLP 2014,[4] a shared task was organized for code-switching detection in Twitter messages and a standard data set was distributed. The data included messages between two languages for four pairs: Chinese-English, Nepalese-English, Spanish-English and Modern Standard Arabic and Arabic dialects.

Conclusion

In this paper, we presented a general overview of the social media text and the language technologies. We started by describing the social media text collection and annotation process, a necessary initial step in any project related to text processing. Later on, we described the text pre-processing step and the NLP tools that are generally used to prepare the data for and build a variety of useful social media real-world applications.

Acknowledgement

This paper was made possible by NPRP grant 9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the author.

References

Ali, Tanveer, Marina Sokolova, Diana Inkpen, and David Schramm. Can i hear you? Opinion learning from medical forums. Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013), 2013

Artstein, Ron and Massimo Poesio. Inter-coder agreement for computational linguistics. Computational Linguistics, 34:553–596, 2008.

Atwell, Eric, Kais Dukes, Abdul-Baquee Sharaf, Nizar Habash, Bill Louw, Bayan Abu Shawar, Tony McEnery, Wajdi Zaghouani, Mahmoud El-Haj. 2010.Understanding the Quran: a new Grand Challenge for Computer Science and Artificial Intelligence. In Grand Challenges in Computing Research for 2010 and beyond. part of ACM-BCS Visions of Computer Science conference. 13-16 April 2010, Edinburgh University

Bollen, Jonah, Huina Mao, and Xiao-Jun Zeng. Twitter mood predicts the stock market. Computing Research Repository (CoRR), abs/1010.3003, 2010.

Bontcheva, Kalina, Leon Derczynski, Adam Funk, Mark Greenwood, Diana Maynard, and Niraj Aswani. Twitie: An open-source information extraction pipeline for microblog text. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013, pages 83–90. INCOMA Ltd. Shoumen, BULGARIA, 2013.

Bouamor, Houda, Wajdi Zaghouani, Mona Diab, Ossama Obeid, Kemal Oflazer, Mahmoud Ghoneim and Abdelati Hawwari. A Pilot Study on Arabic Multi-Genre Corpus Diacritization. 2015. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), Beijin, China, July 2015.

Brown, D. Ralf. Selecting and weighting n-grams to identify 1100 languages. In Ivan Habernal and Vaclav Matousek, editors, Text, Speech, and Dialogue, volume 8082 of Lecture Notes in Computer Science, pages 475–483. Springer, 2013

[3] http://trec.nist.gov/

[4] http://emnlp2014.org

Carletta, Jean. Assessing agreement on classification tasks: The kappa statistic. Computatioanl Linguistics, 22(2):249–254, June 1996

Carter, Simon, Wouter Weerkamp, and Manos Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation, 47(1):195–215, March 2013

Chiao, Yun-Chuang, Olivier Kraif, Dominique Laurent, Thi Minh Huyen Nguyen, Nasredine Semmar, François Stuck, Jean Véronis, Wajdi Zaghouani. Evaluation of multilingual text alignment systems: the ARCADE II project. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

Derczynski, Leon, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In Proceedings of the International Conference on Recent Advances in Natural Language Processing, Hissar, Bulgaria, 7-13 September 2013. ACL, 2013a

Derczynski, Leon, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In Proceedings of the 24th ACM Conference on Hypertext and Social Media, pages 21–30, Paris, France, May 2013b. ACM

Diab, Mona, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. Colaba: Arabic dialect annotation and processing. In LREC Workshop on Semitic Language Processing, pages 66–74, 010

Diab, Mona, Aous Mansouri, Martha Palmer, Olga Babko-Malaya,Wajdi Zaghouani, Ann Bies, Mohammed Maamouri. A Pilot Arabic Propbank; LREC 2008, Marrakech, Morocco, May 28-30, 2008.

Habash, Nizar. Introduction to Arabic natural language processing. Synthesis Lectures on Human Language Technologies, 3(1):1–187, 2010

Jeblee, Serena; Houda Bouamor; Wajdi Zaghouani; Kemal Oflazer. CMUQ@QALB-2014: An SMT-based System for Automatic Arabic Error Correction. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, October 2014.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. A dependency parser for tweets. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1001–1012, Doha, Qatar, October 2014. Association for Computational Linguistics.

Lui, Marco and Timothy Baldwin. Accurate language identification of Twitter messages. In Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM), pages 17–25, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

Lui, Marco and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In Proceedings of the ACL 2012 System Demonstrations, pages 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics

Maamouri, Mohamed, Ann Bies, Seth Kulick, Wajdi Zaghouani, Dave Graff and Mike Ciul. From Speech to Trees: Applying Treebank Annotation to Arabic Broadcast News. In Proceedings of LREC 2010, Valetta, Malta, May 17-23, 2010.

Maamouri, Mohammed, Wajdi Zaghouani, Violetta Cavalli-Sforza, Dave Graff and Mike Ciul. Developing ARET: An NLP-based Educational Tool Set for Arabic Reading Enhancement. In Proceedings of The 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT 2012, Montreal, Canada.

Mohay, George, Alison Anderson, Byron Collie, Olivier de Vel, and Rodney McKemmi. Computer and Intrusion Forensics. Artech House, Boston, 2003

Mohit, Behrang, Alla Rozovskaya, Nizar Habash, Wajdi Zaghouani, Ossama Obeid. The First QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, October 2014.

Nagarajan, Meenakshi, Karthik Gomadam, Amit P. Sheth, Ajith Ranabahu, Raghava Mutharaju, and Ashutosh Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In Web Information Systems Engineering - WISE 2009, 10th International Conference, Poznan, Poland, October 5-7, 2009. Proceedings, pages 539–553, 2009

Obeid, Ossama, Houda Bouamor, Zaghouani, Wajdi, Mahmoud Ghoneim, Abdelati Hawwari, Sawsan Alqahtani, Mona Diab, Kemal Oflazer. MANDIAC: A Web-based Annotation System For Manual Arabic Diacritization. In Proceedings of The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media.

Obeid, Ossama, Wajdi Zaghouani, Behrang Mohit, Nizar Habash, Kemal Oflazer and Nadi Tomeh. A Web-based Annotation Framework For Large- Scale Text Correction. In Proceedings of IJCNLP'2013, Nagoya, Japan.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In Proceedings of Human Language Technologies 2013: The Conference of the North American Chapter of the Association for Computational Linguistics, Atlanta, GA, USA, 9-15 June 2013, pages 380–390. ACL, 2013

Peter, Dominey Ford and Thomas Voegtlin. Learning word meaning and grammatical constructions from narrated video events. In Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non Linguistic Data, 2003.

Pouliquen, Bruno, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, FlavioFuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best. Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. Proceedings of the 5th (LREC'2006), pp. 53-58. Genoa, Italy, 24-26 May 2006.

Pouliquen, Bruno, Ralf Steinberger, Camelia Ignat, Irina Temnikova, Anna Widiger, Wajdi Zaghouani & Jan Žižka. Multilingual person name recognition and transliteration. Journal CORELA - Cognition, Représentation, Langage. Numéros spéciaux, Le traitement lexicographique des noms propres. Available online at: http://edel.univ-poitiers.fr/corela/document.php?id=490. ISSN 1638-5748.

Ritter, Alan, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), EMNLP '11, pages 1524–1534, Edinburgh, Scotland, UK., July 2011.

Rozovskaya, Alla, Houda Bouamor, Nizar Habash, Wajdi Zaghouani, Ossama Obeid, Behrang Mohit. The Second QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), Beijing, China, July 2015.

Sul, Keel Hong, Allan R. Dennis, and Lingyao Yuan. Trading on Twitter: the financial information content of emotion in social media. In System Sciences (HICSS), 2014 47th Hawaii International Conference on, pages 806–815, Jan 2014.

Temnikova, Irina, Zaghouani Wajdi, Stephan Vogel, Nizar Habash. 2016. Applying the Cognitive Machine Translation Evaluation Approach to Arabic. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2016).

Yin, Jie, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. Using social media to enhance emergency situation awareness. IEEE Intelligent Systems, 27(6):52–59, 2012.

Zaghouani, Wajdi and Dana Awad. Toward an Arabic Punctuated Corpus: Annotation Guidelines and Evaluation. In Proceedings of The 2nd workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media.

Zaghouani, Wajdi. Critical Survey of the Freely Available Arabic Corpora. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), OSACT Workshop. Rejkavik, Iceland, 26-31 May 2014.

Zaghouani, Wajdi, Nizar Habash, Houda Bouamor, Ossama Obeid, Sawsan Alqahtani, Mona Diab and Kemal Oflazer. Filtering Dialectal Arabic Text in Two Large Scale Annotation

Projects. The 2nd Workshop on Noisy User-generated Text (W-NUT), December 11 2016, Osaka, Japan. 2016a.

Zaghouani, Wajdi, Ahmed Abdelali, Francisco Guzman and Hassan Sajjad. Normalizing Mathematical Expressions to Improve the Translation of Educational Content. In Proceedings of the AMTA 2016 Workshop Semitic Machine Translation (SeMaT) Collocated with EMNLP 2016 Workshops on November 1st, 2016 Austin, Texas, USA. 2016b.

Zaghouani, Wajdi, Abdelati Hawwari, Sawsan Alqahtani, Houda Bouamor, Mahmoud Ghoneim, Mona Diab and Kemal Oflazer. Using Ambiguity Detection to Streamline Linguistic Annotation, In Proceedings of Coling Workshop "Computational Linguistics for Linguistic Complexity" (CL4LC), Osaka Japan. 2016c.

Zaghouani, Wajdi, Houda Bouamor, Abdelati Hawwari, Mona Diab, Ossama Obeid, Mahmoud Ghoneim, Sawsan Alqahtani, Kemal Oflazer. Guidelines and Framework for a Large Scale Arabic Diacritized Corpus. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2016). 2016d.

Zaghouani, Wajdi, Nizar Habash, Ossama Obeid, Behrang Mohit, Houda Bouamor, Kemal Oflazer. Building an Arabic machine translation post-edited corpus: Guidelines and annotation. In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2016). 2016e.

Zaghouani, Wajdi, Nizar Habash, Houda Bouamor, Alla Rozovskaya, Behrang Mohit, Abeer Heider and Kemal Oflazer. Correction Annotation for Non-Native Arabic Texts: Guidelines and Corpus. In Proceedings of the 9th Linguistic Annotation Workshop, co-located with NAACL in Denver, Colorado, USA, 2015.

Zaghouani, Wajdi, Taha Zerrouki and Amar Balla. SAHSOH@QALB-2015 Shared Task: A Rule-Based Correction Method of Common Arabic Native and Non-Native Speakers' Errors. The Second QALB Shared Task on Automatic Text Correction for Arabic. In Proceedings of the ACL 2015 Workshop on Arabic Natural Language Processing (ANLP), Beijing, China, July 2015.

Zaghouani, Wajdi, Behrang Mohit, Nizar Habash, Ossama Obeid, Nadi Tomeh and Kemal Oflazer. Large-scale Arabic Error Annotation: Guidelines and Framework. in the Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), Rejkavik, Iceland, 26-31 May 2014.

Zaghouani, Wajdi and Kais Dukes. Can Crowdsourcing be used for Effective Annotation of Arabic? In Proceedings of the International Conference on Language Resources and Evaluation (LREC'2014), Rejkavik, Iceland, 26-31 May 2014.

Zaghouani, Wajdi, Abdelati Hawwari and Mona Diab. A Pilot PropBank Annotation for Quranic Arabic. In Proceedings of the first workshop on Computational Linguistics for Literature, NAACL-HLT 2012, Montreal, Canada.

Information Retrieval and Semantic Annotation of Scientific Corpora

Iana Atanassova
Centre Tesnière – CRIT, Université de Bourgogne Franche-Comté
30 rue Mégevand 25030 Besançon, France
iana.atanassova@univ-fcomte.fr

Summary

Scientific papers are highly structured texts and display specific properties related to their references but also argumentative and rhetorical structure. Natural Language Processing can be applied to efficiently explore scientific corpora and develop applications for the Semantic Web, Information Retrieval, Automatic Summarization and Bibliometrics. The organization of scientific papers typically follows a standardized pattern, the well-known IMRaD structure (Introduction, Methods, Results, and Discussion). By analysing the full text of about 80,000 papers of the PLOS corpus, we studied this structure from several different perspectives. Firstly, we performed quantitative and qualitative analyses of citations and their positions in the structure of papers. Secondly, we studied the occurrences of verbs in citation contexts and their similarities across the different sections. Finally, using sentence-based similarity metrics, we quantified the phenomenon of text re-use in abstracts with respect to the IMRaD structure. This research allowed us to establish some of the invariants of scientific papers and the results are useful for implementing novel text mining and information retrieval interfaces taking into consideration the argumentative structure of papers. More specifically, they can be considered as an important element when creating linguistic resources and rule-based methods to perform fine-grained semantic analysis of scientific papers.

Key words: Information extraction, Information retrieval, scientific papers, semantic annotation, text mining, IMRaD

Introduction

Nowadays, we can witness the emergence of a new area of study in Natural Language Processing, which is NLP-enhanced Bibliometrics, or studying the properties of scientific papers and their full text content to gain insight into the evaluation of science. At the same time, the semantic processing of scientific papers is at the heart of many other applications, such as Information Retrieval, semantic publishing, information extraction and aggregation of data from scientific corpora. This increased interest in the application of NLP methods to scientific publications is the result of several important factors:

- the ever growing availability of full text scientific corpora, as a consequence of the Open Access movement;
- the emergence of standardized formats for the representation of the full text content of scientific papers (e.g. JATS, DocBook);
- the recent developments in NLP that have resulted in a number of robust and accessible tools for versatile text processing.

In this context, many studies, workshops and evaluation sessions[1] have been initiated in the recent years, aiming to provide new methods dedicated to the processing of scientific corpora. One important point of interest is the cognitive structure of scientific production and in particular the invariants that may exist in scientific writing that can be of linguistic, discourse, structural or distributional nature.

Levels of processing

Scientific papers are subject to numerous conventions, norms and editorial requirements. They are highly structured texts and often follow a specific rhetorical structure. In experimental sciences, the IMRaD structure (Introduction, Methods, Results and Discussion) (Bertin et al., 2013) has emerged as a standardised pattern that was adopted by a majority of journals during the second half of the twentieth century.

A scientific paper can be considered as a structured document of three parts: metadata (including title, author list, publication date, keywords, abstract, etc.), full text body, and bibliography (see figure 1). Bibliometric studies traditionally focus on only the metadata and bibliography. The elements that contain information in natural language are the title, the abstract, the full text body and the bibliography. Implementing efficient information retrieval and information extraction for all these elements of scientific papers is an important step towards making scientific knowledge more accessible and helping scientists cope with the enormous amount of data produced each day. The information extraction and ontology population are intended to enrich the papers' metadata and thus define new facets for information retrieval.

Concerning the elements of the bibliography and their corresponding in-text citations, a new field of investigation has emerged that aims to characterize in-text citations according to their contexts. Many applications can be envisaged, among which the automatic summarisation (Wang and Zhang, 2017; See et al.

[1] E.g. CL-SciSumm Evaluation Task (https://www.aclweb.org/portal/content/3rd-computational-linguistics-scientific-document-summarization-shared-task), Semantic Publishing Challenge (https://2016.eswc-conferences.org/assessing-quality-scientific-output-its-ecosystem), BIRNDL (http://wing.comp.nus.edu.sg/~birndl-sigir2017/), BIR (https://www.gesis.org/en/services/events/events-archive/conferences/ecir-workshops/ecir-workshop-2017/), WOSP (https://wosp.core.ac.uk/jcdl2017/), CLBib (https://easychair.org/cfp/CLBib2017).

2017), the extraction of relations between authors and the creation of author networks, and the creation of new bibliometric indices.

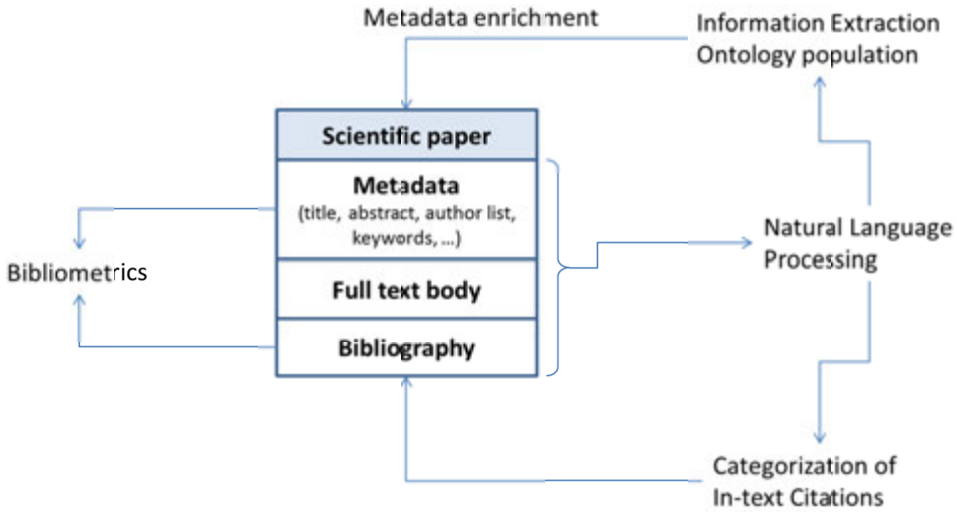


Figure 1. Elements of scientific papers and types of processing

A series of experiments on the PLOS corpus: in search of invariants

We have carried out a series of studies that show some invariants in the density of in-text citations according to their positions in the IMRaD structure (Bertin et al, 2017). The dataset consisted of around 80 000 peer reviewed papers published by PLOS[2] in Open Access. The seven PLOS journals cover different domains and are available in XML format following the JATS[3] schema. The processing stages consist of the analysis of section titles to categorize the sections, the text segmentation into sentences and the identification of in-text citations and their corresponding references. The analysis showed the existence of a strong and stable relationship between the distribution of in-text citations and the rhetorical structure of the papers, and this independently of the journals and domains (Bertin et al, 2017; 2013). This distributional phenomenon is an important factor in the study of citations and their roles in bibliometrics. Moreover, it has direct applications in information retrieval and extraction, as the density of citations in a particular sentence, paragraph or section can be related to its relevance to a user need. In fact, one possible application is the exploitation of these corpora through rich search interfaces (Bertin and Atanassova, 2012;

[2] The Public Library of Science, https://www.plos.org

[3] Journal Article Tag Suite, https://jats.nlm.nih.gov/

Presutti et al., 2014). Another related study addresses the question of the relationships that exist between abstracts, keywords and the full text body of papers (Atanassova et al., 2016). Studying the ways authors summarize their papers is useful to gain insights into which parts of the information in a paper are the most relevant and are considered as the most important by the author.
Other studies were focused on the distributional analysis of citations in IMRaD in relation to the occurrences of different verbs in the citation contexts (Bertin and Atanassova, 2014). While it is natural to find a large number of in-text citations in the introduction and literature review of an article, knowing the verbs and syntactic patterns that introduce these citations is essential to the lexical and semantic analysis of citation contexts.
The production of visualizations of scientific corpora at a large scale is important for the discovery of trends and innovations, and for landscaping a particular domain. Several visualizations of the structure of papers were produced in order to show the various roles that citations take as a function of their position in the rhetorical structure (Bertin et al., 2014). Furthermore, we have explored the visualization of spatial data extracted from corpora in the biomedical domain: we have studied the *PLOS Neglected Tropical Diseases* journal in order to produce geographical maps of the occurrences of spatial data in the corpus related to tropical diseases (Atanassova et al, 2015).

Information retrieval and semantic facets

The Semantic Web assimilates the production of scientific knowledge through ontologies dedicated to the description of scientific papers (CiTO for the characterization of citations, DoCO for the different elements of a document, and BiRO for the description of bibliographic references) (Shotton, 2010). The current publication models allow more and more applications oriented towards the exploitation of scientific corpora at a large scale and the new challenges exist around the automatic aggregation and production of semantic information related to or extracted from publications. The technologies of the Semantic Web play an important role for these tasks, especially to ensure the interoperability between the different formats and systems.
We have implemented two prototypes that use linguistic analyses combined with Semantic Web technologies: an Information retrieval system using semantic facets, and a system for the semantic processing and categorization of in-text citations (in Presutti et al., 2014), that participated in the evaluation track *Semantic Publishing Challenge* 2014 at the *European Semantic Web Conference*. The main objective is to produce data from scientific corpora in order to take into account qualitative information extraction from the papers. The annotation of the set of documents was used to produce data in the form of Linked Open Data in order to identify and characterize the cited papers, auto-citations, multiple citations of a paper, funding organizations, paragraphs and sections containing the state of the art, methods and results used in cited papers, etc. We use

knowledge-based linguistic methods for the annotation, combined with existing tools for Named Entity Recognition and POS-tagging (Chang et al., 2016; Maning, 2011). The results are presented in a semantic search engine based on SPARQL.

Facets in information retrieval are traditionally filters that are available in the user interface and that allow to refine the result list according to predefined categories present in the documents. In the classical model, these categories correspond to classes that exist in the metadata of the documents. We can go further by considering semantic facets that depend not only on metadata but also on the full text content and whose values are obtained by a linguistic analysis that must be carried out during the indexation process. Such a model gives a new possibility for the user to access the semantic content of documents: the search results can be filtered according to semantic categories and rich textual navigation can be supported using a linguistic ontology. A different way of viewing this idea is the enrichment of metadata (Bertin and Atanassova, 2012). For example, the prototype that we have implemented gives the possibility for search and textual navigation in a scientific corpus at the level of the sentence by the choice of categories present in the papers, such as *method, hypothesis, result, conclusion, opinion,* etc. These semantic categories are identified during the indexing by a knowledge-based approach. The interface allows a semantic search, where the relevance of a sentence is a function of both keywords and semantic relations expressed in the sentence.

Conclusion

We have presented an overview of several applications around the exploitation of scientific corpora: a distributional study and the characterizing of in-text citations, a semantic search engine and spatial data visualization. With the growing number of papers published daily scientists need new tool for more efficient access to the information and to be able to rapidly grasp the main ideas of papers. The development of such tools is a recent area of research and has been greatly favored by the Open Access movement. One major challenge before Natural Language Processing is modelling and automating the extraction of argumentative structures for the construction of new textual navigation interfaces for scientific texts.

References

Atanassova, Iana; Bertin, Marc; and Larivière, Vincent. On the Composition of Scientific Abstracts. Journal of Documentation. 72 (2016), 4; pp.636 – 647

Atanassova, Iana; Bertin, Marc and Kauppinen, Tomi. Exploitation de données spatiales provenant d'articles scientifiques pour le suivi des maladies tropicales, Gestion et Analyse des données Spatiales et Temporelles (GAST'2015), 15ème conférence internationale sur l'extraction et la gestion des connaissances (EGC-2015), Luxembourg.

Bertin, Marc ; Atanassova, Iana; Larivière, Vincent and Gingras, Yves. The Invariant Distribution of References in Scientific Papers. Journal of the Association for Information Science and Technology (JASIST). 17 (2017), 1, pp. 164 – 177

Bertin, Marc and Atanassova, Iana. A Study of Lexical Distribution in Citation Contexts through the IMRaD Standard. Bibliometric-enhanced Information Retrieval Workshop at the 36th European Conference on Information Retrieval (ECIR-2014), Amsterdam, Netherlands.

Bertin, Marc ; Atanassova, Iana; Larivière, Vincent and Gingras, Yves. The Linguistic Context of Citations. International exposition: 10th Iteration of the Places & Spaces: Mapping Science Exhibit on "The Future of Science Mapping", 2014

Bertin, Marc ; Atanassova, Iana; Larivière, Vincent and Gingras, Yves. The Distribution of References in Scientific Papers: an Analysis of the IMRaD Structure. 14th International Society of Scientometrics and Informetrics Conference (ISSI-2013), Vienna, Austria.

Bertin, Marc and Atanassova, Iana. Semantic Enrichment of Scientific Publications and Metadata: Citation Analysis Through Contextual and Cognitive Analysis. 18, 7-8. *D-lib Magazine*. 2012

Chang, Angel; Spitkovsky, Valentin I.; Manning, Christopher D. and Agirre, Eneko. A comparison of Named-Entity Disambiguation and Word Sense Disambiguation. International Conference on Language Resources and Evaluation (LREC 2016).

Manning, Christopher D. Part-of-Speech Tagging from 97\ to 100\: Is It Time for Some Linguistics? Conference on Intelligent Text Processing and Computational Linguistics (CICLing). 2011

Presutti, V., Stankovic, M., Cambria, E., Cantador, I., Di Iorio, A., Di Noia, T., Lange, C., Reforgiato Recupero, D., Tordai, A. (Eds.), SemWebEval 2014 at ESWC 2014, *Semantic Web Evaluation Challenge*, Communications in Computer and Information Science (Book 475), Springer, Anissaras, 2014

See, Abigail; Liu, Peter J and Manning, Christopher D. Get To The Point: Summarization with Pointer-Generator Networks. Association of Computational Linguistics (ACL), 2017

Shotton, D. (2010). CiTO, the Citation Typing Ontology. Journal of Biomedical Semantics, 1 (Suppl 1), S6. doi:10.1186/2041-1480-1-S1-S6

Wang, Jie and Zhang, Chengzhi. CitationAS: A Summary Generation Tool Based on Clustering of Retrieved Citation Content. 2nd Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics at ISSI 2017, Wuhan, China.

The Encyclopaedia in the Digital Era: A New Beginning or just the Beginning of the End

Bruno Kragić
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
bruno.kragic@lzmk.hr

During the last 10 to 15 years we have witnessed the transition of encyclopaedias from the printed into web form, most visible through the global spreading of Wikipedia, through her numerous varieties, as the most used encyclopaedia today. However, this opens the question about the change in the encyclopaedic principles and the encyclopaedic concept of knowledge. Besides the obvious change in the modes of use, there is the question regarding the encyclopaedic concept itself: collaborative online encyclopaedia is in itself never-ending and all-encompassing while the traditional, printed one had to define its range, number and dimension of the entries. Thus, while collaborative online encyclopaedia is more propulsive, open or "democratic", the "old" type was closer to the original encyclopaedic concept of the transmission of the knowledge. There are several reasons why online encyclopaedias have replaced printed ones: the financial one – Wikipedia is free – is as obvious as the social one – the reflection of the changes in the information field where crowd-sourcing has in many ways replaced experts. Furthermore, printed, paper encyclopaedias are not as easily and thoroughly searchable as their digital counterparts or, to put it more specifically, they are searchable in a different way. Also, they can be neither easily nor constantly updated. In addition, they are far more limited in size (just to illustrate this, the last printed edition of Encyclopaedia Britannica had about 70,000 entries, while English Wikipedia has more than 5,490,000). But, all this presupposes that the main goal of encyclopaedia is the first-hand information about more or less everything, with prominent place given to the current issues or recent events. Printed encyclopaedia, on the other hand, had another underlying goal, some sort of its grand ambition – to provide an overview of all (relevant) human knowledge and to put it into a coherent, logical order. Finally, in the tradition of the most ambitious such project – that of the Diderot's *Encyclopédie*, it had the ambition not only to order knowledge but to reorder it, to be used as enlightenment's tool. In theory, online encyclopaedia can have similar ambitions, but the aforementioned goal requires a fixed body of articles, needs a clear and stable knowledge frame – to know what is to be put inside and what remains irrelevant. For online users almost everything is relevant and has to be put in encyclopaedia. Collaborative online encyclopaedia as such has no system, thus one can question is it encyclopaedia at all. Although the mapping of the human knowledge, encircling all that has to be put in the encyclopaedia has

always been the mark of utopia, and almost all readers of printed encyclopaedias consulted articles in a more or less similar vein as they consult online encyclopaedias, the encyclopaedia in a printed form, encyclopaedia of the past was a sort of the library distilled down into an organized essence of human knowledge while the encyclopaedia of the present (and future?), the online one, is a sort of warehouse of books and information.

Key words: encyclopaedia, collaborative online encyclopaedia, Wikipedia, knowledge organisation

HUMAN-COMPUTER INTERACTION AND LANGUAGE TECHNOLOGIES

Original scientific paper

A comparative error analysis of English and German MT from and into Croatian

Marija Brkic Bakaric
Department of Informatics, University of Rijeka
Radmile Matejčić 2, 51000 Rijeka
mbrkic@inf.uniri.hr

Nikola Babic
Faculty of Humanities and Social Sciences
Sveučilišna avenija 4, 51000 Rijeka
nbabic@ffri.hr

Luka Dajak
Faculty of Humanities and Social Sciences
Sveučilišna avenija 4, 51000 Rijeka
ldajak@ffri.hr

Maja Manojlovic
Faculty of Humanities and Social Sciences
Sveučilišna avenija 4, 51000 Rijeka
mmanojlovic@ffri.hr

Summary

This paper first gives an insight into the problem of evaluating translations and lists existing approaches. After introducing error taxonomies, a comparative error analysis of Google Translate results is conducted based on the selected taxonomy. The analysis covers translations from English and German into Croatian, and from Croatian into English and German. The intra and inter-annotator agreements, which are rarely accounted for, are reported where applicable. The aim of the paper is better understanding of different notions of error analysis and detecting weak points of such analysis. A preliminary list of guidelines for error analysis is suggested.

Key words: error analysis, machine translation, intra-annotator, inter-annotator agreement, guidelines

Introduction

Quality assessment is one of the most debated topics in translation. There is no single standard for translation quality assessment because quality is context dependent (Secară, 2005). Furthermore, one sentence can be translated in multiple

ways. All this makes machine translation (MT) evaluation inherently subjective. Error analysis is a means to assess translations in qualitative terms. It refers to the identification and classification of individual errors in a translated text. However, like other subjective approaches, it is susceptible to low inter-annotator agreement (Stymne & Ahrenberg, 2012).

Related work, presented in the next section, is divided into research on error analysis in general and research with a focus on Croatian. The research presented in this paper aims at gaining a better understanding of different notions of error analysis and detecting weak points of such analysis. Besides compiling a list of guidelines for error analysis, the research aims at answering what we can conclude about Google Translate (GT) engine for the selected language pairs and what types of errors the system makes most often. Section three gives details on the error analysis conducted on GT engine from English and German into Croatian and vice versa. It is followed by results and discussion. Conclusion summarizes main findings and gives directions for future work.

Related work

Error analysis

Translation error can be defined as a semantic component not shared by source and target texts (Koponen, 2010). This component can be larger than individual words (e.g. compound nouns, names, idioms, etc.). Error analysis gives a qualitative view on the MT system and should be an integral part of MT development (Stymne S., 2011). It can point to strengths and problem areas for a certain machine translation system, which is not possible using automatic evaluation metrics (Stymne & Ahrenberg, 2012). Automatic metrics, as well as some forms of human evaluation, such as fluency and adequacy scoring or system ranking, provide quantitative system evaluation (Stymne S., 2011). However, research community would like to get answers to what kind of errors the system makes most often, whether a particular modification improves some aspect of translation although the overall score is intact, whether one system is superior to another in all aspects of translation or just in some, etc. (Popovic & Burchardt, 2011). Context and extra-linguistic knowledge often subconsciously guide us into correcting certain errors and this alone proves that some errors are less severe than others. The author in (Secară, 2005) emphasizes that in a post-editing scenario precedence should be given to error categories over numerical scores, since the effort put into correcting each error type is as important as the final score. MQM resulted from a thorough investigation of major human and machine translation assessment metrics. It is a general mechanism for declaring specific metrics for general quality assessment and error annotation tasks, since various error taxonomies have been suggested for the task of error analysis (Flanagan, 1994; Font-Llitjós, Carbonell, & Lavie, 2005; Elliott, Hartley, & Atwell, 2004; Vilar, Xu, d'Haro, & Ney, 2006; Farrús, Costa-Jussa, Marino, Poch, Hernández, Henríquez, Fonollosa, 2011; Costa, Ling, Luís, Correia, &

Coheur, 2015). Although error analysis is subjective, Stymne and Ahrenberg (2012) argue it is possible to get a reasonable agreement either when using a simple error taxonomy or when using a more detailed taxonomy and a set of guidelines. The study in (Elliott, Hartley, & Atwell, 2004) emphasizes the importance of assigning weights to different error categories to make them correlate with intuitive human judgements of translation quality. Moreover, the focus of MT evaluation research is gradually shifting towards profiling systems with respect to various error taxonomies (Federico, Negri, Bentivogli, Turchi, & Kessler, 2014). One reason is that parallel data limits system knowledge to the observed positive examples. Another is that majority of automatic metrics provide only a holistic view of system performance. The authors in (Federico, Negri, Bentivogli, Turchi, & Kessler, 2014), therefore, propose a robust statistical framework to analyse the impact of different error types on human perception of quality and on automatic metrics.

Since the task of error analysis is labour-intensive and time-consuming, and requires either professionals or native speakers of a language in question, a lot of effort has been put into automatic error classification (Fishel, Bojar, Zeman, & Berka, 2011; Popovic & Burchardt, 2011). Benefits of coupling automatic and manual error classifications are shown in (Popovic & Arcan, 2016). Furthermore, the authors show that conducting manual error annotation on pre-annotated texts, where reference translations are post-edited translation outputs, can give much better and reliable insights into particular flaws of an automatic error classification tool.

Since different language combinations exhibit different error distributions in the translation output which often relates to the linguistic characteristics of the languages involved and divergences between them (Popovic & Arcan, 2016), the rest of the paper is tied to the research involving Croatian as either source or target language.

Croatian language error analysis

There has been abundant work on error analysis involving Croatian. Tree texts of different types are translated from Croatian into English by GT and errors are analysed on *lexical*, *syntactic*, *morphological*, *semantic* and *punctuation* level in a descriptive manner and corroborated by examples and by *fluency* and *adequacy* judgements (Brkic, Vicic, & Seljan, 2009). Short texts from four different domains and genres are translated from Croatian into English by four translation services, including GT, and evaluated by 48 evaluators on a 1-5 scale according to *fluency* and *adequacy* (Brkic, Seljan, & Matetic, 2011). The evaluation for the opposite direction in (Seljan, Brkic, & Kucis, 2011) included only GT and 50 human evaluations in total. The following four categories are covered: *morphological errors*, *untranslated words*, *lexical errors* (which also comprise semantic errors), and *syntactic errors* (which also comprise punctuation errors). The criterion of *adequacy* is mostly affected by *untranslated words*, while the

criterion of *fluency* is more affected by *lexical* and *syntactic errors*. The research in (Brkic, Seljan, & Matetic, 2011) extends the methodology by including three automatic metrics in the evaluation. GT from Croatian into English is compared to an in-house system and human translations in (Brkic, Basic Mikulic, & Matetic, 2012) by six case-sensitive metrics in the legislative and mixed domains (religion, psychology, computer games). BLEU scores on multiple reference translations and human *fluency* and *adequacy* judgements of English-Croatian GT in the domain of legislation are enriched by error analysis in (Seljan, Brkic, & Vicic, 2012), with a special focus on sentence length. The MT error taxonomy used resembles the one in (Vilar, Xu, d'Haro, & Ney, 2006) when first-level categories are taken into account, with a major difference that *extra words* and *incorrect form* are separated out as categories of the highest level. The criterion of *adequacy* is mostly affected by *semantic* and *lexical errors*, while the criterion of *fluency* is mostly affected by *morphological errors*, but also *missing words*, i.e. *lexical errors*. Similarly, English-Croatian GT is evaluated in legislative and general domains by including three additional automatic metrics and investing the impact of lowercasing, tokenization and punctuation in (Brkic, Seljan, & Vicic, 2013). A new language-pair, i.e. Russian-Croatian, is introduced into the analysis in (Seljan, Tucaković, & Dunđer, 2015) with an additional online translation service, i.e. Yandex.Translate. A comparison with four automatic metrics can be found in (Seljan & Dunđer, 2015b). GT is also evaluated by four automatic metrics for Croatian-English and English-Croatian translations in sociological-philosophical-spiritual domain in (Seljan & Dunđer, 2015a). Better results are obtained for the Croatian-English translation direction.

Experimental setup

Tool and error taxonomy

For the error analysis conducted within this research, the tool BLAST [4] is chosen. The Vilar's taxonomy [7] is used as a starting point with a goal to check its suitability for language directions covered by the study. GT is used as the translation engine. The following abbreviations are introduced and used hereinafter: errs (errors), avg sen len (average sentence length), wo (*word order*), unk (*unknown*), punct (*punctuation*), form (*incorrect form*) E1 (1st annotator), E2 (2nd annotator), de (German), en (English), hr (Croatian).

Test set

There are 24 sentences in our evaluation. The text is constructed for the purpose of this research. It is originally written in Croatian and then manually translated into German and English. The text is in the form of a magazine article which is a reflection on major events in the last year. Croatian original has 513 words, while translations in English and German have 601 and 576 words, respectively. German original has 579 words while its translation in Croatian has 504 words.

English original has 566 words while its translation in Croatian has 478 words. These three texts are then translated by GT into English and German for the Croatian source, and into Croatian for the other two languages. The average translation sentence length is 25 for English, 24 for German, while it ranges from 20 to 21 for Croatian. Croatian as a morphologically rich and pronoun-dropping language has the shortest average sentence length, while English has the longest, due to its poor morphology.

Annotators

The error analysis is performed by four annotators in total who are either native speakers or have a formal university-level education of a language in question finished or nearly finished. Croatian translations are assessed by one final-year student at the graduate study of Croatian who is also native in Croatian and by one native speaker of Croatian. They are given access only to the reference sentences, and not to the source sentences, as we do not want to presume the knowledge of English and/or German. Similar methodology is applied in (Elliott, Hartley, & Atwell, 2004). German translation is assessed by one final-year student at the graduate study of German. English translations are assessed by one final-year student at the graduate study of English and one professional translator who graduated from the same study. Additionally, quality judgements are collected by asking the annotators to rate each translation on 1-5 Likert scale, where 1 means incomprehensible translation, and 5 means perfect translation.

Intra and interannotater agreement

The inter-annotator agreement per each language direction and error category and/or subcategory on a sentence basis is calculated in the first phase of the research. The calculation is performed by the equation in (1), where *all* stands for the total number of annotations by each annotator, and *agree* stands for the number of annotations on which agreement is reached. The practice is taken from (Stymne & Ahrenberg, 2012). Since annotators might agree on the label but not on the position merely because there are no guidelines on how to conduct annotation, only the agreement on the categories is reported, with the presumption that detecting errors is what matters after all, and not their precise position in sentences. Two out of four annotators asked questions prior to annotation. The questions concerned the appropriate positions for annotations of *missing word* and *word order* categories.

$$Agreement = \frac{2A^{agree}}{A1^{all}+A2^{all}}. \quad (1)$$

Results

Table 1 shows that Croatian-English is the best scoring translation direction according to human annotators. The worst scoring direction is German-Croatian. Figure 1 presents the distribution of quality scores per each annotator and language direction. Reluctance in assigning scores 1 and 5 has been observed. Interestingly, out of three translation directions which contain sentences scored 5, two of them also contain sentences scored 1. Both annotators evaluating the German-Croatian translation direction agree that there are some extremely bad sentences, while both annotators evaluating the Croatian-English direction agree that some sentence translations are perfect.

Pearson correlation coefficients are calculated between error frequencies and human sentence scores, and between error frequencies of selected categories and human sentence scores. The results are presented in Table 2. The coefficients between the total number of errors and human scores, as well as between the number of *incorrect words* and human scores are significant at $p<0.05$.

Numbers of errors per each top-level category are presented in Figure 2. The most represented category per all language directions is *incorrect word* category, followed by *missing word* and *word order*. According to the number of errors at the intermediate level of detail, which is not included due to space considerations, the most frequent subcategory is *extra word* for translations into English and German, while the biggest issue for translations from English into Croatian is detected with the *incorrect form* category.

Table 1. Human scores and total number of errors per direction and annotator

Translation direction		**HR → DE**	**HR → EN**	**DE → HR**	**EN → HR**
Avg score	E1	3.33	3.04	2.54	3.08
	E2_1/E2_2	-/-	3.71/3.63	2.58/-	3.17/3.13
	Avg	3.33	3.46	2.54	3.13
# of errs	E1	148	120	319	161
	E2_1/E2_2	-/-	79/96	216/-	133/131
	Avg	148	98.33	267.5	141.66

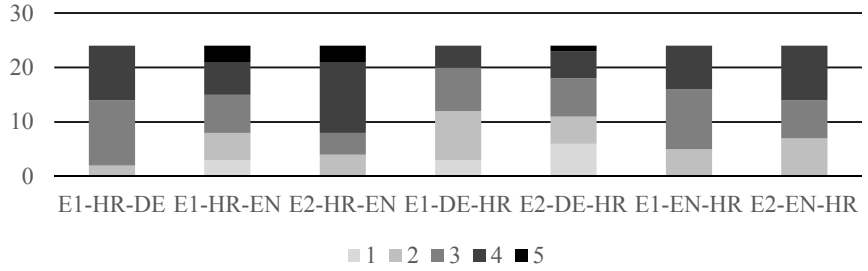


Figure 1. Distribution of quality scores per each annotator and language direction

Table 2. Pearson correlation coefficients between selected categories and human scores

Pearson	errs	missing	wo	incorrect	unk	punct	form	sense
Avg score	**-0.97**	-0.78	-0.1	**-0.97**	0.02	0.39	-0.90	-0.65

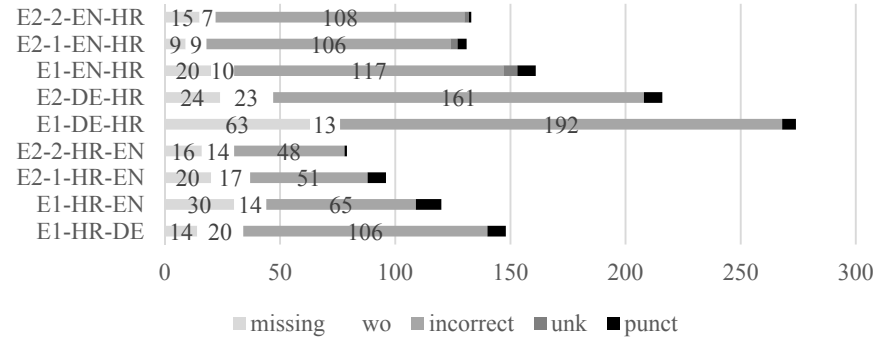


Figure 2. Number of errors per each annotator and translation direction at lowest level of detail

The intra and inter-annotator agreements for the top-level categories are given in Figure 3. Inter-annotator agreement could not be calculated for the Croatian-German language direction since there was no second annotator at disposal. Intra-annotator agreement is calculated only for translation directions involving English and Croatian due to time considerations. Intra annotator agreement is consistently over 60%, except for the *word order* and *punctuation* categories. This is due to low representativeness of these categories. As far as *incorrect words* are concerned, the agreement in translating into morphologically rich Croatian is somewhere between 65 and 75%.

Discussion

The Pearson correlation coefficients indicate that there is a statistically significant relationship between the total number of errors and human scores, as well as between the number of *incorrect words* and human scores.

In the study presented no guidelines are given to annotators. This is done purposefully to better understand different notions of error analysis. Since annotators might agree on the label but not on the position merely due to the lack of guidelines, positions are excluded from the study. If included, the agreement on different categories would be differently affected. A short reflection on the task follows. The annotators had trouble deciding on the number of errors in a phrase, i.e. should a phrase be treated as a unique unit or not (e.g. "Am schlimsten Jahre" can be annotated either as three errors of i*ncorrect form* subcategory or as one error of the same type). Furthermore, they lack determination on deciding whether a *missing word* is *content* or *filter*. One of the annotators showed the tendency to follow references too strictly and mark differences automatically, machine-like. Although at first glance it might seem that human annotators are rather forgiving as far as style is concerned, this can be attributed to the genre of the text. Journalistic texts are broadly represented on the Internet so they make an important part of GT training data. Therefore, the *style* category could be excluded from our further studies, except in highly specialized domains with a pronounced style, e.g. the domain of law. *Punctuation* category is pretty straightforward, i.e. annotators who are not language experts may fail to detect such error, but they will not mistakenly take it for another error type. It could be abandoned from further studies as well, in order to reduce the dimensionality of calculations.

A first draft of the guidelines which we either adopt from related work or compile based on the results obtained in this study is given as follows: (1) only after reading and comprehending translation, check your understanding by consulting source or reference sentence; (2) register all possible errors on a word; (3) mark as few errors as possible to make the sentence grammatically correct and semantically equivalent to the source; (4) if the meaning is affected, wrong preposition should be annotated as a *disambiguation* error; (5) use higher-level categories when it is not possible to use deeper-level categories; (6) mark *filter* word if unsure whether the *missing word* is *content* or *filter*; (7) if unsure whether an error is a *disambiguation* or a *lexical choice* error, consult the corresponding source word and confirm whether it can be used in both senses in order to opt for *disambiguation* error. Agreement per sentence reveals that sentences with many errors are too hard to annotate unanimously and should be excluded from assessment. The authors in (Lommel, Burchardt, Popovic, Harris, Avramidis, & Uszkoreit, 2014) use more than three errors as a threshold.

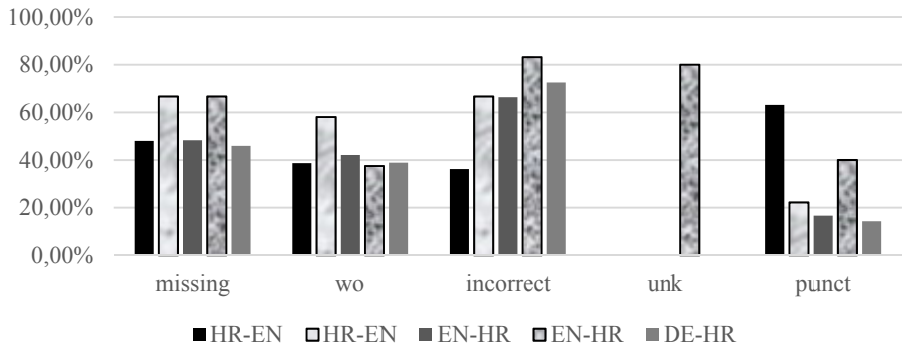


Figure 3. Intra and inter-annotator agreement per translation direction at lowest level of detail

Conclusion

The annotations obtained in this study enable us to detect the system and language-specific distributions of errors. The paper confirms that it is possible to get a reasonable agreement even without any guidelines. However, a first step towards compiling a complete list of guidelines is made. The analysis confirms the intuitive notion that the system best handles translations from a morphologically rich into a morphologically poor language. The opposite direction generates many *incorrect word* errors. A discrepancy detected in the number of *missing word* errors opens up new questions. Should it be attributed merely to the lack of guidelines since annotators might annotate one *incorrect form* error for two errors of types *missing word* and *extra word*? Presenting fine-grained agreement analysis where the kappa values are given for each error category is left for our future work. It would be interesting to show confusion matrices at each level of Vilar's taxonomy within the super-category. Such presentation would be informative enough, i.e. it would suffice to know that *sense,* no matter of what subcategory, might be confused with *incorrect form*. By analysing errors the annotators do not agree on, categories which are most easily confused may be pinpointed, and a list of guidelines may be expanded. Furthermore, it would be interesting to see how automatic error classifications proposed by (Popovic & Burchardt, 2011) or (Fishel, Bojar, Zeman, & Berka, 2011) correlate with the results obtained in this study.

Acknowledgement

This work has been fully supported by the University of Rijeka under the project number 16.13.2.2.01.

References

Brkic, M., Basic Mikulic, B., Matetic, M. "Can we beat GT?" Proceedings of the ITI 2012 34th International Conference on Information Technology Interfaces (ITI). 2012, 381--386.

Brkic, M., Seljan, S., Matetic, M. "Machine Translation Evaluation for Croatian-English and English-Croatian Language Pairs." NLPCS Workshop: Human-Machine Interaction in Translation. Copenhagen: Copenhagen Business School. 2011, 93--104.

Brkic, M., Seljan, S.; Vicic, T. "Automatic and Human Evaluation on English-Croatian Legislative Test Set." International Conference on Intelligent Text Processing and Computational Linguistics. 2013, 311--317.

Brkić, M.; Vičić, T.; Seljan, S. Evaluation of the Statistical Machine Translation Service for Croatian-English. // *International Conference The Future of Information Sciences*. 2009, 319--332.

Costa, Â.; Ling, W.; Luís, T.; Correia, R.; Coheur, L. A Linguistically Motivated Taxonomy for Machine Translation Error Analysis. // *Machine Translation* (2015): 127--161.

Elliott, D.; Hartley, A.; Atwell, E. A Fluency Error Categorization Scheme to Guide Automated Machine Translation Evaluation. // *Conference of the Association for Machine Translation in the Americas*. 2004, 64--73.

Farrús, M.; Costa-Jussa, M. R.; Marino, J. B.; Poch, M.; Hernández, A.; Henríquez, C.; Fonollosa, J. A. Overcoming Statistical Machine Translation Limitations: Error Analysis and Proposed Solutions for the Catalan--Spanish Language Pair. // *Language resources and evaluation.* (2011), 181--208.

Federico, M.; Negri, M.; Bentivogli, L.; Turchi, M.; Kessler, F. F. B. Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. // *EMNLP*. 2014, 1643--1653.

Fishel, M.; Bojar, O.; Zeman, D.; Berka, J. Automatic Translation Error Analysis. // *International Conference on Text, Speech and Dialogue*. 2011, 72--79.

Flanagan, M. Error Classification for MT Evaluation. // *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*. 1994.

Font-Llitjós, A.; Carbonell, J. G.; Lavie, A. A Framework for Interactive and Automatic Refinement of Transfer-Based Machine Translation. // *Tenth workshop of the European Association for Machine Translation (EAMT)*. 2005.

Koponen, M. Assessing Machine Translation Quality with Error Analysis. // *Electronic proceeding of the KaTu symposium on translation and interpreting studies*. 2010.

Lommel, A.; Burchardt, A.; Popovic, M.; Harris, K.; Avramidis, E.; Uszkoreit, H. Using a New Analytic Measure for the Annotation and Analysis of MT Errors on Real Data. // *Proc. of EAMT*. 2014.

Popovic, M.; Arcan, M. PE2rr Corpus: Manual Error Annotation of Automatically Pre-annotated MT Post-edits. // *LREC*. 2016.

Popovic, M.; Burchardt, A. From Human to Automatic Error Classification for Machine Translation Output. // *15th International Conference of the European Association for Machine Translation (EAMT 11)*. 2011.

Secară, A. Translation Evaluation - a State of the Art Survey. // *Proceedings of the eCoLoRe/MeLLANGE Workshop*. 2005, 39-44.

Seljan, S.; Dunđer, I. Automatic Quality Evaluation of Machine-Translated Output in Sociological-Philosophical-Spiritual Domain. // *10th Iberian Conference on Information Systems and Technologies (CISTI)*. 2015a.

Seljan, S.; Dunđer, I. Machine Translation and Automatic Evaluation of English/Russian-Croatian. // *Proceedings of Corpus Linguistics*. 2015b, 72--79.

Seljan, S.; Brkic, M.; Vicic, T. BLEU Evaluation of Machine-Translated English-Croatian Legislation. // *LREC*. 2012, 2143--2148.

Seljan, S.; Brkic, M.; Kucis, V. Evaluation of Free Online Machine Translations for Croatian-English and English-Croatian Language Pairs. // *Proceedings of the 3rd International Conference on the Future of Information Sciences: INFuture2011-Information Sciences and e-Society*. 2011, 331--344.

Seljan, S.; Tucaković, M.; Dunđer, I. Human Evaluation of Online Machine Translation Services for English/Russian-Croatian. New Contributions in Information Systems and Technologies. // *Springer International Publishing*, 2015, 1089--1098.

Stymne, S.; Ahrenberg, L. On the Practice of Error Analysis for Machine Translation Evaluation. // *LREC*. 2012, 1785--1790.

Stymne, S. Blast: A Tool for Error Analysis Of Machine Translation Output. // *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*. 2011, 56--61.

Vilar, D.; Xu, J.; d'Haro, L. F.; Ney, H. Error Analysis of Statistical Machine Translation Output. // *Proceedings of LREC*. 2006, 697--702.

Integrating Localization into a Video Game

Sanja Seljan
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences. University of Zagreb
Ivana Lucica 3, Zagreb, Croatia
sanja.seljan@ffzg.hr

Josip Katalinić
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences. University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
josipkatalinicandroid@gmail.com

Summary

The aim of the paper is to analyse the integration of the localization process within a video game from the engineering perspective, paying attention to tools and resources, processes and project management. Localization is perceived in the context of the global market with regard to language, market needs, time, monetization and project organization. In the paper, the integration of localization into project setting is analysed, then methods of automation of localization, quality assurance and time aspect. The authors suggest several stages of automation in the localization process, such as key-value pairs imported from spreadsheets, dynamic retrieval in the context, separation of code from the content, modularity and point out the importance of quality assurance. Investing in quality assurance of the localization is essential and it is achieved through a number of methods and storage structures where a systematic access to text ensures the visibility.

Key words: localization, video game, automation, project organization, quality assurance

Introduction

Video games, as an interactive media, are growing every day, weather through entertainment, education or business ventures. By expanding popularity, video game localization has become an indispensable part of the game production cycle and approaches to the global marketplace. According to Erbil (2016), media industry has significant influence on the evolution of digital product, comprising movies and video services, digital music, electronic publishing and video games, the last one showing the biggest revenue. Among the mentioned digital products, video games show the tendency of growth (67% of household in USA

play video games; revenue from video games and virtual goods in 2012 are eight times bigger than in 2009). Czech (2013) elaborates on video game industry which is seen as the fastest growing industry in the last five years, due to, among others, development of video game concept and the increase of cultural values in interactive entertainment.

Interactivity of video game media offers us unique experiences that can be perceived through localization, linking it to other cultures. Although localization is perceived from the wider perspective which requires technical, linguistic, project and marketing knowledge and skills, not many researches have elaborated an integrated and a technical perspective. Important question is what language to localize in? How to organize localization? How to implement automation process in localization? Publishing platforms like Google Play, App Store and Steam are also going to be used for exploring tendency of language and its connection to a particular platform. Monetization factor is also important in language selection, as certain markets and languages have greater profitability than others which is derived via market analysis.

Methods that are explained in paper describe systematic integration of the localization, supported via scripts for automation based on industry experience. Current industry is experiencing many shortcomings that are directly connected to the inability of creating reusable system that would have all variables stored in a separate resource file, so nobody needs to go through source code to add translated text into the game.

In the paper, the integration of the localization process within the video game is analysed, from the engineering perspective. In the second chapter the role of the localization process inside of GILT sector is pointed out. In the third chapter, localization is perceived in the context of the global market with regard to language, market needs, time, monetization and project organization. Levels of video game localization are discussed, as well as use of different tools, processes and the role of project management. In the fourth chapter, the authors analyse aspects of the integration of the localization into project setting, methods of automation of localization, quality assurance and time aspect. In the conclusion suggestions for further research and implementation are given.

GILT (Globalization, Localization, Internationalization, Translation) sector

According to W3C, "*localization* refers to the adaptation of a product, application or document content to meet the language, cultural and other requirements of a specific target market (a locale)". Localization refers to, as stated in LISA (2007), linguistic issues, physical modifications, business and cultural issues. Localization often assumes adaptation of settings that usually include number format, date and time zone formats, currency usage, keyboard usage, case conversion, collation and sorting, then symbols, icons and colours, paper size,

adaption of text, abbreviations, graphics and ideas that may cause misinterpretation, physical product adaptation, shortcuts, forms, etc. Localization may require a comprehensive review of logic, visual design or presentation if translation is significantly different from the culture of origin. Localization process can also include change of design, marketing message, object or colour, or paradigm in order to adapt the product to the target culture. Localization goes beyond text structure, including also technical, cultural and marketing strategies, as it is oriented to the final product, "without taking any standpoint on the source text or the relation between the source text and the target text", as in Czech (2013). Localization is often referred to as 'l10n' indicating 10 letters between 'l' and 'n'.

According to Esselink (2003), localization has moved from in-house localization in 80's to internationalization in 90's, focusing on project management, programming and publishing environment using XML, Java and .NET. At the beginning, localization was separated from the products, while today developers create software with source files created for internationalization process. Later on, the expansion of software and documentation has asked for the process of internationalization, which precedes the localization process.

According to W3C, the term *internationalization* refers to the "design and development of a product, application or a document content that allows a localization for a target audience that is different from source product culture, region or language" or, according to LISA, "the process of enabling a product at a technical level for localization". In that sense, it can be said that internationalization precedes and facilitates the task of localization. For internationalization, the numeronym 'i8n' is frequently used. Internationalization assumes the level of abstraction from the specific culture, language or the market. It is performed only once in the development process and can influence the localization process, which impacts ease of technical issues, costs and business processes.

While translation is the process transferring content from one language (original language) to another language (target language) by preserving the context, *localization* process is more specialized and oriented to a final product in order to meet expectations of the global market, as indicated on web pages of Lionbridge. Pym (2005) defines localization as translation + factor X, where factor X includes internationalization, adaptation, controlled writing, use of translation memories, integration of machine translation, project management, i.e. technical knowledge and skills, management and marketing issues.

Šiaučiūnė and Liubinienė (2011) point out that "*localization* encompasses not only activities of traditional translation (terminology research, editing, page layout, proofreading), but also includes multilingual project management, software and online help engineering and testing, conversion of translated documentation to other formats, translation memory alignment and management, multilingual product support, and translation strategy support. One more important difference between translation and localisation is that the former is per-

formed after the product or document is finished and released, and the latter usually runs in parallel with the development of the source document or product."

According to Erbil (2016), *internationalization* is related with developers and copywriters of the content who are engaged in design and development step. In the localization process, various experts engaged in product adaptation are involved, such as engineers, translators, project managers and test engineers. Globalization is related with sales and marketing activities in order to present the product at international market. Translation is often done with the help of software where it is commonly used in form of machine translation (MT) or computer-assisted translation (CAT), requiring background knowledge of these technologies for successful implementation. Software localization tools include the whole range of commercial and free software created for Windows or Macintosh platform. CAT tools provide stable platform for immediate sharing and updating translation memories (Brkić et al., 2009), which can be used integrated with MT technology and terminology management software.

Source: LISA. The Globalization Industry Primer, 2nd edition, 2007.

Figure 1. The global product development cycle

Localization of Web sites often require use of content management systems (CMS) in order to create, manage, store and publish complex sites that often contain functionalities that require an engineering knowledge. Localization also includes other types of software, such as tools to track changes on Web sites, word counting tools, currency converters, etc. Publishing is often performed using XML, Java and.NET environment. XLIFF (XML Localization Interchange File Format) is XML-based format used to store content extracted form original file and the translated file, TMX (Translation Memory Exchange) for the storage and interchange of translation memories and TBX (Term Base Exchange) as a model for terminological databases.
According to LISA, the final stage of this process is globalization which is associated with "all issues to make a product or service really global, integrating also business processes, marketing, sales, customer support, etc. The whole process in known as GILT (Globalization, Internationalization, Localization, Translation) process where various departments (development and planning, design, production and technical department, translation, marketing) and experts are included.

Video game localization

The problem with video games, unlike for e.g. documentation or technical texts containing repetition and being suitable for CAT, is that video games mostly consists of creative texts (e.g. artistic descriptions of landscapes). Video games also do not use controlled languages (except programming language) or any scheme in the text creation phase. Moreover, game makers often employ professional writers who create a story for a particular title. The nature of video game texts is one of the reasons why they are not easily translated using the CAT tools or machine translation, where sublanguage is often used (Seljan, 2000). The second difference between video games and technical translation is the medium. Video games are embedded in information technology. As in Mateusz (2014), video game translation is interwoven with the source code of a particular product. As presented in Šiaučiūnė and Liubinienė (2011), Chroust (2007) elaborates localization as multi-layered process in the form of pyramid performed at different levels:

- technological infrastructure – as the basic level, including all technological and organizational preparation before and during the localization
- grammatical and semantic layer – as the second and third layers, including textual translation and language expressiveness
- graphic and iconic representation layer – which can be changed when localizing software, games or web sites
- business conventions and practice layer – important for e.g. for contractors or the company who can buy the rights to perform localization

- social and communication layer and cultural layers – the last two layers related to specific cultural aspects or specific market.

In today's work, there are numerous solutions to localization, but outsourced localization team is often a solution. Numerous companies and associations offer their services on the market for localization, although sometimes it seems as a major and perhaps unnecessary investment, localization is today considerably important process for increased competitiveness in the global marketplace. Cooperation with the localization can start from the preparation of the software that will use localization, continuing in regular intervals in order to avoid pauses in production, also in Soh et al. (2016). Software preparation through internationalization greatly facilitates the localization team to see and test the context of use, thus increasing the ability to test the integration of localization and the quality of the translation. Quality control through communication with the localization team within the development process, and clear use of unambiguous words leads to successful localization project, as in Cem et al. (2016).

Cardenosa (2006) explains three main approaches in the architecture of the internationalization process:

- when messages, menus and other culturally sensitive factors are contained in the source code of the application – for this case it is necessary to develop different version for each language, which multiplies costs and time
- when messages are extracted into external library, the application is generated from the common code and linked to the specific language library
- the third approach consisting of core application with all functionalities, culturally independent and linked dynamically to external localization resources containing all specific information.

According to Erbil (2016), different game companies follow today different levels of localization, pointing out four levels of localization:

- the first level when *localization is not performed* because of cutting costs. At this level introductory text, content and documentation are not localized. It can be expect that from fan groups to translate the game.
- the second level, called "*box and documentation localization*", assumes that mostly introductory text, cover text and game manual are translated, while in-game text, visual content and dubbing are in foreign language
- the third level is *partial localization*, which includes all in-game texts and visual text translated and localized, while dubbing and effects are in original language
- the fourth and last level is *full localization*, which includes all elements localized (introductory text, visuals on platform when buying the product, text content, documentation and rich content of the game, e.g. dubbing, effects, videos).

Another area of relevant development are online games, which often incorporate chat among players while the game is being played. This implies a need for simultaneous translation of text chat and chat based on speech, as online games have no regard for linguistic boundaries. This raises the scope for research into the application of natural language processing systems.

Benefits of localization

According to data presented by ICT Facts & Figures (2015) development of the mobile gaming market and the adoption of the Internet show that localization contributes to sales in other markets. In addition to monetizing localization, it also enables the convergence of creativity writing to many users who can enjoy localized texts, which would not be possible if the game was released in a language they do not know.

By publishing a localized game to a new market, it is possible to acquire a new community that monitors our product and becomes a future buyer if the video game has multiple sequels or downloadable content. If the budget is limited and does not allow localizing the entire game, this problem can be approached by localizing only the application keywords and thus accessing more actively global market where users potentially use the language of the application, but search in their native language. Keyword localization is a far cheaper and faster solution then localizing the entire game, but a research is required for each language as well as statistics of localization, using tools such as GKWT, Google Trends and App Store Search Suggest, etc.

Users tend to buy more if content is written in their first language and in that way localization is contributing to monetization by opening up markets to numerous customers as potential buyers. According to data on Statista web pages, Asia and especially China rapidly enter the global video game market, which is expected with regard to the population and the economic growth of the middle class. Sales of smartphones are also on the rise, a sales growth of iPhone devices in China exceeds sales in the US and each iPhone owner is a potential buyer of a video game.

Integrating localization

Challenges that localization poses through integration within the game are not negligible and the localization process should be considered when setting up a project. Localization process can start together with the development of the software product and online documentation. For this part, translation memory technology can be used, as it enables cooperative work, sharing and updating and interface localization tools to translate resource files. Late integration of localization within the game is leading to increased pressure on production, since it is necessary to integrate, check the integration accuracy as well as the localization itself, within the short time that is set in the production cycle.

Localization integration is possible to implement in different ways, e.g. a *key-value pair* or ID–localized text. If there is a game containing texts that needs to be localized, it is suggested to use a central place (e.g. Google Sheets) where it is possible to enter IDs of text to be localized. For example, the option is to create for e.g. ID DOCTOR_SPEECH with the value 'Hello' in English, and to have a look at ID DOCTOR_SPEECH as a reference to value that can be variable. In the next step, it is possible to send for translation using a file containing IDs pointing to a text that needs to be localized. The value of the ID varies depending on the language settings within the video game: if the language is set up for the German, DOCTOR_SPEECH will reference the value 'Hallo'.
Modularity can be achieved using IDs characters with defined values that can be replaced inside the programming language. For example, in order to create modular talk with a doctor that depends on his mood, it is possible to create within the Google Sheet ID such as DOCTOR_SPEECH_MOOD with value 'Today I am %mood%'. Using Python programming language with the replace method we can easily replace the first argument with another argument. The mood that is placed through the programming logic can then be replaced by %mood%. We also have to note that translation columns within the Google Sheet need to have brief context of message so translators can use that message and the name of defined values inside IDs to make better understanding of what could be placed inside those values.

current_mood = DOCTOR_SPEECH_MOOD.replace('%mood%', 'sleepy')

In this case the variable current_mood contains the value of 'Today I am sleepy'.
Another positive aspect to IDs access is that it considerably facilitates localization by *separating content and code*, and thus *enables integration of modularity*. There are different localization techniques as (Holovaty 2009) explains, with usage of *gettext* via the standard gettext module that comes with Python in context of Django web framework, showing that it is possible to add translation strings to tell Django that "This text should be translated into the end user's language, if a translation for this text is available in that language." Philips (2015) provides example of index.html Django page with three strings that require localization "Log in", "Username" and "Password". To start with, there is a list of language which the site support inside settings.py with:

```
LANGUAGES = (
    ('en', _('English')),
    ('fr', _('French')),
)
```

After that translation template tags are required, which are accessible by putting `{% load i18n %}` at the top of the template files. There are two i18n translation template tags: trans and blocktrans, where `trans` tag marks a string for translation while `blocktrans` is used if translation require variables (placeholders). Once the text is put inside trans tag, our index.html template will look like:

```
{% load i18n %}
<h1>{% trans 'Log in' %}</h1>
<label>{% trans 'Username' %}</label>
<input id='password' type='text'/>
<label>{% trans 'Password' %}</label>
<input id='password' type='password'/>
```

After the text is wrapped in `trans` tags, it is possible to proceed to creating translation files with

```
python manage.py makemessages -l 'fr'
```

by which .po file is created, which contains actual translations. Each language has its own .po file, usually saved inside of application/locale/<lang_code>/ LC_MESSAGES/django.po. The following information is presented:

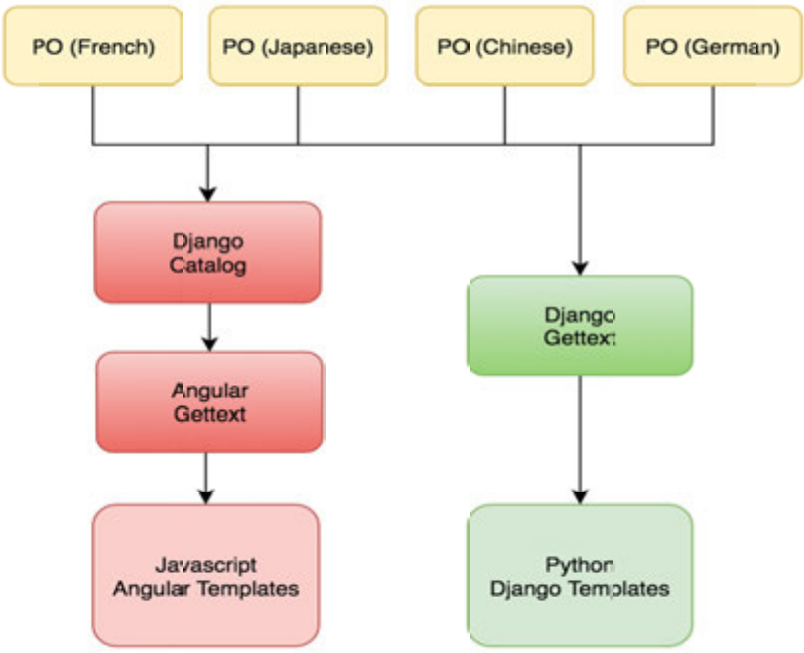
```
#: login.html:2
msgid "Log in"
msgstr ""
#: login.html:4
msgid "Username"
msgstr ""
#: login.html:7
msgid "Password"
msgstr ""
```

After replacing msgstr with translation, it is possible to proceed to creating machine object (.mo) that contains object data referenced by a program:

```
> python manage.py compilemessages
```

The files .mo and .po, after changing the language setting, are presented. Gettext can be used in other web frameworks, like AngularJS as shown in Figure 2 where we see implementation of Angular Gettext with JavaScript Angular Templates based on Django Catalog. Here Django Catalog means Django Message Catalog that corresponds to directories that are created and organized by language with their .po and after compiling .mo files. Standardization of .po

files enables us this power, so we can use it with different web frameworks, Figure 2 also shows on the right side more traditional way of getting translations strings from .po files in the manner what we already discuss previously.



Source: Translation in Horizon. 07.09.2017

Figure 2. Message Substitution

Organization of localization in multiple documents is desirable, in order to avoid creating a large single file of the localization document. For example, it is possible to store the menu localization in one document and button texts in another. By subdividing content into categories, it is easier to find specific localized data and referencing within the code. Separation of the content from the code makes communication easier. The localization process including various experts entails structured and organized workflow in order to obtain good quality product localized on time.

Quality assurance

Ensuring localization quality is a multi-dimensional problem that can be approached through the software design, here video game, and by dividing the product into two concept blocks:

- a block containing all the elements of the user interface and
- the block containing the execution code.

The user interface block contains only localized elements such as text, error messages, windows, menus and so on. The execution code block contains only the application code that will support the localization. Such structure allows testers, as well as programmers, to quickly and systematically test and implement localized video game, saving time and money in the process. In the perfect situation, an original speaker would perform testing of implemented localization. Every problem needs to be described, recorded, evaluated and finally corrected. After performing all corrections, the new version is sent for verification. As this process takes a long time, it is important to plan for it. A crucial factor in ensuring quality of localization is the time, since the localization team should start working together with the development of video game as early as possible, thus avoiding the production delays that may arise with the arrival of a large number of localized texts. An estimation of the required project resources is typically carried out by a project manager in a localization service company. The localization team activity through questioning within the production cycle is a key factor in quality assurance, making the context of localization clearer.

One of the possible acceleration of localization testing within the game can be achieved by setting and grouping localized text within the Google Sheet so that each spreadsheet denotes a particular entity. Therefore, it is possible to have spreadsheet of user interface, text buttons and windows. The search could be facilitated by setting up description of each localized text, for example if we have original English text in column A, then we can have description in column B and localized text in column C. Sanchez and Lopez (2016) propose the following *methods for quality assessment* in localization testing, where testers can choose one or more methods, depending on time and budget:

- step-by-step testing,
- ad hoc testing,
- screen testing,
- emulation.

Testing step-by-step is the most advanced method. It basically consists of the step-by-step testing plan presented in the specific order (e.g. testing dialogues, menus, descriptions, etc.). Testers should follow this plan, trying not to skip any of steps.

Ad hoc testing complements the previous technique. The testers leave the test plan and checklist through playing the game, doing the actions that each standard player would. This type of testing can be useful for finding problems at locations that were not included in the test plan or to focus on details in a similar thinking style as the player.

Testing the screen is done by sending a screen instead of the game with which testers interact. However, this method is only useful if the number of screenshots is not too large, since screen capture for each language takes a time (although sometimes the process can be automated).

Emulation though not a method as such is a useful resource in certain situations. Emulation is a set of additional debug options inside the game that is currently in production, so testers can easily simulate different aspects of game without playing through whole game. The emulation method is frequently used because it saves time and enables checking (e.g. level 5 out of 7). Testers must be able to use these application testing options correctly.

Conclusion

The aim of the paper was to analyse the integration of the localization process within into a video game from the engineering perspective, paying attention to tools and resources, processes, project management and quality assessment. Localization is perceived as a multi-layered process, where attention is given to technical and organizational layers. The authors suggest some aspects of automation in the localization process, such as key-value pairs imported from spreadsheets, dynamic retrieval in the context, separation of the code from the content, modularity and importance of the quality assurance. Investing in quality assurance of the localization is essential and it is achieved through a number of methods and storage structures where a systematic access to text ensures the visibility. The authors also point out simultaneous collaborative work on the same project and need for integration of the localization process from the beginning of the project setting. Localization integration is an important factor that will have to be standardized for fast and efficient integration through the process of internationalization that precedes the localization. The authors suggest continuing the research on standardization of the technical and organizational layers in the domain of localization engineering (also in Soh et al., 2016) and correlating it with output efficiency.

References

Brkić, M.; Seljan, S.; Bašić Mikulić, B. Using Translation Memory to Speed up Translation Process. // INFuture2009 - Digital Resources and Knowledge. Zagreb: Department of Information Sciences, 2009.

Cardenosa, Jesus; Gallardo, Carolina; Martin Alvaro. Internationalization and localization after system development: a practical case. // International Journal Information Theories and Applications, 2006.

Cem, Odacıoğhu Mehmet; Loi, Kim Chek; Köktürk, Şaban; Uysal, Müge Nazan. The Position of Game Localization Training within Academic Translation Teaching. // Journal of Language Teaching and Research, Vol. 7 (2016.), No. 4

Chroust, G. Software Like a Courteous Butler Issues of Localization Under Cultural Diversity. // Proceedings of the ISSS. 5th Annual meeting and Conference for the System Sciences, 2007.

Czech, Dawid. Challenges in video game localization: An integrated perspective. Explorations. // Journal of Language and Literature, Vol. 1 (2013.), No. 1; 3-25

Erbil, Mert. Video Game Localization Factors and Impacts on Digital Purchasing Behavior, Dissertation, 2016.

Esselink, Bert. The evolution of localization. // The Guide from Multilingual Computing & Technology: Localization, (2003.), No. 57 Supplement; 4-7

Holovaty, Adrian; Kalpan-Moss, Jacob. The Definitive Guide to Django. New York: Apress, 2008.

Krammer, Connor. Persona 5: Phantoms of Translation 02.05.2017. http://www.personaproblems.com/ (07.09.2017.)

LISA. The Globalization Industry Primer, 2nd edition, 2007.Mateusz, Sanja. Computer-assisted translation tools and video game rendition. // The Translator and the Computer 2, 2014

Muñoz Sánchez, Pablo; Sánchez López, Rafael. The ins and outs of the video game localization process for mobile devices. // Revista Tradumàtica: tecnologies de la traducció, (2016.), No. 14

Nichols, Brian. The Difference BetweenTranslation and Localization for Multilingual Website Projects. 28.07.2015. http://content.lionbridge.com/the-difference-between-translation-and-localization-for-multilingual-website-projects-definitions/ (25.04.2017.)

Philips, Noal. Translating your site with Django 1.8 03.10.2015. https://medium.com/@nolanphillips/a-short-intro-to-translating-your-site-with-django-1-8-343ea839c89b (06.09.2017.)

Pym, Anthony. Translation and Localization. 21.10.2005. http://www.fut.es/~apym/on-line/talks/localization_bergen_2005.ppt (08.06.2017.)

Sanou, Brahima. ICT Facts & Figures. May 2015. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf (29.04.2017.)

Seljan, Sanja. Sublanguage in Machine Translation. // Proceedings of 23rd Int. Convention MIPRO 2000: Computers in Intelligent Systems CIS + CTS. Rijeka: Liniavera, 2000.

Šiaučiūnė, Vaida; Liubiniene, Vilmante. Video Game Localization: the Analysis of In-Game Text. // Studies about Languages, (2011.), No. 19.

Soh, Mathurin; Nkenlifack, Marcellin; Fotso, Laure P. A New Hybrid Process for Software Development and Localisation. // International Journal of Scientific and Engineering Research, Vol. 7 (2016.), No. 6; 945-953

Steam Hardware & Software Survey. April 2017. http://store.steampowered.com/hwsurvey/ (02.05.2017.)

Translation in Horizon. 07.09.2017. https://docs.openstack.org/horizon/latest/contributor/topics/translation.html (07.09.2017)

W3C. Localization vs. Internationalization. https://www.w3.org/International/questions/qa-i18n (12.6.2017.)

Data Storage Devices in Science Fiction and Fantasy Movies

Domagoj Bebić
Faculty of Political Science, University of Zagreb
Lepušićeva 6, Zagreb, Croatia
Domagoj@edemokracija.hr

Summary

Can we think about, make conclusions, observe, or predict social issues through movies? Through the analysis of selected science fiction and fantasy movies, the goal of this research is to identity the depiction of data storage technology and devices to see if the solutions offered were realistic, possible, and plausible or were just imaginable notions contrived for a movie script. For this purpose, several science fiction and fantasy movies are analyzed: 2001: A Space Odyssey (2001), Star Wars (1977), Superman (1978), Star Trek III: The Search for Spock (1984), Star Trek: Generations (1994), Johnny Mnemonic (1995), Minority Report (2002), and The Avengers (2012). The analysis showed each of the selected movies presented several revolutionary solutions for data storage devices. Several of the movies presented technology that is not yet available but probably will be in the near future. In detecting correlations between realistic technological developments and fantasy movie ideas, the analysis showed that while several movies accurately predicted some of the contemporary technological solutions for data storage, several just made fictional usage of these kinds of devices.

Key words: science fiction, fantasy, movies, data storage device, popular culture, technology

Introduction

In the pilot episode of the original TV series Star Trek, "The Cage" (1965), a race of technologically advanced humanoids called Talosians, after they have examined the databases of the USS Enterprise (Spock: "Tapes, micro-records, everything"), they concluded about humans: "Their method of storing records is crude and consumed much time."

As humans, we have a limited capacity for data storage and therefore remembering memories and experiences is not sufficient. That is why there is a constant need to save memory, to save data, and to invent procedures, devices, or solutions for its storage. One part of those options is for data to be transferred on different media to be remembered / archived as means of transferring knowledge. Whether it's for a family photo album, a computer program, or a

Fortune 500 company's business-critical systems, data storage is a must-have for nearly everyone (Zetta, 2016).

Furthermore, data are important in our lives; it also serves as a significant inspiration in popular culture, especially in movies. Maybe the best illustration of data storage caption in movies is the legendary communicator, from the Star Trek TV series, which, as portrayed in the original series, is acknowledged as the first depicted mobile device (Wikipedia, 2017). This paper focuses on science fiction and fantasy movies in order to determine whether such movies are imaginative representations of reality or are only fictional manifestations of the movie's plot. Many authors were keen to explain the connectivity of technology and science fiction and fantasy films (Isaac Asimov, 1988; Bruce Sterling, 1988; Larry Niven, 1998; Barbara Hambly, 1982; Dan Simmons, 1989; Robert Heinlein, 1973). Also, several authors were involved in near future predictions throughout their science fiction work. In many cases, science fiction and fantasy movies are examples of genres that imply the free use of one's imagination and the depiction of currently unimaginable issues for contemporary culture and society. With this paper, however, the goal is to start identifying a number of segments within such movies that feature the use of data-storing devices, and decide whether these are presented as realistic solutions for data storage or just as imaginable solutions restricted to the movies. With a focus on detecting instances of stored data, transfer of data or back-up, this paper will explain such data devices in several science fiction and fantasy movies. The goal is to determine how science fiction and fantasy movies imagined and depicted data storage technology and whether these were realistic solutions or were only imaginary parts of popular culture. While this is the first notice, the movies for this paper are randomly selected due to their popularity and impact. For this purpose, several science fiction and fantasy movies will be analysed: 2001: A Space Odyssey (2001), Star Wars (1977), Superman (1978), Star Trek III: The Search for Spock (1984), Star Trek: Generations (1994), Johnny Mnemonic (1995), Minority Report (2002) and The Avengers (2012). Although there numerous technology-related films today, these are selected because of their IMDb[1] rating and their popularity and impact. With this paper, the aim is not to explain the technology of the science fiction and fantasy movies themselves, but to provide review of data storage technology movie presentation in last forty years.

[1] The Internet Movie Database (abbreviated IMDb) is the world's most popular and authoritative source for movie, TV and celebrity content. The IMDb consumer site (www.imdb.com) is the #1 movie website in the world with a combined web and mobile audience of more than 250 million unique monthly visitors (IMDb, 2017).

From punch cards to iCloud – An overview of data storage technology

Mark Andrejevic, Alison Hearn, and Helen Kennedy (2015) declare that we are in the "age of big data" and that new analytical methods and businesses are emerging daily seeking to monetize this explosion of emerging data (Andrejevic et al., 2015: 379). Without technology and new, creative technological inventions, it would be impossible to capture, save, and store data. Therefore, technology plays an important, even crucial, role in the multiplying new possibilities for efficient data storage.

As Machles (1999) has pointed out, people are fascinated by the wonderful abilities of modern electronics and many have recorded countless hours of data on computers and other devices such as video recorders, both at home and at work. Furthermore, Machles (1999) explained that the same way of thinking has occurred with the use of compact discs and computer information. Documents that once were painstakingly handwritten or typed were replaced quickly by the use of computer discs, electronic copies, and similar storage media (Machles, 1999: 404). Today, we are witnessing the surge of modern, fast, and efficient data storage technology as every form of technological invention and data storage technology developed through the course of human history has been completely transformed. Today, the cloud is not just making data storage easier and more convenient—it is providing a platform for businesses and services, thereby building the next era of computing (Zetta, 2016).

The article (Gadgets, 2015) describes the chronological development of media for data storage from the eighteenth century up until the end of the twentieth century. As the authors explain in their article, the first technological data storage device consisted of punch cards, from 1725. After that, Alexander Bain, the inventor of the fax machine and the electric printing telegraph, patented punched tape in 1846. Selection tubes were invented in 1946 and magnetic tape in 1950. In the sixties, compact cassettes, the magnetic drum, and the floppy disc were invented. And in the seventies, the LaserDisc (1972) and the compact disc (1979) were invented. From the nineties, DVD and USB discs were significant data storage devices.

All these devices were important technological inventions but, with their development, also their storage improved. Today we are well acquainted with modern data storage media such as memory cards and USB discs but also cloud-based technology. Bryan Clark (2015) explains that hard drive technology is getting cheaper and storage capacity is improving and there is the move toward smaller internal storage is largely due to the expanding use of cloud-based technologies in order to store data, files, photos, videos, and more.

Movies as an important popular culture genre in the explanation of important social issues and concepts

With the rise and popularization of media technologies, popular culture also has undergone a significant rise in popularity. Danesi (2015) defines popular culture as a "culture for the masses" and, because it is oriented toward people, it is one of the best representations of society's development (Danesi, 2015: 25). The possibilities of digital technology have contributed to a higher access to the genres and versions of popular culture to an ever-widening audience. Beer and Burrows (2013) assert that popular culture is at the center of the transformations that have facilitated the accumulation of digital data; but it is also at the heart of the issues and debates that face the social sciences (Beer, Burrows, 2013: 47). That is to say, we need to think about the way in which popular culture is folded into this "performativity of circulation" (Beer & Burows, 2010: 50). As Laura Grindstaff (2008) has explained, probably the best-known strand of sociological research on popular culture is the "production of culture" perspective, which refers to the empirical study of culture-producing organizations within specific institutional contexts.

Today, while there are more and more versions and genres of popular culture, movies present an important and prominent part of popular culture. According to Jarvie (1978), at one time, movies were beneath the notice of academics and they were ranked lower than such acceptable forms of relaxation as detective mysteries (Jarvie, 1978: 374). Claudia A. Barriga, Michael A. Shapiro, and Marissa L. Fernandez (2010) explain that movies, particularly certain genres such as science fiction, medical dramas, and catastrophe films, often contain bits of scientific information with varying degrees of accuracy. There is evidence that the impact of fictional messages about unfamiliar groups may influence our beliefs about them just as much, or more, than nonfictional messages (Slater, 1990), showing that in certain cases fiction may have more influence than educational or informative messages. Anthony M. Hudock, Jr. and Sherry A. Gallagher Warden (2001) claim that several articles have discussed unique aspects of fictional movies, such as the richness and variety of content, emotive substance, and flexibility for use in assignments, making them appropriate for learning and training in the classroom (Chandler, 1997; Gladding, 1994; Gladstein & Feldstein, 1983; Koch & Dollarhide, 2000; Maynard, 1996; Voller & Widdows, 1993). This indicates that movies play a significant role in the perception and imagination of future events, and they can provide solutions and ideas for technological and social issues of contemporary times. That is why this research seeks to examine the possibility of connections between the popular culture representation of realistic issues and needed solutions, in this case for data storage.

The role of movies on human perception and imagination is also explored by Aik-Ling Tan, Jennifer Ann Jocz, and Junqing Zhai (2017) in their analysis, Spiderman and Science: How Students' Perceptions of Scientists Are Shaped by

Popular Media. In this research, the authors concluded that popular media, such as the Spiderman movies, have an influence on how young children perceive science and the work of scientists and that it is important to provide children with images and descriptions of a variety of science careers in order to motivate them to pursue further study in science and possibly science as a career.
Several authors (Erigha, 2016; Laan, 2010) also have wanted to identify the connections between movie presentations and realistic perceptions of common issues. While underlying themes of sci-fi media tackle contemporary, social, political, moral, religious, technological, and environmental issues (Chow-White, Deveau, & Adams, 2015), sci-fi filmmakers also make philosophical conjectures about the world, human existence, and the future (Erigha, 2016: 550). Andrew Milner and Sean Redmond (2015) consider science fiction to be a genre that challenges and critiques the status quo and has the ability to imagine other worlds that work in opposition and contradiction to the habitus of contemporary life (Milner, Redmond, 2015: 4).
In the context of technology presentation and science fiction movies one important name is Frank Herbert and his epic science fiction novel Dune (1965). Dune is considered as one of the greatest science fiction novels of all time (Dune novels, 2009). The first novel also inspired a 1984 film adaptation by David Lynch, the 2000 Sci-Fi Channel miniseries Frank Herbert's Dune and its 2003 sequel Frank Herbert's Children of Dune (which combines the events of Dune Messiah and Children of Dune), computer games, several board games, songs, and a series of followups, including prequels and sequels, that were co-written by Kevin J. Anderson and the author's son, Brian Herbert, starting in 1999 (Wikipedia, 2017). On the other hand, many films are also involved in technological prediction of the future Iron Man (2008 – 2013), The Matrix trilogy (1999 – 2003), Her (2013), Eternal Sunshine of the Spotless Mind (2004), Total Recall (the 2012 remake), Elysium (2013), Prometheus (2012), Ender's Game (2013), Paprika (2006), Tron: Legacy (2010), Wall-E (2008) (Chung, 2014). This paper also contributes to the theoretical explanation of technological presentation of data storage.

Imagination or reality: data storage technology in science fiction and fantasy movies

As has been stated, the goal of this research is to examine a selection of science fiction and fantasy movies to identify data storage technology and devices and determine whether these movie data storage solutions were realistic presentations of possible and realistic data storage solutions or were just imaginable notions contrived for a movie script.
The analysis includes several science fiction and fantasy movies: 2001: A Space Odyssey (2001), Star Wars (1977), Superman (1978), Star Trek III: The Search for Spock (1984), Star Trek: Generations (1994), Johnny Mnemonic (1995), Minority Report (2002) and The Avengers (2012). These productions are pro-

cessed chronologically in order to detect the evolution of technology and movie solutions for data storage. The goal of this analysis is not to discuss or analyse any particular narrative of a movie or story context, but to detect data storage technology and the movie's solution for its storage.

2001: A Space Odyssey (1968)

In 1968, many of the technological solutions depicted in this movie were just ideas but are today realistic, and most of these ideas have come true and are part of contemporary technology. Fifty years ago this movie predicted technological innovations that have become possible in the last decade.

Among the interesting data from this movie is a scene of pilots watching news from video pads (BBC 12 World Tonight); today, of course, we can connect to numerous news sources with tablets or any touch screen. In addition, when it comes to writing down important space station measurements, the pilots are shown writing them down with a pen on a notepad. Another innovative technological solution from this movie is the role of the main computer. The computer—HAL 9000—actually was the ultimate villain. In his final confrontation with the rebellious computer HAL 9000, astronaut Dr. David Bowman (aka Dave) has no other option but to remove HAL's memory from the spaceship's Logic Memory Center, which is portrayed as a small room containing hundreds of memory slots. In the movie, we can see how Dave is taking out the most important ones. Data memory is presented on numbered transparent PVC plates (separately for memory and for logic).

Star Wars (1977)

In the first movie of the Star Wars Space Saga (Star Wars IV: A New Hope) from 1977, the story is concretized around stolen Death Star plans, which are stored on a thin memory card that is hidden in the R2D2 Unit (or just R2). In the movie, this memory card is depicted as a data storage device while a robot (R2) is shown as displaying or projecting the data in the form of a hologram. The memory card presented in the movie Star Wars is a combination of the floppy disc that was available at the time and the compact disc (CD) that was invented only two years after the movie's appearance. Also, at the time the movie was recorded, no technology existed that could present data in the way that is shown in the movie. Today we are using similar technology with CDs, memory cards, or USB discs and projectors that can present data. The movie also predicts a 3D presentation of the data. And while the movie is revolutionary in its explanation and presentation of the universe and distant worlds, it is equally revolutionary in the technological solutions it depicts for data storage devices. Today, 3D presentation technology is still in development, and we can say that this movie predicted this kind of technological data presentation.

Superman (1978)

The crystal in Superman (1978) presents an out-Earth technology that not only contains data, but also acts as a power source that is not known to the human race. Jor-El from the planet Krypton sends this crystal in a spaceship with his son Kal-El to the Earth before the destruction of his planet. Not only does the crystal contain all the recorded data from the destroyed Krypton, it also possesses an indescribable power. Besides data, feelings, and power, the crystal also has the ability to present data: when it is placed on a portal, the crystal projects a hologrammatic message.

This is the text spoken by Jor-El to his son in the 1978 Superman movie:

"The total accumulation of all knowledge spanning the twenty-eight known galaxies is embedded in the crystals which I have sent along with you. Study them well, my son. Learn from them."

In Superman, the crystal is presented as a technology that is not available for humans and humans are unable to use it. That is why this technology for data storage is still not available for humans.

Star Trek III: The Search for Spock (1984) and
Star Trek: Generations (1994)

From the beginning of the Star Trek movie series (1979), we can see extensive usage of the tablets that are called control pads. In the very first movie when Captain Kirk is transported to the Enterprise and its crew is rushing to make the Enterprise functional in a short time due to an unexpected mission to stop an alien force hidden in a cloud of energy moving toward Earth, we can see that most of the crew are using control pads to perform their work assignment.

In The Search for Spock, an original series movie (1984), when Klingon Commander Kruge is contacted by Valkris who wants to transmit data about Project Genesis, she does so via porting with a kind of cassette device. At the time the movie was made, cassette technology was available and familiar, the same technology we can find on a Klingon ship, which leads us to conclude that this is widespread intergalactic technology. In the Next Generations Star Trek series movie, Star Trek: Generations (1994), we can notice that after Captain Picard receives the information that his brother Robert and his nephew Rene have died in a fire back on Earth, he invokes memories of his family by looking at an old-fashioned photo album. In the scene, the data are saved in an old-fashioned way and there is not much technology involved other than the lighting frames of the photo. However, while in this scene there was no technological data storage, in other scenes there were a few revolutionary data storage solutions. For example, the multidimensional Data Room called Stellar Cartography where Captain Picard and Commander Data are trying to understand the destination of Doctor Soran. Stellar Cartography is not a holodeck; rather, it is an interactive multimedia multiscreen database. This technology is later elaborated in other movies,

for example, the X-Men movies, but still is not among contemporary technological possibilities.
Also, another interesting scene occurs when Doctor Soran pays a Klingon to get him safely to the planet of Veridian III; he holds out a designed memory stick, saying: "This contains all the information you'll need to make a trilithium weapon." He emphasizes that the stick is coded so he can get safely to the planet surface of Veridian III, after which he will send them the decryption sequence.

Johnny Mnemonic (1995)
The plot of this movie from 1995 takes place in 2021. In a world of corporations, which control information and access to it, the safest data transistor is the human brain, which, with implants, can gain significant capacity. While the human mind has limited memory and capacity and is not a safe data transistor, in this movie such an enhanced brain is presented as the only safe solution for data storage. The main character, Johnny Mnemonic, has a brain capacity of 80 GB, but with hardware addition (memory doubler) the capacity is increased to 160 GB. A main plot driver of the movie falls into place when Johnny decides to upload 320 GB into his mind, which overloads his present mind. The data held within such an enhanced mind is protected with a picture password taken from random TV pictures, which are selected during the data transmission with a specially designed remote control. These photos can be printed, scanned, and faxed. Therefore, while we are dealing with technology that is still not yet available, this technology is transmitted over media channels available at the time (telephone, scanner, fax machine, etc.). In this movie, there is an interesting demonstration of cyberspace. When the main character, Johnny, connects to cyberspace via a dial-up, the cyberspace realm is presented through two-dimensional animation as a three-dimensional world.
In addition, another interesting data storage solution we can identify in this movie concerns documents, especially passports. A passport is presented as a data-card, which at border control is put in a special reader. The passport is designed as a regular document, but its usage is different. Today, a document and a passport scanner are similar technology.

Minority Report (2002)
The plot of the movie Minority Report (2002) takes place in 2054. Most of the movie's dynamics take place inside the Precrime Unit where visions of future crime are pre-seen by Precognitive persons (Precogs) inside a biopool. These crime visions are stored on a separate interface on plexy plates in the form of multidimensional images. These plates are then inserted in different portals dedicated to data manipulation and additional information and data search, which include the detective's activities, such as defining the surroundings (area and address) where a certain crime is going to take place. To perform these ac-

tions, the detective manipulates the video information on a hologrammatic interface using special data gloves.
The memories that form a life can also be stored on smaller glass boards in the form of a video. The data from these boards are projected as a hologram video.

The Avengers (2012)
The Avengers (2012) deals with the realization of the concept of S.H.I.E.L.D. (Strategic Hazard Intervention Espionage Logistics Directorate), which was mentioned in the movie Iron Man 3. In the scene where agent XX meets with Tony Stark, he gives him a special kind of two-sided laptop as a device for viewing secret data. Another interesting idea is realized in this movie scene and deals with data presentation. On the laptop, there are multidimensional data and folders, which are opened on several hologrammatic displays. We can see that these data are interactive because in one moment, Tony Stark can not only see, but also can accept the hologram's data. While this technology is not yet available today, we can witness the development of 3D printing and highly interactive devices (computers, mobile phones, etc.) that parallel this development.
Another interesting technological solution in this movie is the possibility of easy data transmission. In the movie, data can be transmitted from one computer to another with just a slide of the finger. This presents one possible solution for faster data transmission in the future.

Conclusion

While movies present a number of important genres of popular culture, the aim of this paper was to determine whether science fiction and fantasy movies in particular could predict technological solutions for future issues. In this paper main focus was detecting presentation of data storage technology in popular science fiction and fantasy films in last 40 years. For this purpose several science fiction and fantasy movies were randomly selected for the initial exploration of the link between data storage technology, film and projection of the future. By analysing several science fiction and fantasy movies—2001: A Space Odyssey (2001), Star Wars (1977), Superman (1978), Star Trek III: The Search for Spock (1984), Star Trek: Generations (1994), Johnny Mnemonic (1995), Minority Report (2002), and The Avengers (2012) the goal was to identify and explain how data storage devices were presented in fantasy and science fiction movies. In order to determine if these movie data storage solutions were realistic presentations of possible and realistic data storage solutions or just imaginable parts of a movie script. As shown by the analysis, each of the selected movies presented several revolutionary solutions for data storage devices. The initial analysis revealed that movies like 2001: A Space Odyssey (2001), Star Wars (1977), Superman (1978), and Star Trek III (1984) deal with technological devices for data storage, while, for example, more recent movies like Johnny Mnemonic (1995) use the human brain as a device for data transmission and

storage, while Minority Report (2002) uses human memories as a form of data that can be saved on a technological device. Several movies presented technology that wasn't available at the time, but was invented later on. For example, 2001: A Space Odyssey (1968) provided a large amount of innovative technological solutions like virtual (video) border control, voice ID (voice print identification), traveling in zero gravity, and video conversations (picture phone), tablets—all of which are inventions of contemporary time. Several movies presented technology that is still not yet available but probably will be in the near future. These include, for example, the hologrammatic messages in Star Wars (1977) and Superman (1978), or 3D demonstration of the data and easy data transmission throughout several platforms in The Avengers (2012). Interesting data were included in Star Trek III (1984) and Star Trek: Generations (1994). While Star Trek: Generations (1994) movie presented the first version of a USB device, Star Trek III (1984) had several traditional technological procedures for data storage, like video cassettes and photo albums. As regards any correlation between realistic technological developments and fantasy movie ideas, the analysis showed that while several movies really predicted some of the contemporary technological solutions for data storage, several made merely fictional usage of these kinds of devices. For example in Superman (1978), the data storage device was presented in the form of a crystal, but also it contained indescribable power that was both magical and unrealistic. In addition, Johnny Mnemonic (1995) presented the human brain as an upgraded device and in Minority Report (2002), memories are used as accessed and transferred data. The interesting data is that the newer the films is, the data storage technology is becoming smaller, more abstractedly and more interrogated with the human. For example, in Johnny Mnemonic (1995) and in in Minority Report (2002) we can see that data storage is connected with a human, part of his DNA. While in The Avengers (2012), we have interactive technological solutions for data storage. In contrast to older movies (Star Trek, Star Wars, etc.) which presented data storage technology as devices, in movies that are more recent we have projection of imaginary space for data storage. This paper presents initial idea for further empirical and theoretical research. For further research, the goal is to expand the sample and include more science fiction and fantasy films and to conduct a qualitative and quantitative data study. By this approach, it would be possible to provide empirical data for this concept and to make more solid conclusions about the predictions in the future. With this analysis, it is possible to detect that more recent science fiction and fantasy films project data storage devices as invisible cyber space, similar to cloud computing technology that is present today and will probably be more developed in the future. Steven Hrotic (2014) asserts that science fiction will do as it always has done, create new purposes, not for the purpose of bringing about or preparing us for a technological future, but to teach us to live in the technological present. In that context, mov-

ies are still a part of popular culture, but they are also representations of the human imagination and ideas.

In the context of the issues and further developments in the future (not only for data storage), one can conclude with the Fleetwood Mac song "Don't Stop":

> "*Don't stop thinking about tomorrow,*
> *Don't stop it'll soon be here,*
> *It'll be, better than before,*
> *Yesterday's gone, yesterday's gone"*
>
> *(Fleetwood Mac, 1968)*

References

Andrejevic, Mark; Hearn, Alison; Kennedy, Helen. Cultural studies of data mining: Introduction // European Journal of Cultural Studies. 18(2015), 4-5; 379–394

Barriga, Claudia; Shapiro, Michael; Fernandez, Marissa. Science Information in Fictional Movies: Effects of Context and Gender // Science Communication. 32(2010), 1; 3–24

Beer, David; Burrows, Roger. (2013). Popular Culture, Digital Archives and the New Social Life of Data // Theory, Culture & Society. 30(2013) 4; 47–71

Bryan Clark. From Punch Cards to Holograms – A Short History of Data Storage. (2015). http://www.makeuseof.com/tag/punch-cards-holograms-short-history-data-storage/, (Access date 1 July, 2017);

Chandler, Theresa. An alternative comprehensive final exam: The integrated paper // Teaching Sociology / Hudock, N1; Gallagher Warden, N2. (ed). Youngstown: The Family Journal, 2001, 3

Chung, Becky. 13 Movies That Explore The Future Of Technology. (2014). https://creators.vice.com/en_au/article/53wgp5/13-movies-that-explore-the-future-of-technology (Access date 12 October, 2017)

Communicator (Star Trek). (2017). https://en.wikipedia.org/wiki/Communicator_%28Star_Trek%29, (Access date 1 July, 2017)

Danesi, Marcel. Popular Culture: Introductory Perspectives: 3rd Edition. London: The Rowman and Littlefield Publishing Group, 2015

Dune (novel). (2017) https://en.wikipedia.org/wiki/Dune_(novel) (Access date 12 October, 2017)

Dune novels. (2009) http://www.dunenovels.com/ (Access date 12 October, 2017)

Erigha, Maryann. Do African Americans Direct Science Fiction or Blockbuster Franchise Movies? Race, Genre, and Contemporary Hollywood // Journal of Black Studies, 47(2016) 6; 550–569

Gadgets Fosfor. History of data storage. (2015). http://gadgets.fosfor.se/history-of-data-storage/. (Access date 28 June, 2017);

Gladding, Samuel T. Teaching family counseling through the use of fiction // Counselor Education and Supervision, 33(1994) 3; 191-200

Gladstein, Gerald A.; JoAnn C. Feldstein. Using film to increase counselor empathic experiences // Counselor education and supervision 23(1983)2; 125-131

Grindstaff, Laura. Culture and popular culture: A case for sociology // The ANNALS of the American Academy of Political and Social Science, 619(2008)1; 206-222

Hambly, Barbara. Time of the Dark. Integrated Media: New York, 1982.

Heinlein, Robert. Time Enough for Love. Ace Books: New York, 1973.

History of Data Storage Technology. (2016). http://www.zetta.net/about/blog/history-data-storage-technology, (Access date 28 June, 2017)

Hudock, Anthony; Gallagher Warden. Using Movies to Teach Family Systems Concepts // The Family Journal: Counseling and Therapy For Couples And Families, 9(2001)2; 116-121

IMDb. (2017) http://www.imdb.com/pressroom/?ref_=ft_pr (Access date 12 October, 2017)
Jarvie, Ian, Charles. Seeing through Movies // Phil.Soc.Sci, 8(1978); 374-397
Koch, Gary, and Colette T. Dollarhide. Using a popular film in counselor education: Good Will Hunting as a teaching tool // Counselor Education and Supervision 39(2000)3; 203-210
Machles, David. Electronic Media Storage // AAOHN Journal 47, 10(1999)10; 494-498
Maynard, Peter E. Teaching family therapy theory: Do something different // American Journal of Family Therapy, 24(1996)3; 195-205
Milner, Andrew. Sean Redmond. Introduction to the special issue on science fiction // Thesis Eleven, 131(2015); 3-11
Niven, Larry; Anderson, Poul; Ing, Dean. Man-Kzin Wars. Baen Publishing Enterprises: New York, 1998.
Simmons, Dan. Hyperion Cantos. Doubleday: New York, 1989.
Sterling, Bruce. Islands in the Net. New York: Ace Books, 1988.
Tan, Aik-Ling, Jennifer Ann Jocz, and Junqing Zhai. Spiderman and science: How students' perceptions of scientists are shaped by popular media // Public Understanding of Science 26(2017)5; 520-530
Van der Laan, J. M. Frankenstein as science fiction and fact // Bulletin of Science, Technology & Society, 30(2010)4; 298-304
Voller, Peter, and Steven Widdows. Feature films as text: A framework for classroom use // ELT Journal, 47(1993)4; 342-353

E-SCIENCE

Educating digital linguists for the digital transformation of EU business and society

Petra Bago
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
pbago@ffzg.hr

Nives Mikelić Preradović
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
nmikelic@ffzg.hr

Damir Boras
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
dboras@ffzg.hr

Nikola Ljubešić
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
nljubesi@ffzg.hr

Summary

The digital single market is one of the 10 priorities of the European Commission, recognizing the Internet and digital technologies as an opportunity for customers and businesses in the European Union to contribute to the economy, create new jobs, and enhance Europe's position as a world leader in the digital economy. In order to be able to transform to a digital economy, it is an imperative for the workforce to have the necessary digital skills and competences. Therefore, reskilling and upskilling the workforce enables the workers to develop updated task-specific skills that can contribute to the implementation of digital technologies and thereby help companies move forward in the digital era. Digital linguistics is a new interdisciplinary field of study at the crossroads between linguistics, information sciences, information technology and social sciences. Digital linguistics is not synonymous to computational linguistics or corpus linguistics, although certain skills and research methods my overlap between these disciplines. Although universities offer graduate level programs in complementary disciplines, currently no European university offers a program in the interdisciplinary field of digital linguistics. In this regard, we present a project participating in the development of crucial digital skills and

competences of future employees for the digital economy. DigiLing: Trans-European e-Learning Hub for Digital Linguistics *is a 3-year project funded by the Erasmus+ program of the European Union with the purpose of a) creating an internationally approved model curriculum for digital linguistics at the graduate level, b) training the teachers in relevant disciplines in the use of digital technology with the goal of designing high quality online learning materials, c) designing online courses for core modules and making them open and accessible to a broad network of stakeholders, the widest academic community, and the public at large, d) and disseminating and sustaining the results of the project. By identifying the key skills and competences that a contemporary study program of digital linguistics at the academic level should provide, and by developing and implementing a model curriculum for a digital linguist, the DigiLing project hopes to contribute to the European digital transformation.*

Key words: digital economy, digital single market, digital transformation, digital linguistics, model curriculum, DigiLing

Introduction

The European Commission, with its president Jean-Claude Juncker in the forefront, announced in May 2015 "A Digital Single Market Strategy for Europe". It is one of the 10 priorities of the Juncker Commission, recognizing the Internet and digital technologies as an opportunity for customers and businesses in the European Union to contribute to the economy, create new jobs, and enhance Europe's position as a world leader in the digital economy. The strategy is based on 3 pillars:

- better access for consumers to digital goods and services across Europe,
- creating the right conditions and a level playing field for digital networks and innovative services to flourish,
- maximising the growth potential of the digital economy.[1]

Recent studies have estimated that digitisation of products and services can add more than €110 billion of annual revenue in Europe in the next five years.[2] In order to be able to transform to a digital economy, it is an imperative for the workforce to have the necessary skills. Still in 2016 the EU had 14% of non-internet users, while 44% of the population had an insufficient level of digital

[1] The European Commision. *Communication from the Commision to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – A Digital Single Market Strategy for Europe.* The European Commission : Brussels 2015. 192 final. URL: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52015DC0192&from=EN. (8 June 2017)

[2] The European Commission. *Digitising European Industry*. URL: https://ec.europa.eu/digital-single-market/en/digitising-european-industry. (9 June 2017)

skills.[3] Therefore, reskilling and upskilling the workforce enables the workers to develop updated task-specific skills that can contribute to the implementation of digital technologies and thereby help companies move forward in the digital era.[4]

DigiLing

In this regard, we present a project participating in the development of crucial digital skills of future employees for the digital economy. *DigiLing: Trans-European e-Learning Hub for Digital Linguistics* is a 3-year project funded by the Erasmus+ program of the European Union with the purpose of a) creating an internationally approved model curriculum for digital linguistics at the graduate level, b) training the teachers in relevant disciplines in the use of digital technology with the goal of designing high quality online learning materials, c) designing online courses for core modules and making them open and accessible to a broad network of stakeholders, the widest academic community, and the public at large, d) and disseminating and sustaining the results of the project.

Digital linguistics is a new interdisciplinary field of study at the crossroads between linguistics, information sciences, information technology and social sciences. Digital linguistics is not synonymous to computational linguistics or corpus linguistics, although certain skills and research methods my overlap between these disciplines. Computational linguistics is concerned with modelling and processing of natural language from a computational perspective and the study of appropriate computational approaches to linguistic questions[5], while corpus linguistics is the study and analysis of natural language phenomena obtained from large collections of (machine-readable) texts of both written and spoken language[6]. However, the emerging field of digital linguistics is broader in the sense that it should provide the complete set of scientific, methodological, and practical foundations pertaining to communication in the digital age. This subsumes linguistic knowledge, such as native and foreign language competence, translation related competences and an understanding of language analysis procedures at all levels, and it also entails natural language processing (NLP) skills, particularly at the level of an in-depth understanding of state-of-the-art

[3] The European Commission. *Human Capital: Digital Inclusion and Skills*. *Europe's Digital Progress Report 2017*. URL: http://ec.europa.eu/newsroom/document.cfm?doc_id=44390. (9 June 2017)

[4] Probst, Laurent et al. *Digital Transformation Scoreboard 2017: Evidence of positive outcomes and current opportunities for EU businesses*. The European Commission. January 2017. URL: https://ec.europa.eu/docsroom/documents/21501/attachments/1/translations/en/renditions/native. (9 June 2017)

[5] https://en.wikipedia.org/wiki/Computational_linguistics

[6] https://en.wikipedia.org/wiki/Corpus_linguistics

NLP techniques and basic programming skills. But on top of this "traditional" interdisciplinary blend between linguistics, information sciences, and information technology, digital linguistics has several additional foci which justify its claim for a field of its own right.[7]

The first is digital content authoring, which could at first glance be understood as one of the linguistic competences formerly known as text production skills developed through text and discourse studies. But the digital age has brought profound changes to the ways texts and other types of content are produced. Digital news media have revolutionized journalism and brought new paradigms into the concepts of journalistic research, credibility, authenticity, authorship and accessibility. The personalization of digital services means that content, including web sites, ads, user manuals and posts on social platforms, is produced in a targeted and user-centered fashion, whereby the cyber-identity of the target user is not to be confused with their real-world identity.[8]

These issues can only be adequately addressed by bringing in the sociological, psychological, and cognitive perspectives, and by putting communicative behaviour in digital media into the centre of study. Another aspect of content authoring is related to multilingual contents and activities such as translation, localization, subtitling and interpreting. While traditionally the providers of multilingual services were the ones generating content, contemporary translators compose texts by selecting from available hits offered by translation memories, machine translation engines and other multilingual resources. From the cognitive point of view, as Pym (2013) points out, the process of [content] generation has been transformed into the process of selection, where the issue of critical assessment and trust has become paramount.[9]

The issues of trust, identity, authorship and reuse inevitably lead to questions concerning intellectual property rights and data protection, but also ethical aspects of communication in digital media. The legislative framework which attempts to regulate rights related to language data is lagging behind.[10]

Therefore, digital linguistics as a field of study combines insights and perspectives from different disciplines and does not overlap with computational linguistics, nor for that matter with digital humanities, sociolinguistics, corpus linguistics or machine translation, though it may inherit methods and tools from all of the above.[11]

[7] Vintar et al. (2017)

[8] Ibidem.

[9] Ibidem.

[10] Ibidem.

[11] Ibidem.

The survey of labor market needs was conducted from January through March 2017[12], receiving 81 responses from companies in eight different countries, with the majority coming from the five countries of partner institutions. The needs analyses among employers reveal important trends regarding textual content processing and multilingual communication amongst European enterprises. The results of the survey were used to identify the key skills and competences that a contemporary study programme at the academic level should provide in order for its graduates to be highly employable language professionals.[13]

Currently European universities offer a variety of graduate level programs[14] in the fields such as computational linguistics, applied linguistics, digital humanities, interpreting and translation, language and communication technologies, natural language processing, information sciences, computer sciences, digital media management, digital communications, social informatics etc. However, no university offers a graduate level program in the interdisciplinary field of digital linguistics. The goal of the project is to develop a programme which would combine linguistic- and translation-oriented subjects with technologies and language processing subjects, and add the cognitive, psychological and sociological knowledge to create a new graduate profile, that of a digital linguist.

DigiLing Objectives

DigiLing will bridge the gap between employers' needs and employees' skills through achieving the following objectives:

- Create an internationally approved model curriculum for digital linguistics at the graduate level,
- Train the teachers in relevant disciplines in the use of digital technology with the goal of designing high quality online learning materials,
- Design online courses for core modules,
- Disseminate and sustain the results of the project.

Create a model curriculum for digital linguistics

One of the objectives of DigiLing is to create an internationally approved model curriculum for digital linguistics at the graduate level by combining existing and new courses. To identify the necessary skills and competences a digital linguist should hold and to pinpoint gaps in existing curricula, a trans-European survey among employers and end-users was conducted. A highly skilled university graduate holding a master's degree in digital linguistics possesses knowledge and understanding about language and communication from several comple-

[12] http://www.digiling.eu/deliverables/

[13] Vintar et al. (2017)

[14] http://www.mastersportal.eu/

mentary disciplines. A master's graduate in digital linguistics should have the following skills and competences:

- Language competence in at least two languages,
- An understanding of the way written and spoken language works at all levels of linguistic analysis,
- An understanding of the principles of multilingual communication, including skills in intercultural mediation, translation, interpreting, localization and multilingual content authoring,
- Skills in the compilation of digital language resources, such as corpora, lexica, acoustic databases and similar, including competences in methodological design and technical implementation of LR compilation,
- Skills in analysing and processing natural language, including the ability to design and develop own tools as well as implement existing ones in order to analyse or process language data,
- Basic understanding of digital media from the sociological, psychological and legal perspective,
- Ability to perform independent research and acquire new skills,
- Ability to work in interdisciplinary/multilingual teams.[15]

Train the teachers in the use of digital technology
Another objective of the project is to train the teachers in relevant disciplines in the use of authoring tools and in the design of high quality online learning materials. This will be achieved through a face-to-face workshop for higher education staff in e-authoring.

Design online courses for core modules
An additional objective of DigiLing is to design online courses for selected modules covering many of the key topics in digital linguistics, as well as localizing, evaluating, testing, and implementing the courses. Courses will be designed in compliance with open e-learning standards, and will be accessible under the Creative Commons license. The content will be localized into all partner languages (except Czech) and International Sign language to facilitate inclusion into new national or joint study programs and to provide accessibility to special needs students. Courses will be cross-evaluated by partners (teachers and students), participants of the DigiLing summer school and national ECTS accreditation bodies.

Disseminate and sustain
The final objective of DigiLing is to disseminate and sustain the results of the project. DigiLing results will be publicized to a broad network of stakeholders,

[15] Vintar et al. (2017)

the widest academic community, and the public at large including special needs students. Short- and medium-run sustainability will be achieved by using existing and acknowledged platforms for our DigiLing hub: the international CLARIN network and the University of Ljubljana infrastructural centre of Language Resources and Technologies (CJVT). Medium- and long-run sustainability will be achieved through national accreditation of online courses and the curriculum.

Target audience

The project directly targets an estimated audience of 2000 people, with potential benefit for many more:

- Students of partner universities studying or planning to study at any language- or IT-related study programme, including (General or Applied) Linguistics, (General or Specialised) Translation, Intercultural Communication, Natural Language Engineering or Processing, Information Technologies, Information Sciences, Informatics, Computer Science and similar,
- Teachers and researchers of partner universities in the relevant fields,
- Companies, organisations, public institutions and other users of digital language services.

Conclusion

The European Commission announced in May 2015 "A Digital Single Market Strategy for Europe" as one of the 10 priorities of the Juncker Commission, recognizing the Internet and digital technologies as an opportunity for customers and businesses in the European Union to contribute to the economy, create new jobs, and enhance Europe's position as a world leader in the digital economy. In order to be able to transform to a digital economy, it is an imperative for the workforce to have the necessary digital skills and competences. Digital linguistics is a new interdisciplinary field of study at the crossroads between linguistics, information sciences, information technology and social sciences. Digital linguistics is not synonymous to computational linguistics or corpus linguistics, although certain skills and research methods my overlap between these disciplines. Although universities offer graduate level programs in complementary disciplines, currently no European university offers a program in the interdisciplinary field of digital linguistics. In this paper we presented a project participating in the development of crucial digital skills and competences of future employees for the digital economy. *DigiLing: Trans-European e-Learning Hub for Digital Linguistics* is a 3-year project funded by the Erasmus+ program of the European Union with the purpose of a) creating an internationally approved model curriculum for digital linguistics at the graduate level, b) training the teachers in relevant disciplines in the use of digital technology with the goal of

designing high quality online learning materials, c) designing online courses for core modules and making them open and accessible to a broad network of stakeholders, the widest academic community, and the public at large, d) and disseminating and sustaining the results of the project. By identifying the key skills and competences that a contemporary study program of digital linguistics at the academic level should provide, and by developing and implementing a model curriculum for a digital linguist, the DigiLing project hopes to contribute to the European digital transformation.

References

DigiLing. http://www.digiling.eu. (9 June 2017)

Probst, Laurent et al. *Digital Transformation Scoreboard 2017: Evidence of positive outcomes and current opportunities for EU businesses*. The European Commission. January 2017. URL: https://ec.europa.eu/docsroom/documents/21501/attachments/1/translations/en/renditions/native. (9 June 2017)

Pym, Anthony. (2013) Translation Skill-Sets in a Machine-Translation Age. Meta 583: 487– 503.

The European Commission. *Digitising European Industry*. URL: https://ec.europa.eu/digital-single-market/en/digitising-european-industry. (9 June 2017)

The European Commision. *Communication from the Commision to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions – A Digital Single Market Strategy for Europe.* The European Commission : Brussels 2015. 192 final. URL:http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52015DC0192&from=EN. (8 June 2017)

The European Commission. *Human Capital: Digital Inclusion and Skills*. *Europe's Digital Progress Report 2017*. URL: http://ec.europa.eu/newsroom/document.cfm?doc_id=44390. (9 June 2017)

Vintar, Špela et. al. *Labour market needs survey and the DigiLing model curriculum*. URL: http://www.digiling.eu/deliverables/. (9 June 2017)

Training, Consulting and Teaching for Sustainable Approach for developing Research Data Life-Cycle Management expertise in Switzerland

Basma Makhlouf-Shabou
Information Science Dept. // Geneva School of Business Administration,
University of Applied Sciences and Arts Western Switzerland
Rue de la Tambourine 17, Bât. B, 1227 Carouge, Switzerland
basma.makhlouf-shabou@hesge.ch

Summary

Developing research in different fields is defendable as well as necessary for the development of disciplines and knowledge construction in general. The progress of qualitative and quantitative approaches is based on hard competitiveness and high level of innovation. This increases the need of a rigorous management of research process which should be more and more accurate and traceable to ensure a good data management approach. Considering this context, European Council stipulates a directive to require a good research data management in order to reinforce the ability of researchers to conduct properly their research activities (European Commission, 2016). For example the H2020 projects requires a Data Management Plan (DMP) since January 2017. In Switzerland, this tendency was clearly confirmed. Swiss researchers have been submitting their proposals to funding agencies without any requirement for research data management so far. However, the Swiss National Science Foundation will require a DMP since October 2017. Researchers are not prepared. They don't know how it impacts their work and are looking for solutions to comply with these new requirements. This paper draws a general portrait of a recent Swiss project on this subject: data life cycle management applied on research data: DLCM[1]. It presents, first, an over view of the main objectives and major dimensions of DLCM project and second, it will focus on one those latest which is dealing with training, consulting and teaching in the field of research data management.

Key words: research data, national services, research data life cycle management, data governance, research data training, research data consulting

[1] https://www.dlcm.ch/

Introduction

We understand the notion of research data governance as a global and exhaustive management of data arising from the research process in order to guarantee its use, security and optimization, from its creation and capture to its disposal (UK Data Archive, 2017). In the Swiss context, the research data management should be studied considering two levels: cantonal and federal. Regarding the cantonal level, public universities and high schools are mainly managed by the 26 cantonal government. The federal level concerns especially the seven federal administration's departments and is not synonymous of national level, under which the federal institutes of technology or "polytechnics" are supervised. Added to those main two levels, we have thus a third-level: an inter-cantonal level that proposes a network of high schools in different cantons of Switzerland as well as the University of Applied Sciences and Arts of Western Switzerland. The data management issue affects all levels (cantonal, inter-cantonal and federal), and also the public and private sectors, like academic institutions or private laboratories. Despite this particular fragmented environment, researchers needs remain generally comparable. The issues related to the various aspects of research data processing remain the same (DMP designing, security of sensitive data, long-term preservation, etc.). In addition, some technical solutions are very expensive and require a specialized expertise. In this context, the idea of national services was argued and some initiatives were encouraged in order to develop national solutions to enhance Swiss researchers developing such expertise and share their resources to comply with standards and good practices as well. In this same context, openness and sharing research data has become essential to obtaining funding for research projects. The DLCM Project is one of those initiatives which aiming to provide scientific communities a research data national services. It started at the end of 2015, and is funded by the RCSU (Rectors'Conference of the Swiss Universities) and is conducted under the leadership of the University of Geneva, especially the Information system service. It should finish in 2018. The following sections will first specify the objectives of the project after a brief presentation researcher's needs. Second, the dimension of training, consulting and teaching in the field of research data management will be presented.

Initial exploration of the needs

As mentioned, the research need examination was the starting point for the project to be able to design accurately its main deliverables. To do so, an explorative approach was adopted. Based on a series of semi-structured interviews with some 50 researchers and heads of research departments working in some thirty disciplines at six institutions, we were able to identify mainly four ranges of needs.

1. The first are guidelines and tools. There is a lack of practical guidance. When we talk about guidelines, policies or management programs, we mainly think of tools such as the DMP (Data Management Plan), the tools of analysis, coding of the information or the collected data, and all that is related to the preservation and publication. This includes also, research data governance policies.
2. The second is all that concerns the processing of information in the first phase of the research process, before reaching the preservation and the disposal. It refers to processes, tools, methods and best practices needed to active data management.
3. The third is about the management of the collected data, the results and the publications, including the reuse and sharing of research data, in particular it's various modes and channels.
4. Finally, the qualified staff, which has a dedicated expertise, developed especially in the management of research data, also represents something that is missing. Generally, the universities do not have an important staff of archivists or records managers. And the focal point into universities became the librarian who has certain knowledge of this domain. However, this later remains not sufficiently prepared to assume this role because his professional profile as a librarian does not allow him to provide appropriate and advanced expertise needed or requested by researchers.

Objectives and deliverables

Based on need analysis, the project objectives are focused around five fields:

1. Guidelines & Policies – key tools commonly used and recognised at an international level;
2. Active Data Management – everything linked with the records management;
3. Publication & Preservation – the perpetuation and the publication for a future reuse;
4. Consulting and Training – the perpetuation and the knowledge transfer in this field;
5. Outreach & Dissemination – the publication and the promotion of everything that will be developed in the context of the project.

To do so, the implementation of DLCM project was articulated on 6 main tracks (dimensions) as shown in Figure 1: 0) project management; 1) guidelines and policies, 2) active data management; 3) publication and preservation (Burgi, Blumer and Jelicic, 2017); 4) training and consulting; 5) dissemination publication (Makhlouf Shabou, Burgi, Blumer and Echernier, 2016).

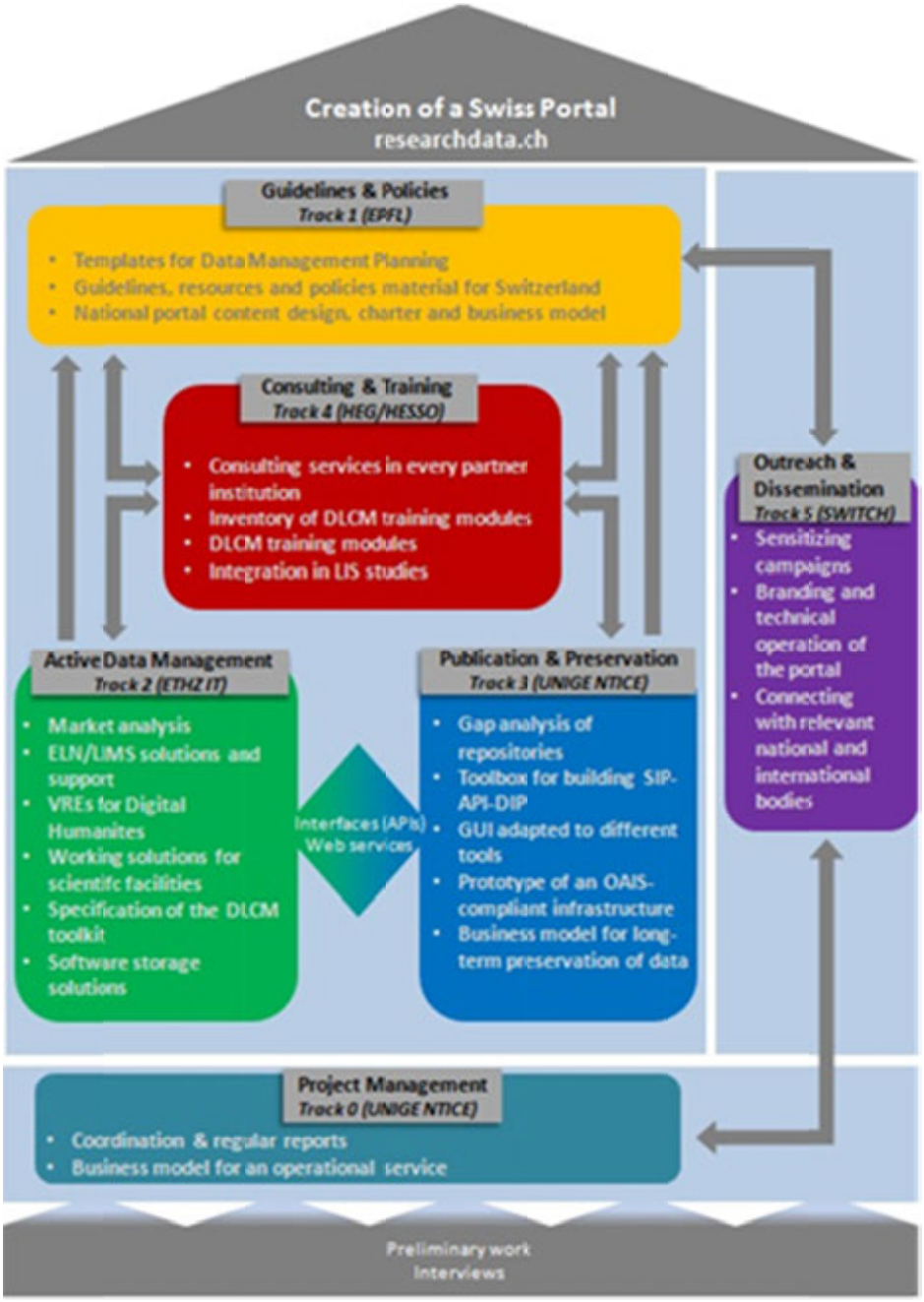


Figure 1: DLCM organisation in 6 tracks

In the following section, we will present the track 4: one central dimension dealing consulting, training and teaching in the field of research data management (Figure 2).

Consulting, training and teaching for the knowledge transfer in research data management

The consulting, training and teaching dimension aims to facilitate the creation, centralization and exchange of know-how in the management of research data in each partner institution first and spread this knowledge at national and international level. It is divided into three sets of activities: 4.1) development of advisory and support services in each partner institution (*consulting*), 4.2) collection and development of training modules (*training*) and 4.3) integration of teaching modules in the curriculum in information sciences (*teaching*). Implementation of these activities is carried out in close collaboration with other dimensions and partner institutions of the DLCM project (Figure 2).

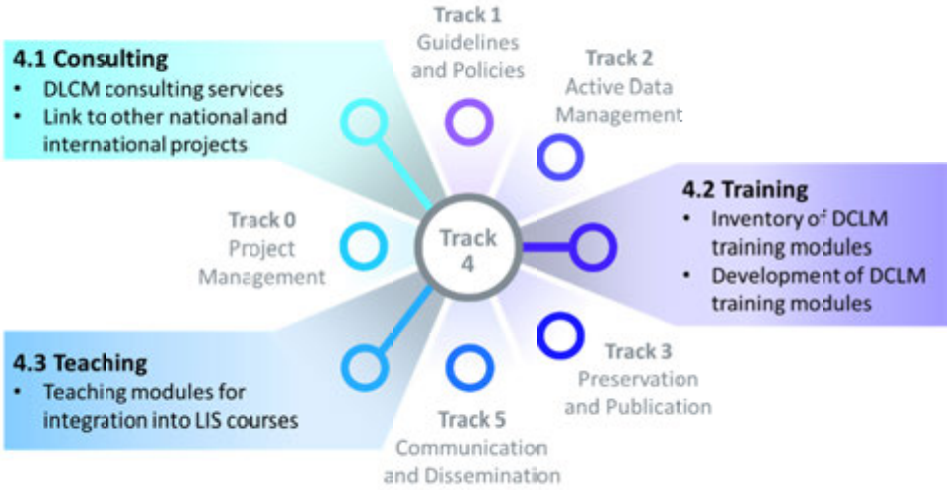


Figure 2: Objectives & deliverables in Track 4 of DLCM Project

Consulting

For this theme, we have planned two deliverables: (i) Implementation of consulting services and (ii) coordination between project partners to pool efforts and synergies of similar projects.

The first period focuses mainly on consulting activities, which will be offered to partner institutions and their respective users. The objective is to develop a clearinghouse at the national level using the "DICE+ model" (Digital Copyright in Education) and to offer workshops along with a hub for field expertise. To assume the relevant impact of those two axes (training and consulting), the track

4 of Project worked on a coordination desk of to guarantee the accessibility and the visibility of those services at national level. Consulting services will be established at the institutional and inter-institutional levels respectively, and will cover all the major steps of DLCM, including processes, tasks and tools. Finally, all these services will be coordinated at interinstitutional level by a central office: Coordination Desk. A pilot of coordination desk is established into the University of Applied Sciences and Arts of Western Switzerland (HES SO). This offers a reel research data national services presence through an extended academic network.

The objective of the DLCM coordination desk is to offer a unique point of contact to all academic institutions for guidance and support on research data management practices. It would do so by having a resource person and teams of focal points and subject-matter experts work together to answer requests from the Swiss research community. The coordination desk will also provide a strong instrument for mobilizing and coordinating the resources needed to achieve this objective. The primary target groups for the coordination desk are researchers and research teams, project managers and administrators, teachers and assistants, as well as students. Secondary target groups are information management professionals and consultants, including librarians, archivists, records managers, IT specialists, etc.

Training

By training, we mean learning process that allows an individual to acquire the knowledge and skills needed for the professional activity (Swiss DLCM Project, 2017). These training activities, which are intended primarily for users of partner institutes, will be developed with the objective of ensuring an adequate transfer of knowledge and a mastery of the skills and expertise related to the management of research data in the different institutions, irrespective of their size or the complexity of their structures (Makhlouf Shabou & Echernier, 2016). As shown in Figure 2, the training proposes two deliverables: (i) an inventory of existing training modules; and (ii) the creation of research data training modules.

Two types of training modules are planned: general and advanced. The first, targets to introduce the fields of research data management to beginner scientific communities to coach and help them assuming the basic activities in research data processing. Those introductive modules cover as well principles and method needed for complying high data quality requirements. The advanced modules will address aspects of research data management that have not been covered so far and will be tailored to the specific needs of partner institutions and their respective users.

Added to conception and realization of training modules, two catalogues will be developed: one for research data existing training modules and another for re-

search data trainers and this on the DLCM project partners and also at the international level.

Teaching

The third and last set is teaching that refers to academic programs and courses that could be proposed to train-the-trainer who will be future professionals and specialists of research data management (Doctoral seminars, Masters, Bachelors, etc.) (Swiss DLCM Project. 2017). It aims to integrating the knowledge and skills of this project into the master's programs in information sciences. Initially, we propose the promotion of DLCM field to ensure their adoption and rapid application among future professionals, and secondly awareness of students of information science curricula in order to ensure the sustainability of the expertise and the transfer of knowledge for future generations. Since September 2017, the Information Sciences Department of Geneva School of Business Administration of University of Applied Sciences and Arts Western Switzerland offers to the Master of Information Science students a specialization in Data governance. This specialization includes 16 periods (hours) devoted to research data. This was the first implementation step of teaching deliverables. In addition, Master students in this Department have to propose a DMP for their proper Master research. Other courses related to research data are planned to be added into the same programme during the following years.

Conclusion

The DLCM project creates a real opportunity for developing sustainable and tangible solutions at a national level to implement research data life-cycle management (DLCM) in Switzerland. As mentioned the funding agencies are requiring a research data management ability to ensure the appropriate research data governance of the whole life cycle of researches and also to increase the exploitability of research outcomes. The DLCM offer to researchers the opportunity to exchange valuable materials, tools, methods, infrastructure to be able to acquire the needed knowledge for managing research data from the creation to disposition and the preservation of those data. It specifies the way to perform the capture and creation processes, the techniques and methods to structure and share them, and the requirements of comply standards and good practices preservation. This will indeed reinforce the quality of the work of researchers, since it allows the traceability and the documentation of the research processing.

References

Blumer, Eliane; Burgi, Pierre-Yves. Data Life-Cycle Management Project: SUC P2 2015-2018. 17th December 2015. http://www.ressi.ch/num16/article_110 (17th May 2017)

Burgi, Pierre Yves; Blumer, Eliane; Jelicic, André. Research DLCM: Creating the framework for a successful research data management in Switzerland. 13th September 2016. https://prezi.com/jwd2ilsrwruc/research-dlcm/ (19th May 2017)

Makhlouf Shabou, Basma; Burgi, Pierre Yves; Blumer, Eliane; Echernier, Lydie. Le projet DLCM : une approche globale pour une meilleure gouvernance du cycle de vie des données de la recherche en Suisse. Tunis, Colloque international sur les bibliothèques et archives à l'ère des Humanités numériques (CIBAHN), 19 octobre 2016.

Makhlouf Shabou, Basma; Echernier, Lydie. La gouvernance du cycle de vie des données de la recherche en Suisse : transfert & pérennisation. Montpellier, Journée de la section Aurore "Archiver la recherche : responsabilités partagées", 23 juin 2016.

European Commission - Directorate-General for Research & Innovation. H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020. 26th July 2016. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (19th May 2017)

International Organization for Standardization. *Space data and information transfer systems - Open archival information system (OAIS) - Reference model*. Geneva, ISO 14721, 2012.

Ray, Joyce M. (ed.). Research Data Management: Practical Strategies for Information Professionals. West Lafayette, Purdue University Press, 2014

Swiss DLCM Project. 2017. https://www.dlcm.ch/ (19th May 2017)

UK Data Archive. Create & manage data: research data lifecycle. 2002-2017. http://www.data-archive.ac.uk/create-manage/life-cycle (19th May 2017)

DIGITISATION, RECORDS MANAGEMENT AND DIGITAL PRESERVATION

Original scientific paper

A Model for Long-term Preservation of Digital Signature Validity: TrustChain

Vladimir Bralić
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
bralic.vladimir@gmail.com

Magdalena Kuleš
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
magdalenakules@gmail.com

Hrvoje Stančić
Department of Information and Communication Sciences
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
hstancic@ffzg.hr

Summary

When archiving a digitally signed document an issue arises once the certificate used in the signature expires (or possibly the certificate authority stops functioning). Once this happens, the signature can no longer be confirmed and tampering with the document is possible. This paper presents a model for long-term preservation of digitally signed documents using blockchain technology. The authors propose a semi-open system in which only certain institutions can create new entries but any interested party can view the records and confirm their authenticity.

Key words: digital signature, blockchain, archive, certificate authority, long-term preservation, TrustChain, TRUSTER

1. Introduction

Digitally signed documents are now seeing widespread use in almost any online, digital document storage, management and preservation systems. They are replacing the traditionally sealed and signed documents. While the digital signature system is now well-documented, safe, and easy to implement, an issue arises with the outdated certificates. For better understanding, it is necessary to define what a digital signature is. According to one definition, digital signature

is "a code, generally created using a public key infrastructure (PKI) associated with a digital object that can verify the object has not been altered and, in some contexts, may be used to authenticate the identity of the sender"[1], while the other defines digital signature as "cryptographic transformation of data which, when associated with a data unit, provides the services of origin authentication and data integrity and may support signer non-repudiation"[2]. Digital signatures have two aspects. Firstly, they guarantee the integrity of a document. This means they guarantee the document contents match that at the time of digital signing. Secondly, they guarantee authenticity[3]. A digital signature can be traced back to a specific person or institution using a certificate authority. In this manner, one can be sure that the document was created by the stated author, and that the document could be used as a legally binding contract. The first aspect is time independent (one can always confirm the document in question has the appropriate signature by recalculating the hash of the document and comparing it to the one in the digital signature. However, the second aspect is time dependent. Most digital certificates expire, and unless renewed by their owners, i.e. document creators, cannot be confirmed. They are also reliant on certificate authority institutions still being in operation. It is certainly conceivable that some certificate authorities might go out of business or close down at some point in the future. Once this happens, it will be impossible to confirm that their certificates are genuine. Currently, most archives depend on trust to confirm outdated digital signatures. One has to trust the archive (or another institution) which preserves the document that the signature was valid at the time of archiving and that the document has not been tampered with. Another common solution is to use a time stamp service, which significantly extends the lifetime of a signature but, much like the digital signature itself, is not a permanent solution. To improve this situation we propose a system based on the blockchain technology[4] that might eliminate the need to trust archiving institutions by storing control hashes of digital signatures in an immutable and publicly readable blockchain. By using such a system, any interested party could confirm that a digitally signed and archived document has indeed remained unchanged and

[1] A Glossary of Archival and Records Terminology, Society of American Archivists, https://www2.archivists.org/glossary/terms/d/digital-signature (Accessed 30.05.2017).

[2] InterPARES Trust Terminology Database, http://arstweb.clayton.edu/interlex/en/term.php?term=digital%20signature (Accessed 30.05.2017).

[3] ISO 15489 defines that "an authentic record is one that can be proven: a) to be what it purports to be, b) to have been created or sent by the person purported to have created or sent it, and c) to have been created or sent at the time purported", ISO 15489-1 Information and Documentation – Part 1: General, p. 7.

[4] Blockchain – an open-source technology that supports trusted, immutable records of transactions stored in publicly accessible, decentralized, distributed, automated ledgers. InterPARES Trust Terminology Database, http://arstweb.clayton.edu/interlex/index.php (Accessed 20.04.2017).

that its signature was valid at the time of the blockchain record creation. We call this system TrustChain. TrustChain is being developed as part of the TRUSTER Preservation Model (EU31) research study at the InterPARES Trust international project and is one of several solutions to the problem of long-term preservation of digitally signed documents being considered by the research group.

2. The TrustChain concept

The system we propose is based on cooperation between multiple archival (or other interested) institutions. While there is no technical reason why a single institution could not run the needed software and hardware components, the trust in the envisioned system is in direct relation to the number of independent participating institutions. If a single institution runs the whole system, that institution is capable of manipulating records and would need to be trusted implicitly. This is the situation we have today. We are bypassing this need to trust a single institution by requiring multiple institutions to confirm the validity of a digitally signed document before writing it into an immutable blockchain. In principle, our approach uses a blockchain as described in the Bitcoin Whitepaper (Nakamoto, 2008) but we do not include the proof-of-work concept so our solution is perhaps more similar to the original Haber and Stornetta timestamp linking and random-witness solutions (Haber and Stornetta, 1991). We have merged both approaches and designed the system for use specifically for preserving (timestamping) digital signatures by a trusted union of (archival) institutions.

The core of the system is a blockchain containing hashes of digital signatures. Any interested individual or institution can request a record to be added to the blockchain but only the authorised nodes are allowed to write the new record into the blockchain (after confirming validity of digital signature(s)).

TrustChain nodes are servers maintained by institutions participating in the TrustChain project. These servers accept new record requests, process them, write them into the chain and keep the blockchain stored and available to be read by interested parties. Communication between a party requesting a new record to be added and nodes can be achieved via a specialized TrustChain client software or a web interface provided by the nodes themselves. Similarly, a party interested in confirming the validity of a document with an expired signature would contact a node, read the blockchain, find the relevant entry and compare it to the document that needs signature conformation. Finding the relevant block in the blockchain would be achieved by an indexing system that relies on the document metadata stored in the blockchain. This indexing system might be part of the TrustChain nodes or it might be outside of the system (since the blockchain is freely readable). The basic architecture of the TrustChain system is shown in Figure 1.

While TrustChain cannot extend the life span of a digital certificate, it would provide a guarantee that the document and its signature have remained unchanged since the TrustChain entry was created. Since the digital signature

contains the name of its owner, this can be used to confirm the creator of the document at a later date.

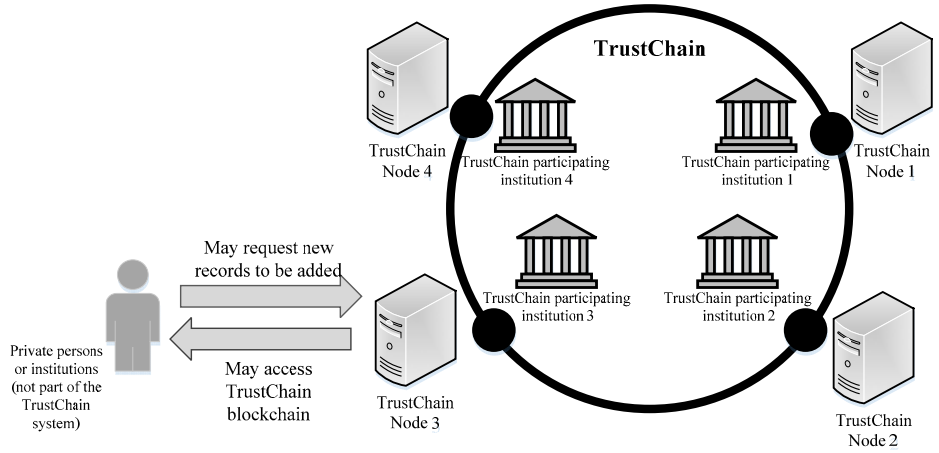


Figure 1. Basic concept of the TrustChain model

3. The TrustChain model processes

The process of adding a document to the TrustChain begins with the interested party selecting a digitally signed document that needs to be preserved. The digital signature of this document needs to be validated by the relevant certification authority. This check is performed at this point as well as later (by multiple TrustChain nodes). The document itself is to be stored outside of the TrustChain system, as the TrustChain stores only a control hash of the digitally signed document. If the full documents were to be stored in the blockchain, it would increase the blockchain to an unmanageable size quickly. It would be possible to build the TrustChain on top of a dedicated database system, similar to BigChainDB (McConaghy et al, 2017), but this is not our intention at this time. As it is, TrustChain is a system that complements other digital document and records management systems, digital archives, or repository systems and does not replace them.

The software preparing the TrustChain record calculates the hash, which is stored in the TrustChain system. As stated earlier, this can be a standalone application that communicates with the TrustChain nodes' API or a web service provided by the nodes. A link to the document (stored in an outside service), a timestamp and any relevant metadata (entered by the user) are added to the hash and a TrustChain record is formed. The record can then be forwarded to a TrustChain node for inclusion into the blockchain. This process is shown in Diagram 1.

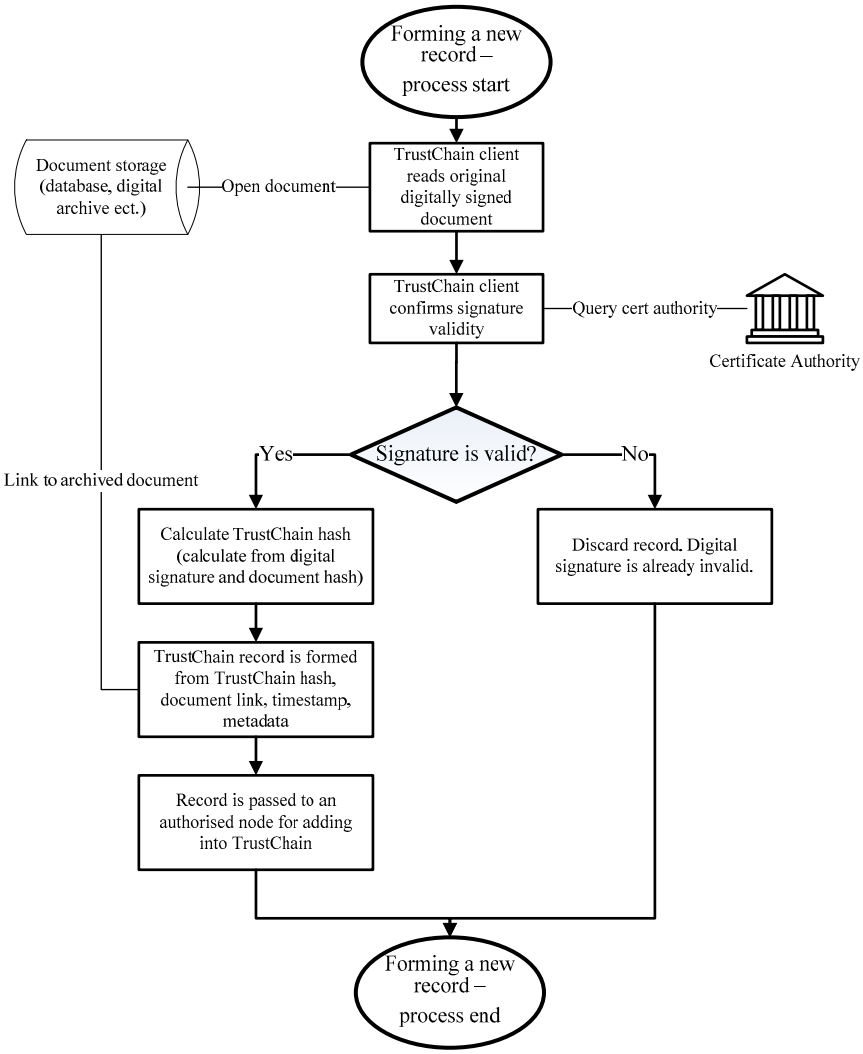


Diagram 1. Adding a document record to the TrustChain

The node will check signature validity and create a new record to be added to a joint record que (which will later form a new TrustChain block). Signature validity is to be confirmed automatically by the node accepting the block. The exact process by which this is achieved will depend on the format of the document (file) and its digital signature. Which documents and signatures can be checked will depend largely on the participating institutions and their requirements. Generally, this step involves checking that the document has not been tampered with after signing (by recalculating the document hash and decrypting the sig-

nature hash) and sending the certificate to the certificate authority for validation. Describing this process in detail is beyond the scope of this paper (this will be addressed at a later stage in the development) but there are many existing industry solutions to this problem. For example, a very common digital signature is the x.509[5] and checking the validity of such a signature is easily achieved by using .NET System.Security.Cryptography.X509Certificates class[6] or Java java.security.cert class[7]. Since the proposed blockchain solution doesn't depend on or store the document or the signature and only interacts with them while checking signature validity this part of the system is independent and we except it to grow and change to be able to accommodate future file formats or digital signature types.

It should be noted that the system makes no effort to eliminate documents signed by the compromised certificates. This is out of scope for TrustChain. The security of a certificate is obviously a responsibility of the certificate owner and his or her certificate authority. The best the TrustChain can do is to make use of protocols such as the Online Certificate Status Protocol[8] to identify the revoked certificates.

The process of adding records to a block and writing that block into the blockchain is left exclusively to TrustChain nodes (ran by trusted providers, most likely archival institutions that are members of the TrustChain system). At this stage of development, the nodes act in a round robin system. Once a node comes to its turn it collects new (candidate) records from a queue and attempts to validate all signatures. If a signature fails, the record is discarded as invalid and new records are collected. Once a sufficient amount of valid records is found, they are added to a block. We still do not add this block to the blockchain. Before this happens, we also require a certain number of other nodes to confirm signature validity of all records. The required number depends on the total number of available TrustChain nodes and the required level of reliability (the more nodes rechecking the records, the more reliable the vote will be). Since the number of participating institutions is not known, at this, early stage we will assume that all participating institutions maintain a node and they all vote on every block. Should the number of institutions rise to a number where having everyone vote becomes a performance issue, a smaller, randomly selected, subset of nodes can vote for each block. This subset should change for every block. If the majority of voting nodes agree that the block is valid it can

[5] IETF RFC 5280, https://tools.ietf.org/html/rfc5280 (Accessed 29.9.2017).

[6] MSDN Library, .NET Development, Framework Class Library, System.Security.Cryptography. X509Certificates Namespace (Accessed 29.9.2017).

[7] Java™ Platform Standard Ed. 7 Online documentation, Package java.security.cert. (Accessed 29.9.2017).

[8] ITEF RFC 6960, https://tools.ietf.org/html/rfc6960 (Accessed 29.9.2017).

be added to the blockchain (after having its hash calculated from its contents and the previous block's hash). Otherwise, the block is discarded and the records that formed it are returned to the new records queue (Diagram 2).

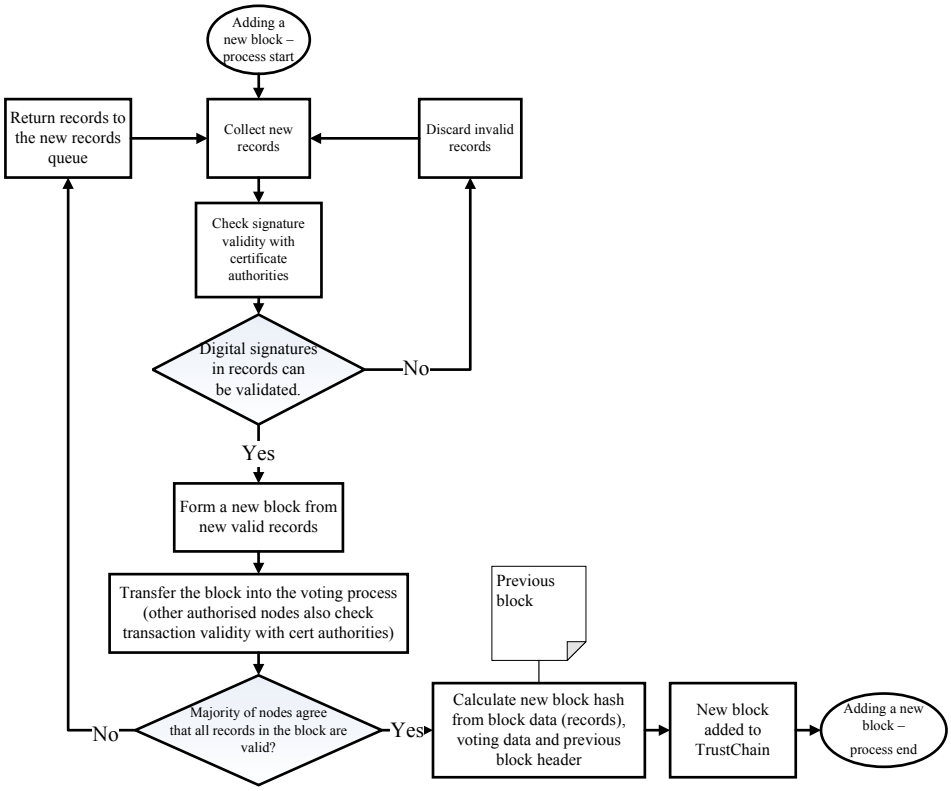


Diagram 2. Adding a new block to the TrustChain

The confirmation process of the (expired) digital signatures begins with finding the relevant records in the TrustChain blockchain. For this, the TrustChain relies on the recorded document metadata. Special services that allow searching the blockchain will need to be built as part of a standalone TrustChain clients or web services. Since the blockchain is written in pure text form, it is also possible to download it and search it without a specialized tool but this might prove troublesome for some users.

Once the relevant record is identified, all that needs to be done is to recalculate the hash from the original document and compare it to the one written in the TrustChain. If these hashes match, one can reliably claim that the document and its signature have remained unchanged since the date indicated by the blockchain record timestamp (Diagram 3).

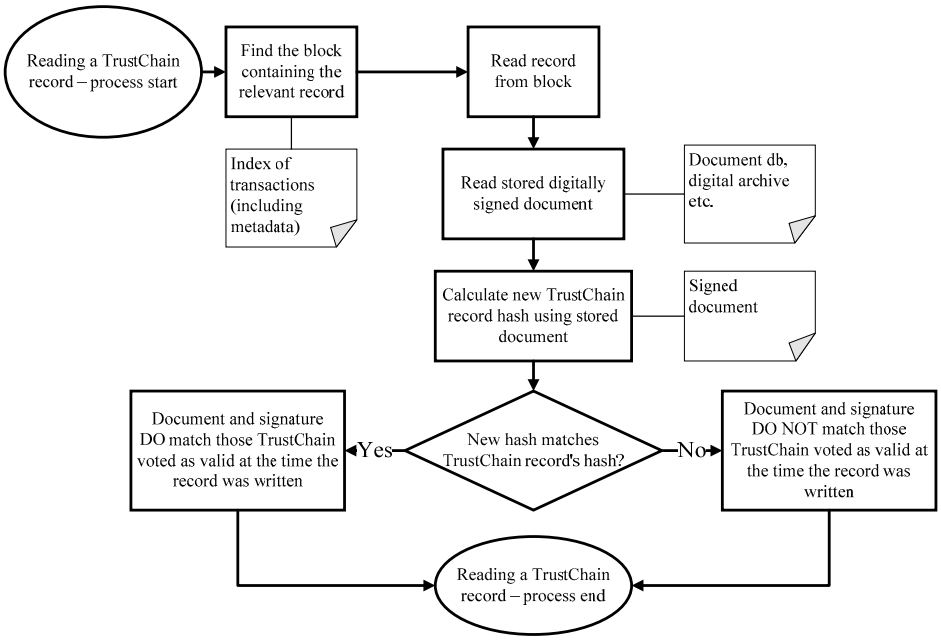


Diagram 3. Reading a record from the TrustChain

4. The TrustChain model's record and blockchain format

This chapter describes the core data structure of the TrustChain system – its blockchain. At this time, the TrustChain blockchain is designed as a plain text file with its contents written in JSON format. We decided on the JSON format because of its widespread use, Internet service-oriented design and human readability (IETF RFC 7159, 2014). Like all blockchains, the solution proposed here is comprised of multiple blocks that form an immutable chain by including a hash which is calculated from the current and the preceding block. These blocks can be further broken down into three sections: header, records and votes. These sections are further described in order of their creation during the process of writing new data into the blockchain.

The record section is at the beginning. Each record contains information about a single digitally signed document and its hash. Document hashing algorithm to be used is most likely SHA256, but it could change over time. This hash, along with a document link and any relevant metadata, are the most important parts of a TrustChain record. Record structure is presented in the following code.

```
{
"id": "<record hash to be used as uID>",
"version": "<model version>",
"record": {
      "timestamp": "<transaction timestamp>",
      "certAuthName": "<cert auth name>",
      "certAuthID": "<cert auth id, if available>",
      "certAuthApiLink": "<cert service link>",
      "data": {
            "TrustChainHash": "<document cert hash>",
            "docLnk": "<document link>" },
      "metadata": {
            "docRefCode": "<document reference code>",
            "docTitle": "<title of issued document>",
            "docCreator": "<name of document creator>",
            "docCreationDate":"<document creation date>" }
         }
}
```

Most of the fields are self-explanatory but few needs to be clarified. The "version" key refers to the version of the data model used to create this record. It is certainly conceivable that the format of the record will change over time and it is also possible to have standalone clients which might not have been updated regularly. Because of this, different versions of the records might appear in the same block and that is why the record data model version needs to be recorded on a record level.

The metadata subsection, which relies on the ISAD(G)'s essential set of elements[9], needs to contain all the information necessary to index and later find the document record in the blockchain. On the other hand, it would not be wise to overburden the blockchain with unnecessary information or fields that will often be left blank. This section requires fine balance and is still under review by the group.

The metadata section might also contain information pertaining to the archival bond (Lemieux and Sporny, 2017). While the primary purpose of the TrustChain is not to store complete documents (or records) and their metadata, since this is clearly a task for an external storage or archival solution, it would certainly add to their functionality. It could be used to add archival bond information to storage systems which otherwise might lack such features. In their paper, Lemieux and Sporny propose the use the OP_RETURN field of the Bitcoin transactions to store archival bond metadata. Since we propose a specialized system with its own blockchain, an implementation of the archival bond syntax could be significantly simpler. Aligning to the original proposal, a

[9] ISAD(G), General International Standard Archival Description, Second Edition, ICA, Ottawa 2000, p. 9, http://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_EN.pdf (Accessed: 30.06.2017).

subfield of the metadata section could be implemented which can contain the needed information and the additional signatures and links to the original transaction which a Bitcoin OP_RETURN field implementation requires can be omitted. The only data needed to be stored would be the ontology used and whatever specific fields it requires to insure the archival bond remains unbroken. An archival bond subfield can be added to the metadata information and any document that requires an archival bond can make use of it (implementing an appropriate ontology).
Multiple record subsections form the records section of a block. Following this section is the votes section. This section is again comprised of multiple vote subsections. The whole section is formed at the time of block creation by the originating node, but nodes that are indicated as voting nodes enter their vote in the "is_block_valid" field. The structure of this section is presented in the following code.

```
{
"nodeID": "<unique id of the node voting>",
"nodeSig": "<signature of voting node, hash of vote data>",
"vote": {
      "blockCandidate": "<id of the voted block>",
      "is_block_valid": "<true | false >",
      "timestamp": "<voting timestamp>"}
}
```

The voting system is based on a simplified version of the system used in the BigChainDB system (McConaghy et al, 2017). Once the originating node receives the responses from voting nodes (filled out vote fields in its block), it confirms the votes as valid by checking the voting node signatures. This method is under review by the group as it adds a public-private key element into TrustChain whose purpose is to avoid reliance on such systems. However, since the signing occurs in the voting section it is not needed to reconfirm it at a later date to validate document record included in the block. It is also easy to maintain certificates for the participating and previously participating nodes using the TrustChain infrastructure itself (in this case the TrustChain acts as a certificate authority for identification and authentication of the voting nodes). This might be an acceptable compromise. Since adding blocks to the TrustChain is a closed system, another possible method of insuring vote safety would be to skip this field altogether and implicitly trust the voting node responses. This would require security measures at the network and system levels to guarantee that the votes originated from the nodes in the TrustChain network, and have not been changed. Describing such a solution is beyond the scope of this paper but remains as a possibility during further development of the system.
Once the votes subsection has been filled in with sufficient votes confirming the block's validity, the final hash of the block is calculated and the block is written

into the blockchain. The hash of the block is calculated from the entirety of the current block and the header of the previous block (its own hash and id). This is presented in the following code.

```
{
"blockHash": "<hash of block>",
"blockID": "<block order number>",
"block": {
   "previousBlockHash": "<hash of previous block>",
   "timestamp": "<timestamp of block creation>",
   "nodeID": "<unique id of the node creating the block>",
   "records": [
      { <record1> }
      { <record2> }],
   "votes": [
      { <vote1> }
      { <vote2> }],
         }
}
```

Finally, the complete architecture of the TurstChain blockchain is illustrated in the Figure 2.
As stated, this blockchain would be freely available for downloading to any interested party but only the TrustChain members (authorised nodes) would be allowed to write new data. They may do so to fulfil their own archiving requirements or at the request of any person or institution which needs such a (trusted) service.

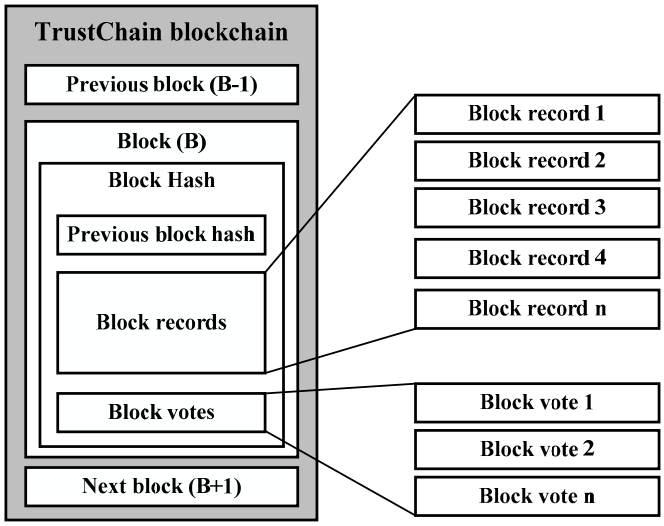


Figure 2. The TrustChain blockchain architecture

5. Existing standards for long-term preservation of digitally signed documents – time stamping

Long-term preservation of digitally signed documents is not a new concern. The first commercial application of an RSA digital signature has made an appearance in Lotus Notes[10], almost three decades ago. One way to address out-of-date certificates is to digitally time stamp the signed documents. According to Volarević and Stančić[11] this process is standardised by the international ISO/IEC 18014 and the ETSI 319 422 standard and, while it does provide a solution, it requires the document to be periodically restamped. On the other hand, the model proposed here attempts to eliminate this requirement. The technical aspect of the ETSI standard is based on the RFC 3161 – Time-Stamp Protocol (TSP). This RFC describes the process of time stamping a document by adding a trusted date and time to the document that are protected by a public key of the time stamp issuer. Since the time stamping process is reliant on the trusted time stamp providers, which use their own private and public keys to prove the document was indeed untouched since the time of stamping, they suffer from problems similar to the digital signature itself. They can be used to extend the lifetime of a digital signature but are not a permanent solution. The RFC itself makes note of this in chapter 4: Security Considerations[12].

Point 2 of chapter 4 states: "When the TSA[13] private key has been compromised, then the corresponding certificate SHALL be revoked". A revoked certificate invalidates all the existing time stamps (or tokens). In this event the RFC does not provide a reliable way of distinguishing between the time stamp tokens which are valid from the ones which are compromised and suggests that an audit trail of all tokens (stamps) should be kept in attempt to distinguish between the two. In any situation after such an event, all the documents (even valid ones) would have to be time stamped again, which will be a problem for those whose original digital signature certificates have expired. TrustChain nodes also use the public-private key system to sign their votes, but since every entry into the TrustChain block requires multiple nodes to confirm the entry validity, multiple nodes would have to have their private keys compromised at the same time for an attacker to be able to write an invalid entry into the blockchain. Depending on the number TrustChain nodes this makes such an attack on the system highly impractical.

Point 3 of chapter 4 raises another limitation that does not affect TrustChain. This point states: "The TSA signing key MUST be of a sufficient length to al-

[10] IBM developerWorks. The History of Notes and Domino.

[11] Volarveić, Ira; Stančić, Hrvoje. Standards for electronic time stamps and the possibilities for their application in archival practice.

[12] IETF RFC 3161 - Time-Stamp Protocol (TSP).

[13] Time stamp authority.

low for a sufficiently long lifetime. Even if this is done, the key will have a finite lifetime". This limitation comes from the fact that keys are expected to become vulnerable after a certain period, even if no vulnerabilities inherent to the cryptographic algorithm are present, because of increased processing power. The relevant RFC suggest that as a key length or algorithm reaches the end of its lifetime the documents should be time stamped again with new keys. ETSI 319 422 refers to ETSI TS 119 312[14] to define how long the lifetime of a certain key length (and cryptographic algorithm) is. This document only attempts to predict key length/algorithm durability for up to 10 years or up to the year 2030 (and most combinations do not last even that long). This is insufficient for archiving needs in the context of long-term preservation since many records maintain relevance for much longer and are legally required to be preserved (e.g. 70 years or permanently). As in the previous point, the TrustChain node private keys will also, inevitably, become obsolete and invalid and will need to be changed after a long period but, in the case of TrustChain, this will not affect existing records. This is because TrustChain stores its records in an immutable blockchain and the key is only relevant at the moment of block addition, the fact that certain key length/algorithm combinations will become obsolete will not require "restamping" of old entries as it does in the case of time stamping described by the relevant ETSI standard and RFC document.

In contrast to ETSI 319 422 and RFC 3161, which require periodical restamping, TrustChain provides a (more) permanent solution by writing its entries into a blockchain. Considering this, it is obvious that while the goal is similar, TrustChain is more than a time stamping service – it provides a way to securely store its entries without the need to restamp them. This might be an improvement but it comes at a price. In its current form, TrustChain assumes the existence of a network of trusted (archival) institutions. Not every party might be willing to trust these institutions, or an insufficient number of them might be willing to participate in such a system. It should be noted that the time stamping standards also require existence of institutions that will provide the service but because they require a single institution per service they can be considered easier to implement (and indeed are since many time stamping services already exist). One of the additions to the system currently being considered by the authors and the InterPARES Trust's research group is the addition of a third-party time stamp (instead of simply using TrustChain node clocks) to TrustChain records. This would insure the system encompasses the advantages of both solutions but would further complicate TrustChain as regards of the number of participating institutions and required third party services.

[14] ETSI TS 119 312 V1.1.1 (2014-11) – Electronic Signatures and Infrastructures (ESI); Cryptographic Suites.

6. Conclusion and further work

We have presented a possible solution relevant for the long-term preservation of digitally signed documents. The proposed system is not an archival or digital preservation system for the documents themselves, but rather a standalone system which works in concert with such systems in order to provide the ability to reliably store information on the expiring digital signature validity (or the validity of signatures whose certification authorities no longer exist), without having to trust any single institution. To achieve this goal, the blockchain technology can be relied upon. The proposed system relies on the involvement of a group of trusted institutions that are interested in implementing such a system. Once such a group is identified and the system is implemented, it can be made available to any interested party. The single largest downside of the model is that it only solves the problem for the documents with valid digital signatures. It does not directly provide a solution for the existing documents whose certificates have already expired. These would need to be resigned before having their records written into TrustChain, or a separate blockchain-based solution for storing the validated signatures should be developed and connected to the TrustChain solution.

The proposed model is a prototype and is one of a few possible solutions to the problem of long-term preservation of digitally signed documents being currently pursued by the authors. Further research of this model will include consultation with archival institutions about their willingness to participate in such a project, refining model details with regards to archival institutions' needs, publishing a full project whitepaper, development of a working prototype and testing it in various recordkeeping and archival preservation situations.

Acknowledgements

The research presented here is a part of a much broader research study "Model for Preservation of Trustworthiness of the Digitally Signed, Timestamped and/or Sealed Digital Records (TRUSTER Preservation Model)" which is part of the international multidisciplinary research project InterPARES Trust, http://www.interparestrust.org.

References

A Glossary of Archival and Records Terminology, Society of American Archivists, https://www2.archivists.org/glossary/ (Accessed 30.05.2017)

ETSI EN 319 421 V1.0.0 (2015-06) - Electronic Signatures and Infrastructures (ESI); Policy and Security Requirements for Trust Service Providers issuing Time-Stamps (draft). European Telecommunications Standards Institute, 2016, (Accessed: 4.9.2017)

ETSI EN 319 422 V1.1.1 (2016-03) - Electronic Signatures and Infrastructures (ESI); Time-stamping protocol and time-stamp profiles. European Telecommunications Standards Institute, 2016, (Accessed: 4.9.2017)

ETSI TS 119 312 V1.1.1 (2014-11) - Electronic Signatures and Infrastructures (ESI); Cryptographic Suites. European Telecommunications Standards Institute, 2014, (Accessed: 4.9.2017)

Haber, Stuart; Stornetta, W. Scott. How to Time-Stamp a Digital Document. *Journal of Cryptology*. 3(1991), 2, pp. 99-111.

IETF RFC 3161 - Time-Stamp Protocol (TSP), Internet Engineering Task Force (IETF), 2001, https://www.ietf.org/rfc/rfc3161.txt (Accessed: 4.9.2017)

IETF RFC 5280 - Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile, Internet Engineering Task Force (IETF), 2008, https://www.ietf.org/rfc/rfc5280.txt (Accessed: 29.09.2017)

IETF RFC 6960 - X.509 Internet Public Key Infrastructure Online Certificate Status Protocol, Internet Engineering Task Force (IETF), 2013, https://tools.ietf.org/html/rfc6960 (Accessed: 29.9.2017)

IETF RFC 7159 - The JavaScript Object Notation (JSON) Data Interchange Format, Internet Engineering Task Force (IETF), 2014, https://tools.ietf.org/html/rfc7159 (Accessed: 30.6.2017)

InterPARES Trust Terminology Database, http://arstweb.clayton.edu/interlex/en/index.php (Accessed 30.05.2017)

ISAD(G), General International Standard Archival Description, Second Edition, ICA, Ottawa 2000, http://www.ica.org/sites/default/files/CBPS_2000_Guidelines_ISAD%28G%29_Second-edition_EN.pdf (Accessed: 30.06.2017)

ISO 15489-1 Information and Documentation – Part 1: General, ISO, 2016, https://www.iso.org/standard/62542.html (Accessed: 6.9.2017)

ISO/IEC 18014-1:2008: Information technology – security techniques – time-stamping services – part 1: Framework. ISO/IEC, 2008, https://www.iso.org/standard/50678.html (Accessed: 6.9.2017)

Lemieux, Victoria L.; Sporny, Manu. Preserving the Archival Bond in Distributed Ledgers: A Data Model and Syntax. *Proceedings of the 26th International Conference on World Wide Web Companion*. Perth: WWW '17 26th International World Wide Web Conference, 2017, pp. 1437-1443.

McConaghy, Trent; Rodolphe, Marques; Muller, Andreas; De Jonghe, Dimitri; McConaghy, T. Troy; McMullen, Greg; Henderson, Ryan; Bellemare, Sylvain; Granzotto, Alberto. BigchainDB: A Scalable Blockchain Database. Berlin: BigchainDB GmbH, 2016. https://www.bigchaindb.com/whitepaper/bigchaindb-whitepaper.pdf (Accessed: 20.04.2017).

Microsoft: MSDN Library, .NET Development, Framework Class Library, System.Security.Cryptography.X509Certificates Namespace. https://msdn.microsoft.com/en-us/library/system.security.cryptography.x509certificates(v=vs.110).aspx, (Accessed: 29.9.2017)

Nakamoto, Satoshi. Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf (Accessed: 15.04.2017).

Oracle: Java™ Platform Standard Ed. 7 Online documentation, Package java.security.cert. https://docs.oracle.com/javase/7/docs/api/java/security/cert/package-summary.html (Accessed: 29.9.2017)

The History of Notes and Domino. IBM developerWorks. International business machines, 2007, https://www.ibm.com/developerworks/lotus/library/ls-NDHistory/ls-NDHistory-pdf.pdf (Accessed: 6.9.2017)

Volarević, Ira; Stančić, Hrvoje. Standards for electronic time stamps and the possibilities for their application in archival practice. Arhivi i domovinski rat. Zagreb: Hrvatsko arhivističko društvo, 2016, pp. 425-435.

Preservation of Spatial Information

Göran Samuelsson
Mid Sweden University
851 70 Sundsvall, Sweden
goran.samuelsson@miun.se

Summary

This article highlights the importance of preservation of spatial information, i.e. information clearly connected to a geographic position, regardless of whether it is a matter of GIS or CAD/BIM information. The author consider through his review and compilation of texts, interviews and workshops that we have existing tools for appraisal/selection, metadata and preservation formats that with minor additions are fully useful for creating a reliable digital continuum. The future work consists of – appraisal/selection and preservation – of spatial information and is extensive, and therefore the author believes the work should be initiated with immediate effect.

Key words: Spatial information, preservation, BIM, GIS, Metadata

Introduction and background

As society grows more digital, and more and more of the necessary societal sectors become increasingly information-dense and dependent on information, it is more important than ever that research focuses on what is currently a vital part of all businesses: information. It would make sense to assume that in information society there is a strategy for long-term information preservation and future use. This applies to all societal sectors, but this article highlights its importance within the field of spatial information, i.e. information clearly connected to a geographic position, regardless of whether it is a matter of geographical information or CAD[1] BIM[2] information. The work towards a more long-term preservation of spatial information is still in its infancy. Any difficulties are related to the complexity of the information; information objects often have multiple connections to other information, and there are frequently various types of file formats. Since the 1950s, international standards have been developed within the field of geographical information, which is what makes it possible to exchange geographical information today. For the past 20 years, several countries have increased their efforts and together developed business concepts,

[1] Computer-aided design (CAD) is the use of computer systems (or workstations) to aid in the creation, modification, analysis, or optimization of a design. In Narayan, K. Lalit (2008).

[2] BIM definitions in: Abbasnejad B, Moud H. 2013.

information models, objects and formats for the transfer of geographical information.[3] Today, Sweden has a national strategy for geographical information, the purpose of which is to increase and facilitate collaboration within the field of geographical information.[4] In Europe, the strategy also includes the implementation of the EC directive geographical information infrastructure (INSPIRE), which is a European collaboration aiming to develop synchronized national geographical information. The directive was adopted in 2007, the aim of which was to develop a European geographical information infrastructure. It is expected to be completed by 2019.[5] Even if these discussions and projects are aiming for a synchronized management of geographical information, it is still a work in progress nationally and internationally. All businesses in all societal areas are struggling with the issue of a long-term strategy for the preservation of information supply over time. Similarly, the process of developing strategies for the coordination of CAD/BIM greatly increased during the last 5 years. BIM comes from the world of CAD, and is connected to the introduction of 3D technology. Since 2012 there is an ISO standard, or rather a technical specification – ISO/TS 12911:2012 Framework for Building Information Modelling (BIM) guidance – which includes a definition of the framework for the development of guidance for building information modelling (BIM). The standard is applicable to the development of guidelines for the modelling of buildings and structures of all scales, for groups including several structures as well as single parts of a structure.[6]

BIM is a model that is not only used by architects and constructional engineers to draw buildings. It can also be used for all types of structures, i.e. also roads, bridges and all sorts of land. BIM can be used to digitally develop one or more virtual and precise models of an object. It can be used as a support through all design stages and facilitates analysis and control, compared to manual processes. When these computer-generated models are complete, they include the exact geometry and information needed for design, production and purchase activities, through which the building comes into being. But the important part is really the information – the creation of a shared information model and a set of rules and regulations for the exchange of the model or parts of it. Selected parts of this information will accompany the object's management period for perhaps hundreds of years. In the US, home of what is perhaps the most developed construction industry, more than 70 percent of companies were already using BIM by 2012.[7] Strongly contributing to this interest is that, basically, everyone prof-

[3] http://e2.relationbrand.com/stanli/_Nyhetsbrev-Nyhetsbrev_Nr_7_Oktober_2004/mail.html

[4] https://www.geodata.se/upload/dokument/strategi/geodatastrategi_2012.pdf. Retrieved 2017-03-11

[5] Bartha, G., & Kocsis, S. 2011.

[6] Keenliside, S., & Beange, M. 2016.

[7] Dodge Data & Analytics, 2013

its from its introduction – public administration as well as the construction industry. In England, the government expects a 20-30 percent reduction of administration costs for buildings as a result.[8] All over the world efforts have increased, and a number of national strategies have been developed.[9] In the last 10 years, a number of projects world-wide have dealt with the issue of preservation, but it has not had any profound effect on everyday work in most areas of business.[10]

Purpose

An overall aim of this article is to compile a state of the art of the knowledge we have, for further work within the ISERV-project11 and the subproject that studies the management of information in large infrastructure projects (road, rail, bridges, etc.). Specifically, this article will contribute with knowledge of and answers to the following questions:

- What principles of appraisal and selection are in place?
- What preservation formats are available for information objects and metadata?

At the same time, the result will help to develop a digital continuity model for managing spatial data in Sweden over time.

Problems

Currently, management of spatial information is an important part in a large number of businesses both producing and administrating information. It is a matter of databases that are purely spatial that are used in connection which all types of planning – in centralized management, energy, environment, town planning, school, care, etc. All organizations dealing with location-bound infrastructure, service and care in e.g. a municipal area need access to spatial information. But there are also many administrations who produce a great deal of

[8] Ministry of Business, Innovation and Employment (MBIE), Building and Construction Productivity Partnership. http://www.buildingvalue.co.nz/sites/default/files/Productivity-Benefits-of-BIM.pdf. Retrieved 2016-12-10

[9] New Zealand issued a national handbook in BIM in summer 2014; Australia 2012, buildingSMART Australasia, http://buildingsmart.org.au/advocacy/the-national-bim-initiative-nbi/#.WR_8_VPyjGI. Retrieved 2017-03-10

[10] Shaon, A et al (2011)

[11] The purpose of the ISERV-project is to develop prerequisites of the development of e-services and archival practices. The project is coordinated by a research group at Mid Sweden University and will also be carried out with participants of the Västernorrland region The project focuses on the capturing, managing and reuse of information within organizations. https://www.miun.se/iserv/.

data – data that then remains on respective administration.[12] This means that the quality and format of the spatial information could vary. Because the technical administrations/functions of an authority have been early adopters of the digital tools, processing has often been digital for decades, while the formal decision documents have been printed and archived. For the last few years, some authorities have also expanded functionality and added the option of an e-archive. There is frequently a connection to a business system. Because few authorities have developed central e-archive solutions, it means that the authorities are running a risk of major interoperability problems in the future, as well as difficulties developing integrated and complete analysis and planning based on the information the authority as a whole has at its disposal. In a study, we have highlighted the problem areas that the authorities consider most urgent to deal with when it comes to spatial information:

- Lack of general coordination – fragmented and inaccessible information making any attempt of a comprehensive analysis difficult.
- Lack of reliable metadata and "working databases" that can sometimes be a cause for uncertainty and/or situations of legal doubt, where information on paper and digital material has not been synchronized.
- Lack of long-term preservation strategies. The growing volumes of spatial information require carefully prepared decisions in terms of what to preserve and how it should be preserved.
- Lack of guidelines for which information objects/storages that must, should, or can be preserved – suggestions for selection and priority are also connected to the development of specification for the information package arranged by the National Archives (so-called FGS).
- Inadequate information models that also include aspects of preservation (and information security aspects) within the field of spatial information.
- Lack of appropriate format for the long-term information supply.[13]

The lack of general strategies dealing with the long-term preservation of spatial information could have consequences directly affecting citizens' right to information in the long term. It also counteracts current investments in the digitalization of Swedish e-government data in general. Spatial information is an important part of the everyday work of government services – if it is not satisfactorily preserved it will affect other information collected. A very real example of this is when those in charge of spatial information at Stockholm County Council are no longer able to follow the current geographic growth of the city and analyse growth in connection to the health of the citizens – there is no information for the last 30-40 years.[14]

[12] Inspire 2013: Infrastructure for Spatial Information in Europe Member State Report: Sweden, 2010-2012

[13] Samuelsson, G & Svärd, P 2011

[14] http://folkhalsoguiden.se/amnesomraden/folkhalsoarbete/statistik/folkhalsa-pa-karta/

Method

Due to the nature of the study, methods with a mixed approach have been used, consisting of literature review, workshops, interviews and questionnaire. The literature review is used to map the research problems in previous studies/texts on GIS/BIM and information management. This review leads to the identification of the state of art and also eventually the research gap. In connection with formulating the problem, we carried out a questionnaire, primarily aimed at coordinators and producers of geographical information. We have also carried out a couple of workshops and semi-structure interviews of users of spatial information.

Result

In this section, we aim to compile the parts of the information on GIS/BIM and preservation, which we gathered through the literature review, the questionnaire and interviews / workshop. It can be considered an attempt to analyse and use what is already known to compile a basis for a continuity plan/model[15] and a more operative checklist.

Standardization is always a work in progress, connected to the group working on all geographical information standards in ISO (ISO/TC 211 – Geographic information/Geomatics), aiming to develop a standard specifically for the preservation of geographical information "Standard for the Preservation of Geospatial Data and Metadata: ISO 19165". The group will have to carry on working for some time yet, before being able to deliver something substantial that can be used by any authority. While the standard will be adapted and based on the OAIS standard, it does not seem to take the records or business standard ISO 15489, nor the metadata standard ISO 2308, into consideration.[16]

Our compilation has received the following subheadings:

- Appraisal, Retaining and Disposing of information
- Metadata
- File format/database format

Appraisal, Retaining and Disposing of information

Spatial information is information originating from some type of activity, which is why this type of material, in the formal archive sense, does not differ from social acts, building permits, municipality board minutes, etc. This means that just like with other information flows, it is necessary to map out the processes generating the information (See Fig 1). The easiest way of doing it is to follow activities in connection to a business analysis, as described in DIRKS (Design-

[15] MacLean, Margaret; Davis, Ben H (eds) 1999

[16] Kresse, W m.fl 2015.

ing and Implementing Record Keeping Systems) and in ISO 15489.[17] When it comes to information appraisal, "*Guidance for process-oriented information mapping*" by Swedish Civil Contingencies Agency (MSB) and the National Archives is a good support.[18] Record planning includes a number of activities aimed at determining what business information and types of documents can be found in the organization's business processes – and how they should be handled during their lifetime. The latter is described in further detail in the final step Archival description, the content of which is primarily based on RA-FS 2008:4.[19] The appraisal process takes its starting point in the Archives Act and its portal sections, the needs of your own business and an analysis of other societal needs.[20]

Map and describe processes	**Appraisal and classification of information**	**Set up a documentation plan**	**Organize archival description**
Set up classification structure		Decide types of documents	Take inventory of information storage
Connect to system and organization		Decide basic metadata	Present and describe archives
		Decide metadata handling	

Figure 1. Business processes analysis

Metadata

For all types of information, the importance of metadata is increasing. The more frequent our migration, use and distribution of information, the more important accurate metadata becomes. In the archival context, the standards ISO 23081-

[17] https://www.records.nsw.gov.au/recordkeeping/advice/dirks/methodology. See also Sahlén, T 2016.

[18] Riksarkivet 2012: Guidance for process-oriented information mapping. Retrieved 2017-05-01.

[19] National Archive Sweden (2008): Föreskrifter om verksamhetsbaserad arkivredovisning (RA-FS 2008:4) https://riksarkivet.se/rafs?pdf=rafs/RA-FS%202008-04.pdf. Retrieved 2017-04-30

[20] The major laws and regulations that affect recordkeeping practice of Swedish public organisations are: The Freedom of the Press Act SFS 1949:105; The Public Access to Information and Secrecy Act SFS 2009:4001; The Administrative Procedure Act SFS 1986:223; The Archives Act SFS 1990:782; The Personal Data Act SFS 1998:204; The Public Sector Information Act SFS 2010:566; The National Archives Regulations 'RA-FS

1:2006 and ISO 23081-2:2009 emphasize connections and context, rather than more general standards, which also applies to spatial information. This means that to use the industry standard in an archival project on spatial information is desirable, which in our case agrees with the European Collaboration INSPIRE. The list below is a selection of the items found in the national metadata profile.[21]

- *Name* – What is the official name of the document? map, photo of location 2012, building area etc.
- *Description* – A few lines describing the document, to help the receiver to interpret it. Do add a few descriptive pictures and links to any additional sources for those interested in finding out more if possible.
- *Type* – Is it e.g. lines, polygons, dot patterns, point clouds or text files?
- *File format* – E.g. SHP, DWG, TIFF, ICF .
- *Collection method* – What methods were used to put together the document? E.g. flight data, image interpretation, basic inventory etc.
- *Quality* – Describe the quality. How close are the points of measurement? Flying altitude? Sources of error?
- *Geographical boundaries* – Does it, for example, cover all of the municipality or 100 m on each side of a road?
- *Other* – A free text field to add that which doesn't quite fit elsewhere. Is there anything else the user needs to know? Maybe the eastern part of the area has not been updated the same way as the other parts? Maybe some objects have the incorrect height value z=0, when really was no information about height? That is information that could be entered here.
- *Urgency* – When was it created? And for how long was the layer maintained/updated?
- *Delivery date*
- *Person in charge* – Name and contact information of the person in charge.

This list should be added with more traditional archival metadata describing the aim and context of the information created according to ISO 23081 and the framework found in PREMIS.[22] Metadata Model ISO 23081 is for authoritative forms of transaction and has a recordkeeping focusses on the relationships between business activities, the people involved in them and the records that are produced from and by them.[23] A proposal is that this model (Fig.2) gets an interchangeable optional part. The BIM models describe object and many infor-

[21] Swedish Standards Institute (SIS): Geodata – Nationell metadataprofil – Specifikation och vägledning –SS-EN ISO 19115:2005-geodata.se Version 3.1.1

[22] http://www.loc.gov/standards/premis/. Preservation Metadata is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability.

[23] Sue McKemmish, Glenda Acland, and Barbara Reed, (2000).

mation flows today's have an information related to objects. Object should be able to replace or complement the People function.

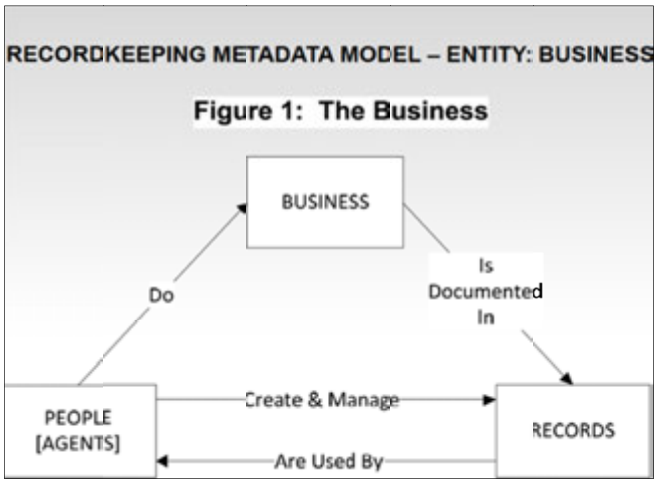


Figure 2. ISO 23081-1, Metadata for Records

File format/database format

Traditionally, Sweden has often been quite "strict" when talking about digital archives and file formats. Special archival formats should be used for this very purpose. It is often explicitly stated that no business formats should remain. In this case, I believe that we need to be more pragmatic and allow the most common business formats in the archive, at least for the foreseeable future. What is important is that the formats are accompanied by detailed documentation about the organization and structure of the format, and of course a reliable organization in terms of rights and regulations, defining who can collect and add information. However, major business actors such as ESRI, who deal with geographical information, have in the last few years started to understand the importance of being able to exchange information. They have developed guidance documents and instructions for their spatial information databases, describing extraction of datasets or entire databases in XML.[24] We have also noted that the Danish National Archives has introduced an updated version of their archive adaptation to the GML standard.[25] A recommendation directly following the definition of which information to preserve, is to transfer the information in

[24] http://www.digitalpreservation.gov/formats/fdd/fdd000295.shtml#sign

[25] GML is short for Geography Markup Language (GML); it is the XML structure that the Open Geospatial Consortium (OGC) has developed to describe geographical elements.
Danish National Archives 2016: https://www.sa.dk/wpcontent/uploads/2016/05/Geodata_anvisning_2016.pdf

question to a more dedicated ”archival server” and start working on developing a delivery/export feature, making constant transfer of spatial information in XML/GML possible from a number of different business processes to the e-archive/archival server. The Danish National Archives have also participated in the European Project E-AR, which have worked with preservation and use of geodata.[26] In the fall of 2017, the first more dedicated standard for the preservation of geodata will also be launched – ISO Standard for the Preservation of Geospatial Data and Metadata: ISO 19165. It will be synchronized with the OAIS model but will still require a lot of work of the business. When it comes to CAD/BIM formats, there are still different voices – some advocate that you must have at least a few different formats depending on the information objects and purpose[27]. However, the dominant view is that the IFC [28] format with continuous updating is more than enough to preserve information over time and create good conditions for interoperability both now and in the future. The project that worked the most with BIM and preservation is – DURAARK, which stands for "Durable Architectural Knowledge", which has produced a report on how to work with preservation and complex BIM-models.[29]

Conclusion

The interpretation, which can be drawn today, is that we, without major problems, should start the tangible work that is currently needed to be carried out with appraisal and preservation – of spatial information. This should be initiated with immediate effect. We have sufficient tools in the form of appraisal and business analysis and preservation format to begin a systematic archiving of spatial information.

References

Abbasnejad B, Moud H. (2013): BIM and basic challenges associated with its definitions, interpretations and expectations. Int J Eng Res Apps. 3:287–294.

Bartha, G., & Kocsis, S. (2011). Standardization of geographic data: The european inspire directive. European Journal of Geography, 2(2), 79-89.

BIM Alliance Sweden (2012): Arkiveringsrekommendationer. Retrieved 2017-04-15
http://www.bimalliance.se/library/2445/del_3_bakgrund_preliminar_utgava_uppdaterad_20120602

E-Ark 2016: http://www.eark-project.com/29-user-stories-scenarios/122-pilot-5. Retrieved 2017-05-05

Evolve Consultancy: BIM Collaboration formats. http://www.evolve-consultancy.com/resource/bim-brief/bim-collaboration-formats. Retrieved 2017-05-05.

[26] E-Ark 2016: http://www.eark-project.com/29-user-stories-scenarios/122-pilot-5

[27] Evolve Consultancy 201: BIM Collaboration formats. Retrieved 2017-05-05; Practical BIM – http://practicalbim.blogspot.se/2013/06/ifc-what-is-it-good-for.html.

[28] buildingSMART: IFC Overview summary. Retrieved 2017-01-20. http://www.buildingsmart-tech.org/specifications/ifc-overview. Business Collaborator Ltd; https://www.groupbc.com/blog/2015/02/10/open-bim/. BIM Alliance Sweden; http://www.bimalliance.se/soek/?q=arkivformat.

[29] Duraark (2015); Tamke, M. (2016)

Dodge Data & Analytics, 2013: "The Business Value of BIM in North America: Multi-Year Trend Analysis and User Ratings (2007–2012)," https://www.construction.com/about-us/press/bim-adoption-expands-from-17-percent-in-2007-to-over-70-percent-in-2012.asp. Retrieved 2017-03-11

Duraark (2015): D7.3 Use case long term Archiving. Retrieved 2017-05-05

Inspire (2013): Infrastructure for Spatial Information in Europe Member State Report: Sweden, 2010-2012. Retrieved 2017-05-05.

InterPARES Trust (2016): Policies for recordkeeping and digital preservation. Recommendations for analysis and assessment services. Retrieved 2017-05-02 https://interparestrust.org/assets/public/dissemination/EU04_20160811_FinalReport.pdf.

ISO 19165 Standard for the Preservation of Geospatial Data and Metadata

ISO 23081-1 2006: Information and documentation – Records management processes – Metadata for records – Part 1: Principles

ISO 23081-2 2009: Information and documentation – Managing metadata for records – Part 2: Conceptual and implementation issues.

ISO/TS 12911:2012: Framework for building information modelling (BIM) guidance

ISO 19165 Standard for the Preservation of Geospatial Data and Metadata

ISO 15489-2 2001: Information and documentation – Records management – Part 2: Guidelines

Keenliside, S., & Beange, M. (2016). A Comparative Analysis of the Complexities of Building

Information Model (ling) Guides to Support Standardization. International Journal of 3-D Information Modeling (IJ3DIM), 5(3), 18-30.

Kresse, W; Masó, J (2015): *Development of an ISO-Standard for the Preservation of Geospatial Data and Metadata: I*SO 19165. In Photogrammetrie – Fernerkundung – Geoinformation 2015(6):449-456

Library of Congress (2017): Sustainability of Digital Formats: Planning for Library of Congress Collections. ESRI Geodatabase XML. Retrieved 2017-05-12. https://www.loc.gov/preservation/digital/formats/fdd/fdd000295.shtml#sign.

MacLean, Margaret; Davis, Ben H (eds) (1999): *Time & Bits: Managing Digital Continuity*. Getty Publications. ISBN 0-89236-583-8.

McKemmish, S; Acland,G and Reed, B, (2000): Towards a Framework for Standardising Recordkeeping Metadata: The Australian Recordkeeping Metadata Schema. Monash University,http://www.infotech.monash.edu.au/research/groups/rcrg/publications/framewrk.html. Retrieved 2017-03-15.

Ministry of Business, Innovation and Employment (MBIE), Building and Construction Productivity Partnership (2012): http://www.mbie.govt.nz/about/whats-happening/news/document-image-library/nz-bim-productivity-benefits.pdf. Retrieved 2016-12-11.

Narayan, K. Lalit (2008): *Computer Aided Design and Manufacturing*. New Delhi: Prentice Hall of India. p. 3. ISBN 812033342X.

National Archives of Denmark (2016): Anvisning i aflevering af geodata til Rigsarkivet. https://www.sa.dk/wp-content/uploads/2016/05/Geodata_anvisning_2016.pdf. Retrieved 2017-05-18

National Archives Sweden (2008): https://riksarkivet.se/rafs?pdf=rafs/RA-FS%202008-04.pdf

National Archives Sweden (2012): Guidance for process-oriented information mapping. Retrieved 2017-05-01

Sahlén, T (2016): Information management in the public and private sectors.

Shaon, A et al (2011) Long-term sustainability of spatial data infrastructures: a metadata framework and principles of geo-archiving. In Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011).

Swedish Standards Institute (SIS) (2005): Geodata – Nationell metadataprofil – Specifikation och vägledning –SS-EN ISO 19115:2005-geodata.se Version 3.1.Retrieved 2017-04-11.

Samuelsson, G & Svärd, P (2011): E-Government Developments and The Challenges of Managing Geodata. Linz : Trauner Druck GmbH & Co KG (Schrifenreihe Informatik 37).

Tamke, M. (2016). Enabling BIM for the full Lifecycle of buildings. In no 3, Geospatial World.

Digital Archives: Towards the Next Step

Arian Rajh
Agency for Medicinal Products and Medical Devices /
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb /
Highflott
Zagreb, Croatia
arian.rajh@halmed.hr

Summary

The author discusses what has to be taken into account when designing or upgrading digital archives today. After specifying standards associated with archival functions, this article brings preliminary communication related to upgrading of digital archival system implemented in the Agency for Medicinal Products and Medical Devices in Croatia.

Key words: digital archives, METS, OAIS, PAIMAS, PAIS, PREMIS, transfer projects

Introduction

There are numerous organisations which have running systems for management of digital records and which are *archiving* their digital content. Many systems were developed according to Open Archival Information System (OAIS) reference specification,[1] using its information model and implementing all or several of its functions. However, implementing just one standard and ignoring other affiliated standards[2] leads to archiving capability fixed to an original digital environment. This misses the point of OAIS concept and its preservation function. In order to facilitate technology-independent long-term preservation function, as well as transfer and exporting to other and future environments, digital archiving should be harmonised with producer-archives interface methodology and functional and system-based archival standards. Digital records and their associated metadata are constantly evolving. Export of content should correspond to one or several metadata and packaging standards.

Most recent archival standards should be included in designs of new systems or in upgrade plans for existing archives. Organisations should fine-tune OAIS functions in previously developed OAIS-compliant digital archives so they

[1] ISO 14721:2012 Space data and information transfer systems – Open archival information system – Reference model

[2] ISO 20104:2015, ISO 20652:2006, ISO 13527:2010, ISO 16363:2012

could support transfer projects to other OAIS environments. Some examples are transfers from public creators' digital archives to principal archival authorities' repositories, as well as other acquisitions performed by archival authorities.

DAIS analysis, findings and plans for the future

Digital Archival Information System is the digital archival system implemented in the Agency for Medicinal Products and Medical Devices (Zagreb, Croatia) during EU-funded IPA project "Preparations for eCTD and Implementation of DAIS" (9.2013-11.2014).[3] Digital Archival Information System (DAIS) was developed according to OAIS information and functional models. Generic interoperability methods were added to DAIS in its EU-financed development phase in order to enable its communication with existing business applications. During annual upgrades between 2015 and 2017, or the second phase of upgrading software, system integration was reinforced by refining generic methods or creating additional methods and additional workflows. DAIS was connected with archival management system so it could enable archiving and archival description of digitised and digitally-born records; connected with DPU Scan Xino internal digitisation system; Quality management system etc. Implementation of internal digitisation in the Agency was done according to authenticity related requirement that included adding identity and integrity metadata and excluded software interpolation from the digitisation process. Archiving of digitally-born records in DAIS comes down to the declaration of a document as a record and its protection by the IBM FileNet Enterprise Records component, placing a record into the structure of creators' fonds and adding ISAD (G) archival description by the archival management system.[4] Digitised records as archival information packages (AIPs) have their preservation description information metadata set up to be the properties of their content and AIPs are being additionally described in an archival sense in the same manner as conventional paper records. The archival description used in the Agency is based on ISAD (G) and ISAAR (CPF)[5] standards so it can be formatted as EAD XML finding aid.[6] Archived digitised records and digitally-born records are

[3] This project was financed by the European Union. The Agency for Medicinal Products and Medical Devices was the beneficiary of the Instrument for Pre-accession Assistance (IPA) programme. The project was implemented by Ericsson Nikola Tesla and AAM Management Information Consulting (author of this article was project manager for the beneficiary).

[4] ISAD (G): General International Standard Archival Description, 2nd ed., 2000. The Agency for Medicinal Products and Medical Devices is aware of changes in archival description. At this stage, the Agency is observing the development of the Records in Contexts standard.

[5] ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families, 2nd ed. Ottawa, 2004.

[6] Besides archival description functionality, application "Centrix Pismohrana" or archival management system developed by IT company Omega Software has some additional and

also being transferred from the active storage to the archival storage. Ingest function, developed in the framework of migration and ingest module of DAIS in the first phase, was automated for internally digitised records in the second software upgrade phase. Export function and preparations for transfer process were developed in the second phase. Export from DAIS is initiated by the archival management application, but the outcome is not harmonised with most recent standards. For the upcoming phase or cycle of DAIS upgrading, related to export and transfer functionalities, it is planned to keep the existing workflow and to adjust information packages or the outcome of this workflow to most recent standards.[7]

Further improvement should be done toward usage of metadata standards like Metadata Encoding and Transmission Standard (METS). METS is XML container for descriptive and administrative metadata, metadata about file groups, package structural map and associated executable behaviours. In its descriptive and administrative sections, it may contain pointers to external metadata. Harmonisation with METS could be done on a level of export of packages as their associated description – in other words, it should be possible to export AIPs as groups of digital objects with METS XML metadata.

DAIS system should be tested against ISO 16363 requirements by using customised self-auditing questionnaire. Harmonisation with ISO 16363 requires at least partial implementation of PAIMAS, PAIS and XFDU. PAIMAS is Producer-archives interface methodology abstract standard aiming to define relationships between producer and the archives or producer's archives and archival authority or any secondary archival environment in our case. It covers OAIS *administration* and *ingest* functions in parts of negotiating submission agreement and ingesting submissions. The PAIMAS process consists of four phases: preliminary phase, with feasibility checks, formal definition phase, which is further elaborated in PAIS standard, transfer phase and validation of submission information packages (SIPs) phase. SIPs – taken from the perspective of public archives as something submitted to public archives – are defined precisely according to PAIS standard. PAIS process consists of the creation of abstract SIP model, the creation of SIP model related to particular producer – public archives transfer project, the creation of SIPs and their transfer and validation. PAIS defines which data and metadata should be provided to archives and in which order. SIPs implementation is supported by using XFDU as packaging standard. (XFDU remained linked mostly to space agencies' domain to a greater extent

customised functionalities linked to archival description and description of a creator, e.g. computer generation of EAD-formatted finding summary inventory. Please see references, Rajh 2016.

[7] Team for this upgrade project will be assembled by of the archival unit and IT unit of the Agency for Medicinal Products and Medical Devices and Ericsson Nikola Tesla's solution manager, developers and testers.

than other standards from OAIS family.) Besides the OAIS itself, the other standards from OAIS series should be implemented in the Agency for Medicinal Products and Medical Devices on the level of packages because the export-for-transfer function was already built in DAIS and it is feasible to modify its output.
Besides harmonisation with ISO 16363 standard, with the holistic approach at the system level, PREMIS standard should be implemented on the same level, by defining repository entities in the organisation.[8] After that, PREMIS should be implemented on the object level, by embedding PREMIS metadata into XML (e.g. METS XML). The Agency should define its PREMIS entities and their properties. The Agency's intellectual entities as single units would include case files, dossiers, databases and books. The Agency's stored objects would include representations and files. Agents, rights and events should be defined in the Agency's archival management manual also. At the end, this will enable additional self-auditing of adherence of DAIS to PREMIS requirements.

Table 1: Digital archive's lifecycle on the example of the Agency for Medicinal Products and Medical Devices

Lifecycle phase	Implementation	Quality measure
1st phase – development of system and system customisation	development of digital archives based on OAIS standard; establishing communications with existing IT islands; management of archived digital records	system level: OAIS, content level: PDF/A
2nd phase – upgrade of software modules and functions	strengthening communication and interoperability with IT applications; building new linked systems and workflows	software modules level, e.g. forensic quality of digitisation process
3rd phase – upgrade of system	fine-tuning of OAIS functions and harmonisation with additional archival standards; management of archived authenticated digital records	system level: PAIMAS, PAIS, XFDU, METS, PDF/A, blockchain technology

DAIS's preservation planning and archival storage functions are facilitated with an automated conversion of declared textual records into PDF/A-1b file format by using particular software. In order to enhance these OAIS functions, it will be necessary to develop internal LTP procedure for conversion from existing archival file formats to archival file formats of next generations.
Besides OAIS ecosystem upgrade requirement, the examined records creator is using electronic signature solution valid only for internal purposes. It will be necessary to adopt the present-day "non-paper" concept of authenticity in the digital environment and to implement a solution which will be easier to maintain and which will be valid outside Agency's boundaries. Regarding obtaining

[8] PREMIS Data Dictionary for Preservation Metadata, http://www.loc.gov/standards/premis/

authenticity of records for the long-term, a substitute of original e-signatures by blockchain technology should be considered.[9]

Conclusion

Digital archives have their lifecycle as any software and hardware do and upgrades on the level of their parts (modules) or entire systems should be anticipated. Croatian Agency for Medicinal Products and Medical Devices carried out several upgrade 2nd phase projects (Table 1) on system modules and functions level after finishing system development. It is very important to plan long-term preservation procedures related to internal conversions/migrations and transfers to external environments, as digital objects and their storage or digital archives are becoming technologically obsolete legacy systems. To be able to move organisational content to tomorrow's digital archives, it is necessary to plan transfer projects according to additional OAIS standards and other archival standards. Preservation cycles should, therefore, include projects in one OAIS system and transfers of packages between different OAIS-compliant systems. Although it is expected to use current technologies for archives management, the goal is to go beyond these technologies and to focus on the preservation function.

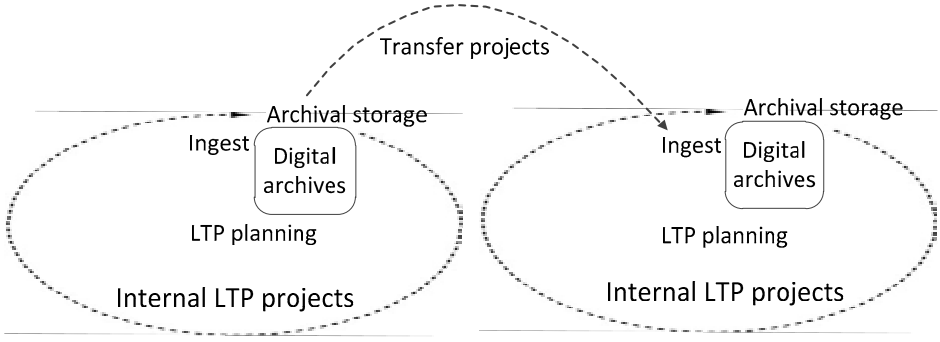


Figure 1: Digital content preservation cycles – a view through OAIS glasses

[9] Among other choices, please see Blanchette (2006) and Stančić (2016) from references.

References

Blanchette, J.-F. *The digital signature dilemma.* // Annales des Télécommunications, 61 (7–8), 2006, 908–923. http://polaris.gseis.ucla.edu/blanchette/papers/annals.pdf (February 2017)

ICA. ISAD(G): General International Standard Archival Description, 2nd ed. 2000

ICA. ISAAR(CPF) International Standard Archival Authority Record for Corporate Bodies, Persons and Families, 2nd ed. 2004

ICA. ISDF International Standard for Describing Functions, 2011

ICA. ISDIAH International Standard for Describing Institutions with Archival Holdings, 2011

ICA. Records in Contexts CM. Consultation Draft v.01, September 2016

ISO 14721:2012 Space data and information transfer systems - Open archival information system - Reference model

ISO 20104:2015 Space data and information transfer systems - Producer-Archive Interface Specification (PAIS)

ISO 20652:2006 Space data and information transfer systems - Producer-archive interface - Methodology abstract standard (PAIMAS)

ISO 13527:2010 Space data and information transfer systems - XML formatted data unit (XFDU) structure and construction rules

ISO 16363:2012 Space data and information transfer systems - Audit and certification of trustworthy digital repositories

METS, http://www.loc.gov/standards/mets/ (February 2017)

PREMIS Data Dictionary for Preservation Metadata, http://www.loc.gov/standards/premis/ (February 2017)

Rajh, Arian. Različite okoline, ista struka: mogućnosti primjene informatičke tehnologije u arhivima za izradu računalno generiranih opisa gradiva. // Arhivski vjesnik. 59 (2016), 1; 99-116.

Stančić, Hrvoje. Long-term Preservation of Digital Signatures // Technical and field related problems of traditional and electronic archiving / Gostenčnik, Nina (ur.). Maribor: Pokrajinski arhiv, 2016. 481-491

Chain of Archival Requirements. Usability of Digital Records in the Context of e-Services in Sweden

Erica Hellmer
Department of Information Systems and Technology
Mid Sweden University
Holmgatan 10, 85170 Sundsvall, Sweden
erica.hellmer@miun.se

Summary

This paper presents an ongoing case study within the ISERV research project at Mid Sweden University, examining the process of changes in the records management process when paper-based turns digital and also examining the challenges in the change within the profession of the archival domain. The findings of this study will contribute to an understanding of the creation of sustainable e-services where the information will continue to be usable, with an archival requirement perspective over time.

Key words: digital records management, digital preservation, e-services, ISO 15489, ISO 40100, usability, accessibility

Introduction

E-government facilitates administration, making it more efficient, open, and more comfortable for both citizens as well as organizations. The *EU eGovernment action plan 2016-2020* is accelerating the digital transformation of governments by suggesting principles with the purpose to coordinate the modernization of the new digital government, and to "modernize public administration, achieve cross-border interoperability and facilitate easy interaction with citizens"[1].

Within Sweden, governments and municipalities endeavor to create and implement an efficient electronic government in order to provide public services to citizens and to create a digital records management. Along the rapid ongoing technical development and the continuing demand of service and use and reuse of information the development of e-services and e-archives is increasing. In the digital agenda of Sweden, one challenge is to have a digital records management that supports the organizations, but also a records management that includes the archival requirements. Nevertheless, e-services created today that in-

[1] European Commission (2016) p. 4

clude, or will include, the profession of archivists or other specialists within the information domain are seldom created by the archivists themselves. Users, external and internal, of e-services are often elucidated as well as usability issues regarding the design of public e-services but rarely the accessibility and preservation of the information that these new e-services will handle.[2]
Changing from paper-based records management towards a digital records management creates challenges and also a changing role of the archivist or other information professionals in preserving the evidential value. The changing role also includes making access for present and future use when moving from the role of a guardian to actively intervene in the records creation process and shaping the collective and social memory.[3]
This ongoing case study is examining e-services created today with a usability and preservation perspective and also the new active role of information professionals. The process that will be studied is a new e-service that will be created by The Swedish Companies Registration Office (Bolagsverket) and focuses on managing digital financial information, in this case annual reports. By using the standard XBRL (ISO 40100), the former paper-based process of annual reports will turn digital. This is an initial study, and a part of the ISERV research project[4], in examining the domain of e-services and will contribute to the creation of e-services with e-archive requirements.

Research problem and questions

New decisions regarding implementation of new techniques or new standards set demands on both current and future records management and records (re)usability.
The process of annual reports is today paper-based where different types of information in a variety of formats are submitted, controlled, scanned, and preserved. This process is to be computerized through a digital service. Examining the process of how this digital service will function in an authority which will create the digital service, and also examining the users of this service, will provide an ecological and representative case study where we gain insight in the construction of e-services. It is also relevant to examine the profession of the archival domain when the work process is changing.

[2] Goldkuhl G, Röstlinger, A (2010)

[3] Millar (2010)

[4] The main goal of ISERV research project is to develop prerequisites of the development of e-services and archival practices. The project is coordinated by the a research group at Mid Sweden University and will also be carried out with participants of the Västernorrland region The project focuses on the capturing, managing and reuse of information within organizations. https://www.miun.se/iserv/

The primary research questions are the following:

- How is usability and preservation considered when paper-based records management turns digital?
- What does the use of new e-services implicate to the long term records management at The Swedish Companies Registration Office (Bolagsverket)?
- How is the role of information professionals affected?

Limited companies are obligated to annually report to The Swedish Companies Registration Office according to the annual accounts act (1995:2554). These reports include directors' reports, profit-and-loss accounts, balance sheets, and notes.[5] An annual report is therefore not only one document and retaining a complete chain of information is relevant.

It is also relevant that the aggregation of records of an organization meet the recordkeeping requirements in order to provide links between records, i.e., meaning and evidence. A record's authenticity, integrity, and understandability derives from its relationships with other records, i.e., the aggregation of records.[6] Within the records management context, the international standard 15489-1:2016 provides guidelines to "ensure that authoritative evidence of business is created, captured, managed and made accessible to those who need it, for as long as it is required".[7] The standard sets out requirements for records which are requirements for evidence of business activity. These records requirements are context-dependent, and may pertain to any records process, may apply to whole functions, industries or jurisdictions, and should be linked to particular functions, activities or work processes. According to the standard the processes for creating, capturing and managing records should be integrated to procedures, records systems and should be supported by policies. For ensuring continuing usability it is suggested in the standard to prepare a plan to enable continuing access and usability.

Research design

The design for this study is a qualitative case study consisting of semi structured interviews with key personnel at The Swedish Companies Registration Office with the purpose of collecting data of the process of the annual reports. Interviews with two limited company will also be carried out in order to gain knowledge regarding the work process, how and which information is structured when these annual reports is transferred to the Swedish Companies Registration Office. Today, limited companies prints, signs, and posts their annual reports to the Swedish Companies Registration Office. It is relevant to study the

[5] Annual accounts act (1995:2554)

[6] Hofman (2005)

[7] ISO 15489 (2016) p. VI

whole chain of information in classical structure in recordkeeping i.e. when it is created, captured, organized and made accessible.
The study will follow the process of the new e-service at the Swedish Companies Registration Office from planning to implementation and ultimately do a reflection of the whole process. This initial study is investigating how the role of information professionals is affected by a new work process: focusing on both when the process of annual reports turns digital, but also what the use of the new e-service will implicate to the long term records management at The Swedish Companies Registration Office. As it is today, the process of annual reports is mostly paper-based and in an initial interview with an archivist at The Swedish Companies Registration Office it is described as outdated. However, the interviewee envisions that an implementation will simplify the re-usability, preservation, and open up for new functions. The interviewee ads further that it is of great importance to be able to reuse information both within The Swedish Companies Registration Office as well as outside of the authority. The international standard 15489-1: 2016 can be used as an audit tool, a checklist of the new digital work process at Bolagsverket.
Figure 1 shows a traditional records management with traditional steps of responsibilities and functions from creation to archives management. The figure is also inspired by the records continuum model, created by Frank Upward[8], with a post-custodial perspective where the role of archivist are more active in the organizations where the records are created and used.

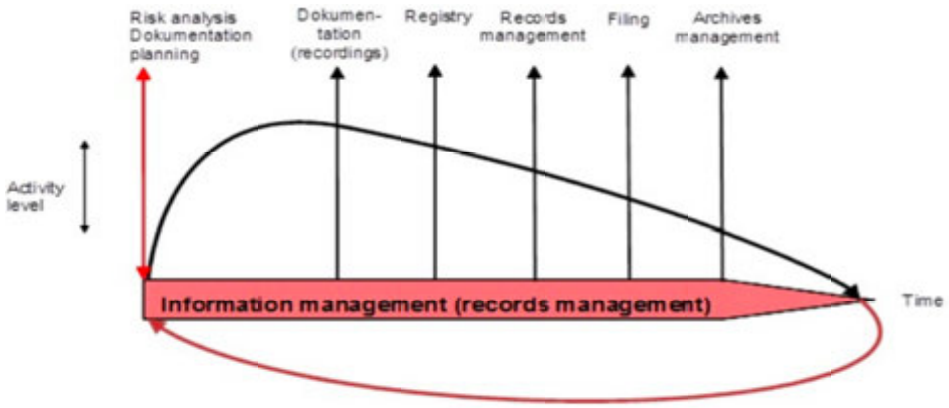


Figure 1. Components of digital information management

[8] Upward (1996)

The continuum model can be used as a supporting process when evaluating and exploring the different steps in the recordkeeping process at The Swedish Companies Registration Office today and after the implementation of the digital service.
Records continuum thinking and practice has brought records managers and archivists under a "recordkeeping umbrella". It has focused on unifying and multiplying purposes and frameworks for accountable recordkeeping which may enable access to useable evidence of social and business activity. A records continuum approach enables partnerships with stakeholders in business and information domain. Previous collaborative partnership with records managers, archivists and standard setters resulted in standard development. There is an interest of collaboration with IT professionals, archivists, records managers, system managers or other stakeholders in order to develop coherent information architecture and metadata structures.[9]

Preliminary results

Results of this ongoing case study suggest that e-services created today need to be efficient, accessible, and reusable. It indicates a more dynamic and continuing records management that is opened and advocates reusability. When examining this digital service at The Swedish Companies Registration Office, which has a well-established and structured process in managing paper-based annual reports, it will open up for new ways to analyze and measure processes and records.
A preliminary observation regarding the profession of information professionals that handles, and will handle the information of the annual report through the e-service indicates a need of intern education regarding the understanding and development of e.g. metadata and making data searchable and accessible. It indicates that the profession of the archival domain can contribute with a proactive and holistic approach. In the context of the digital agenda and the modernization of the new digital government the ISERV research project will elucidate not only the practical way when developing new e-services and the active role of the archive profession but also whether archival principles do in fact apply.

[9] McKemmish, 1997

References

Annual Accounts Act (1995:1554). Swedish code of statutes. Webpage: http://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/arsredovisningslag-19951554_sfs-1995-1554 Retrieved at 19/5 2017

European Commission. Communication from the commission to the European parliament, the council, the European economic and social committee and the committee of the regions. EU eGovernment Action plan 2016-2020. Brussels, 19.4.2016. COM (2016) 179 Final.

Goldkuhl, G; Röstlinger, A. Development of public e-services – a method outline. // Paper accepted to the 7th Scandinavian workshop on E-Government (SWEG-2010) January 27-28 2010, Örebro University

Hofman, H. The archive. // Archives: Recordkeeping in Society. McKemmish, S, Pigott, M, Reed, B, Upward, F. (eds). Wagga Wagga, NSW: Centre for Information Studies, Charles Sturt University. (2005)

International Organisation for Standardisation. Information and documentation – Records management – Part 1: Concepts and principles. 15489-1:2016.

McKemmish, S. Yesterday, today and tomorrow: a continuum of responsibility.// Proceedings of the Records Management Association of Australia 14th National Convention, 15-17 Sept 1997, RMAA Perth

Millar, L.A. Archives principles and practices. // Principles and practices in records management and archives. Series Editor: Geoffrey Yeo. (2010)

Upward, F. Structuring the Records Continuum – Part One: Postcustodial principles and properties. // First published in Archives and Manuscripts, 24 (2) (1996)

PERSONAL DIGITAL INFORMATION MANAGEMENT

Gluing Provenance to Dispersed Personal Content and Creating Contemporary Personal Archives

Arian Rajh
Agency for Medicinal Products and Medical Devices /
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb /
Highflott
Zagreb, Croatia
arian.rajh@halmed.hr

Krešimir Meze
Think Big Hub
Zvonimira Furtingera 1, Zagreb, Croatia
kresimir.meze@thinkbighub.com

Summary

The article discusses the necessity of providing archival support to creators of personal archives. The authors state the need to overcome the fragmentation of contemporary personal fonds, bind together overall personal content and facilitate long-term preservation. The article focuses on the preliminary communication of personal digital archiving requirements and presents the development of the prototype software product.

Key words: Personal Digital Archiving (PDA), digital legacy, prototyping

Introduction

Personal archives or fonds are entireties of records created and accumulated by individuals during their lives. They appear to have distinct qualities compared to other types of archives because the transactional value of the records is not their prevalent value (Hobbs 2001, 128). According to Hobbs, personal archives are "archives of character" and they are "documenting our complex inner humanity" (Hobbs 2001, 135). The same definition should also relate to personal digital archives.

"[T]hey accumulate as the person goes about his or her own private work and life's activities and are ordered (or not) to suit the individual's proclivities and needs." (Eastwood 2016, 19). Contemporary digital personal archives,[1] in the

[1] "The term 'personal digital archiving' refers to how individuals manage or keep track of their digital files, where they store them, and how these files are described and organised. People keep personal archives for reasons that may be simultaneously sentimental, practical, and even accidental.", Redwine, G. ibid. Furthermore, "There is an emerging scholarly discipline, personal in-

majority of instances, will hardly end in public archives where they would be arranged and kept according to the professional standards. Even with the best will in the world, the accumulation of digital personal records and content surpasses the capacities of professional processing in archival institutions. Personal archives will not be structured and arranged by archivists and there is a considerable risk of loss of their content. Today it is recognised that public institutions are creating just a part of future memory and that private records creators accumulate socially and culturally relevant archival materials too. This adds severity to the risk of continuously losing potentially valuable archival holdings.[2]

Personal archives have particular ingest policy, organisation, structure and a notion of value. They are also different from institutional papers concerning the transfer to archival institutions and their processing. Maybe it's time to change our approach to personal archives – if they are different from the typical archives that archival professionals process in the archival institutions, and if their accessions to archival institutions are not likely to happen. Preserving potentially valuable archival holdings should be the main drive for archivists to intervene and assist individuals in their preservation efforts. "The archival profession needs to concentrate on developing new mechanisms for education the public about how to care for their personal and family archives" (Cox 2009, 107). Because of large production of personal materials, we propose an additional way to do it – with a help of ICT tools with built-in professional knowledge. This proposal is in line with contemporary archival mission. There are a number of tools for content creation, capturing, storing, sharing and recycling available to individuals. However, there is a lack of tools which might ensure content cohesion - tools for gluing records together, gluing provenance to the content and for organizing records and saving their interpretative potential for the future. Provenance refers to the archival principle of separating material of different origins (and grouping material of the same origin).

Archival tools for the organisation of personal fonds are unavailable to personal records creators (end users). This leaves personal accumulations unstructured,

formation management, archivists need to begin both to dig into and to influence (and there is probably a natural connection to the other emerging area, digital curation).", Cox, R. J. Digital Curation and the Citizen Archivist. //: Digital Curation: Practice, Promises & Prospects. Uni. of North Carolina, 2009, p. 108.

[2] More creators archive their records than in the past and creators of personal archives are included in social power dynamics. Their small narratives and records could become important in the future for seeing the bigger social picture. That is the reason why archival profession should proactively offer methodologies and tools for assembling of personal archives. If personal archives are better organised and arranged, it will be easier to use the legacy of their creators in the case of acquisitions in the future. After all, "the more control an individual has over their personal archives, the greater their ability to save only what they intend", Redwine, Gabriella. Personal Digital Archiving. DCP Technology Watch Report 15-01, December 2015, Digital Preservation Coalition, http://dx.doi.org/10.7207/twr15-01 (accessed January 2017), p.2.

provenance links decaying and content scattered. Our research questions are linked to the quality of emerging tools for personal digital archiving (PDA) and to characteristics of personal archives:

1. *How to create personal digital archives today?* More exactly, what kind of user and professional archival requirements should PDA tools support in order to facilitate personal digital archiving and to mitigate the risk of losing personal content?
2. *What is the purpose of creating personal archives?* What characteristics should contemporary PDA tools and personal archives have in order to promote that purpose?

From PDA requirements to PDA model and prototype

To answer the first set of research questions we analysed and divided user and professional requirements to those that address long-term preservation of content and those that address content accumulation. After requirements analysis, we proposed mechanisms for their implementation in two stages: on a conceptual level of model and on the functional level of actual prototype tool. We are at the stage of developing the prototype now and we are preparing it as open source software and commercial SaaS solution. Later in this article, we will show use cases which will be covered by our prototype and which should also, in our opinion, be covered by other potential ICT solutions.

In Table 1 we have listed the requirements that PDA software model and the prototype should satisfy. These requirements are central problems that PDA conceptual model and tool should solve; implementation mechanisms on the model level are functional solutions to the stated problems and implementation mechanisms on prototype level are software functions. We are using the methods from Table 1 for the development of the prototype.

The first content-related requirement (IA) deals with the usage of various cloud file hosting services and mail solutions. The personal digital archiving solution should be able to serve as the portal or central point of access to personal content stored in various environments (e.g. digital repositories, mail services, social media platforms, professional and scientific social networking sites, shared content). We imagine digital personal fonds[3] as portals – this is based on the idea of fonds as intellectual constructs or conceptual aggregations in archival theories of Barr, Cook and Yeo (Barr 1988; Cook 1993; Yeo 2012). Creators should be able to have access to their records and have them readable and usable over the long term. Creators' records include files of various file formats,

[3] Fonds – "[t]he entire body of records of an organization, family, or individual that have been created and accumulated", Glossary of archival and records terminology, https://www2.archivists.org/glossary/terms/f/fonds (accessed 26/7/2017)

e.g. textual formats, image formats, multimedia container formats.[4] Content LTP requirement (IB) presumes proactive monitoring of the content. Although certain qualities of the content should be assured before archiving, in the act of creation of the content, a tool for personal digital archiving should monitor the targeted properties and enable the creator to act at the right time. PDA prototype was not intended to be used for conversion of file formats – our viewpoint is to use functionalities and tools already developed by other IT companies and to focus on archival issues.

Table 1: Methods used for functional model and prototype of proposed PDA solution

Requirements		**Model level implementation – concepts**	**Prototype level implementation – functions**
I	*Requirements related to digital content*		
A	Provide access to personal content	Portal that enables a central access point for selection and processing of content for personal fonds	*Add account* function for the integration with various software solutions by using open web services
B	Content long-term preservation (LTP)	Relying on OAIS and related standards for monitoring all content and export of information packages	*Analyze&report* function; *Export* function
II	*Requirements related to aggregation of digital content*		
A	Establish virtual provenance of fonds	Binding all selected content into the personal fonds by using the pre-set structure	*Select units for my fonds* function; *Create/update my fonds* function
B	Facilitate interpretative potential of fonds	Relying on archival professional standards for description, formatting standards, cognitive services for indexing, linked data and semantic web for selected descriptions	*Add archival description* function (with advanced features, e.g. linking with cognitive services for photos in the fonds)*;* *Make selected description available* function

Besides these two *content-oriented requirements* (IA, IB)*,* there are two requirements *linked to content accumulation* or fonds (IIA, IIB). Binding widespread content into personal fonds (IIA) is the central concept around which our PDA tool should be built. The tool should propose the structure of personal fonds to end user and the placement of the archived item in the structure. User modifications of the structure should be allowed because there are differences between various personal fonds and the final structure depends on many factors related to creator and material. The archival description should be added (IIB) for the purposes of taking intellectual control of the content, for retrieval and for

[4] The technical registry PRONOM, accessed 25/7/2017 https://www.nationalarchives.gov.uk/aboutapps/PRONOM/default.htm

preservation. It should be possible to track identity, integrity and captured context(s) of creator's records. We are considering using cognitive services could for additional content analysis and indexing. All of this brings us back to the second set of research questions which is about a purpose – or purposes – of personal archives. Records creator creates personal fonds to keep his or her records and contents together, to keep personal memories and evidence.[5] This purpose is expected. However, although personal archives are relatively isolated from the environment, archives, in general, show a tension between availability and unavailability. Openness for usage is, of course, closer concept to public records creators' holdings. But it is linked with any accumulation of memories. To some extent sharing of selected content of personal archives is similar to the writing of "personal" memoir writing or diaries.[6] Any writing is supposed to be read; not just by the author; it is something fused with its very existence. The perplexing situation with archived items of personal provenance resembles the situation with autobiographical prose; they want to be found. So, this secondary purpose is something linked with the desire to participate in wider social memory. Sometimes records creators want to be present in a society and want to contribute, promote him or herself, provide or add something related to his or her point of view. In the end, a local historian would be very glad to retrieve and to access photographs or other evidence of locally significant historical event taken and provided by private records creator. We have considered this second purpose also in the course of prototyping. That is why we have designed characteristics that contemporary personal archives need to enable participation in the shaping of the new archival landscape of tomorrow. It could also be convenient for accruals of material into an existing personal archive or for sharing archived items with family members who live apart (which could lead personal records creators to the creation of virtual family archives). Providing access to selected descriptions and then to photographs, videos and other records, could be facilitated by using linked data and semantic web concepts. Appropriate archival description assures interpretative potential of the fonds and enables preservation of personal heritage for future use by the creators or his or her heirs. It's like sharing action… but with archival qualities.

[5] "With the proliferation of networked information and communication technologies (ICTs), the definition of personal archives has expanded to include the collections of potentially any individual with an archival impulse to document his or her life.", Acker, Amelia; Brubaker, Jed R. Death, Memorialization, and Social Media: A Platform Perspective for Personal Archives. // Archivaria 77 (2014), p.3.

[6] Please see the example taken from a contemporary website called Public Diaries. This service hosts private and public diaries. Many authors decide to make their diaries openly accessible, https://www.my-diary.org/surf/ (accessed 26/7/2017).

Having both sets of research questions in mind, we decided to cover and automatize PDA use cases of *personal fonds creation* (creation and arrangement of archival information packages and archival units, creation of descriptions)

1. *intended participation in wider memory with selected materials*
2. *preparation for preservation of information packages with the export function*

We think that PDA tools, in general, should implement these fundamental use cases. The desirable contemporary PDA tool should be easy to use and it should reflect the smart integration of ICT and professional knowledge into daily life. Based on this analysis of requirements and these use cases, we[7] are creating the prototype of PDA solution called "Legacy Sky"[8] which will enable building personal fonds and preservation of personal digital legacy. PDA tools in general should solve the problem of dispersing storage and facilitate access to digital content stored on any digital repository. It should help users to preserve their digital content and targeted properties of the content. It should help users to organise their content. In one sentence, the tool should bind the content scattered through various spaces and add archival structure and description. From that point onwards, content takes the form of personal fonds. It can be preserved as a coherent unit from that point on. Storage and access to content, as well as preservation issues, are connected with requirements related to digital content and structuring and adding the description with requirements related to aggregation of digital content.

PDA prototype and its possible impact

We strive to implement previously stated PDA requirements and develop PDA tool prototype. We assume that innovative linking of different cloud services (repositories, social networks etc.) into one archives will provide new value for personal records creators. For prototyping PDA software product we are using lean-based project management method that combines several different lean canvases as well as state-of-the-art *scrum* methodology.[9] The prototype is based on web portal connected with content storage providers through open web services. It is also based on several straightforward software functions which provide a solution for previously identified requirements.

[7] Highflott (http://www.highflott.com) and Think Big Hub (http://www.thinkbighub.com/).

[8] http://www.legacysky.com

[9] Business model canvas by Alexander Osterwalder and Yves Pigneur; lean canvas for start-ups, adapted by Ash Maurya. The canvas was used for developing Legacy Sky start-up as the joint venture between our companies. For the development of our prototype we are using *scrum* software development methodology (GitHub Boards, Agile project management https://www.zenhub.com/). We plan to include early adopters in our *scrum* development process.

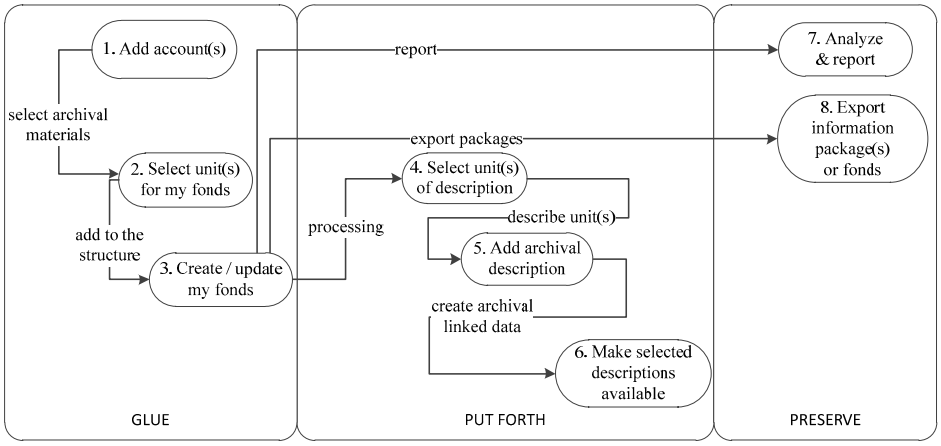


Figure 1: High-level functional model of the PDA tool Legacy Sky prototype

Archival functions of the tool that "glue together" personal content are: (1) add cloud service account and (2) select units for fonds. The functions in control for description and (consequent) visibility of archival materials of personal fonds are: (3) create/update fonds, (4) select units for description, (5) add description (including metadata created by cognitive services for the description of images) and (6) make selected descriptions (and units) available. The functions that support the proactive preservation of archives are related to (7) analysis of file formats and (8) export of information packages (in OAIS sense) and they are still work in progress. The tool should propose an archival structure of the personal fonds for the organisation of personal content[10] and lead end users to add the archival description in an easy manner. The offered generic structure should facilitate placement of records by the creators and facilitate growth and maintenance of personal fonds. Although generic structure should be suggested by PDA tools, they should support certain flexibility during the development of specific structures, especially knowing that this flexibility is quite common in the archival processing of personal fonds. Advanced users should be able to adjust the structure so that it could match their *character* better (Hobbs 2001) or

[10] E.g. personal records, correspondence, notes, records created in business activities, auxiliary records in Škalić-Štambuk, M. Vrednovanje arhivskog gradiva u osobnim arhivskim fondovima. // Arhivski vjesnik 38 (1995), 88-89. Another generic structure – biographical and personal records, correspondence (except messages inseparable from particular professional business activity records, cooperation or project records), professional and other activities, the body of work/œuvre, ownership and financial records, testimonies of others, collections, varia (with subseries formed by creators). Lucic (2014) mentioned several possible structures for personal archival fonds and family fonds. But, there is no one "perfect" structure, so we are "stuck" with generic structures that could be applied for the majority of archives.

present the life and work of creator[11] in a more apparent way. Users should be able to add subseries for sorting out closed documentary units. Subseries of business-related records, cooperation and projects series should be harmonised with creator's activities. Development and usage of such IT tool would provide many users with a helpful possibility to preserve their content recorded in various repositories for the future. Innovative linking of various online services by using such a tool is going to create additional value in PDA domain. If tools are built on professional concepts and methods, individuals who create and store their records and content will preserve them in the proper way. With creators participating in their archives' environments, this could have an impact on the preservation of future memory in a proactive way.

Conclusion

The risk of losing memory relates to conventional and digital personal archives. M. Lučić showed in her book about personal archives that 43% of personal fonds *preserved* in Croatian cultural institutions consist of one box of material. This analysis was based on the example of 1949 personal archives preserved in Croatian National Archives and other cultural institutions in Croatia.[12] Analysis of completeness of personal archives showed that 7% of them relate to one year of creator's life and 5% of them relate to the period of 2-10 years.[13] Lučić stressed out a deficit of material and/or structural asymmetry of preserved personal archives. The principle of provenance was not respected when materials of the same creator were scattered in different physical institutional repositories and without linking (Lučić 2014, 64-66). We share Lučić's concern for personal archives and consider potential harm very critical due to extensive production and proliferation of digital materials, as well as their volatility. Usage of different repositories without having a portal that glues the pieces together is the same as having one particular creator's legacy in custody of various institutions. Everybody creates digital content, manages these materials through various repositories and experience obsolescence problems severe enough to require some action. "We are all archivists", as Richard J. Cox stated in one section heading of his 2009 article.[14]

[11] Lučić, M. Osobni arhivski fondovi: Arhivistički pogled na prikupljanje, obradbu i interpretaciju rukopisnih ostavština u baštinskim institucijama. Zagreb: HDA, 2014, p. 123.

[12] Lučić, ibid., p. 57.

[13] This analysis was based on General guide for archival fonds and collections in the Republic of Croatia (2006) finding aid, Lučić, ibid., p. 58.

[14] Cox, R. J. Digital Curation and the Citizen Archivist. //: Digital Curation: Practice, Promises & Prospects. Uni. of North Carolina, 2009, p. 107.

We can agree with Hobbs's perception of personal archives as different from public records and with Cox's perception of archivists' role[15] in their digital preservation. Postcustodial archivists ought to be advisors. The definition of archives should be broadened to include not just results of practical activities of institutions but also someone's personal legacy.[16] There are difficulties of processing personal archives by archivists in formal archival institutions and these difficulties are even more severe in a domain of personal *digital* archives. Archival processing done by the professionals in archival institutions cannot possibly meet the needs of all records creators. That is why we considered an approach based on the principle of helping people to help themselves. This approach comes down to creating new tools with built-in archival logic for records creators. Contemporary personal digital archives should be created by using current technologies and professional logic but in a creative manner and according to the needs of existing personal archives creators. They are quite different from 19th Century gubernatorial organisations. This was related to the first research question. The purpose of creating a personal archive, to answer the second question, lies in the preservation of content for any other personal purposes, including content sharing and thus participating in wider information realm.

PDA enables taking timely actions, it envisages new role of archival knowledge and ICT tools in the contemporary world, and it answers the requirement to support a new form of digital meta-literacy. At last, reinforcement of creation of personal archives by using archival logic already built in computer applications, so users are not burdened with professional principles, represents a possible further integration of ICT into people's daily life. What does it mean? A person doesn't have to be famous to have his or her personal archives arranged professionally and described properly. Capability to archive his or her own personal content in a form of fonds and even to offer archival material to the community represents archiving democratization (similar to data democratization). There is too much digital content nowadays and archiving should be facilitated by ICT tools with built-in archival logic. Users should be able to keep their digital legacy and even *participate* in the creation of the digital landscape of tomorrow.

[15] "In this role archivists will function more as advisors rather than acquirers, educators giving their knowledge away rather than protecting the secrets of a guild, and advocates rather than reactors in seeking to preserve the portion of the documentary universe that possesses archival value.", Cox, ibid., p. 108.

[16] Please see Melina Lučić's elaboration of a sharp distinction between public archives and personal archives (or legacy) in more traditional archival theories (Jenkinson, Casanova, Brenneke, Duranti) and merging of these notions to one definition of archives in more recent or more flexible theoretical approaches (Lučić 2014, 16-21).

References

Acker, A.; Brubaker, J. R. Death, Memorialization, and Social Media: A Platform Perspective for Personal Archives. //: *Archivaria* 77 (2014), 1-23.

Barr, D. The Fonds Concept in the Working Group on Archival Descriptive Standards Report. Archivaria 25: includes supplement, "The Archival Legacy of the Department of the Interior" (Winter 1987-88)

Barrett, H. Personal archives: Documenting the stories of our lives. 27.6.2014. Financial Times. https://www.ft.com/content/b1bf5044-f5ba-11e3-afd3-00144feabdc0 (February 2017)

Cook, T. The Concept of the Archival Fonds in the Post-Custodial Era: Theory, Problems and Solutions. Archivaria 35 (1993), pp. 24-37.

Cox, R. J. Digital Curation and the Citizen Archivist. //: Digital Curation: Practice, Promises & Prospects. University of North Carolina School of Information and Library Science, 2009, pp. 102-109.

Cox, R. J. Personal Archives and a New Archival Calling: Readings, Reflections and Ruminations. Duluth, Minn.: Litwin Books, 2008

Eastwood, T. A Contested Realm: The Nature of Archives and the Orientation of Archival Science. U: MacNeil, H.; Eastwood, T. (ur.). Currents of Archival Thinking, 2. izd., Santa Barbara: Libraries Unlimited, 2016., pp. 3-23.

Guide to preserving your digital legacy. February 2015. Saga legal services. https://www.saga.co.uk/saga/media/Legal/Digital%20legacy%20guide%20Feb%2015/Digital%20Legacy%20guide.pdf (February 2017)

Gränström, C. Access to current records and archives, as a tool of democracy, transparency and openness of the government administration, Arh. vjesn., god. 42 (1999), pp. 79-92

Hobbs, C. The Character of Personal Archives: Reflections on the Value of Records of Individuals. // *Archivaria* 52 (2001), 126-135.

ICA. Records in Contexts CM. Consultation Draft v.01, September 2016

ICA. ISAD(G): General International Standard Archival Description, 2nd ed. 2000

InterPARES project. http://www.interpares.org/ (February 2017)

ISO 14721:2012 Space Data and Information Transfer Systems – Open Archival Information System – Reference Model

ISO 20104:2015 Space data and information transfer systems – Producer-Archive Interface Specification (PAIS)

Lučić, M. Osobni arhivski fondovi: Arhivistički pogled na prikupljanje, obradbu i interpretaciju rukopisnih ostavština u baštinskim institucijama. Zagreb: HDA, 2014.

Maurya, A. Running Lean: Iterate From Plan A to a Plan That Works, 2012.

MacNeil, H.; Eastwood, T. (ur.). Currents of Archival Thinking, 2. izd., Santa Barbara: Libraries Unlimited, 2016.

Muller, S.; Feith, J.A.; Fruin, R. Manual for the arrangement and description of archives: drawn up by direction of the Netherlands Association of Archivists. Chicago, IL : Society of American Archivists, 2003. http://hdl.handle.net/2027/mdp.39015057022447 (July 2017)

Osterwalder, A.; Pigneur, Y. Business Model Generation: A Handbook for visionaries, Game Changers, and Challengers, 2010

Redwine, G. Personal Digital Archiving. DCP Technology Watch Report 15-01, 2015, Digital Preservation Coalition, http://dx.doi.org/10.7207/twr15-01 (accessed January 2017)

Škalić-Štambuk, M. Vrednovanje arhivskog gradiva u osobnim arhivskim fondovima. // *Arhivski vjesnik* 38 (1995), 81-91.

Yeo, G. The Conceptual Fonds and the Physical Collection. Archivaria 73 (2012), 43-80.

E-ENCYCLOPAEDIA

Original scientific paper

Epistemological Value of Contemporary Encyclopedic Projects

Ivan Smolčić
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
ivan.smolcic@lzmk.hr

Jasmina Tolj
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
jasmina.tolj@lzmk.hr

Zdenko Jecić
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
zdenko.jecic@lzmk.hr

Summary

Contemporary encyclopedic projects are a continuation of a multi-century evolution of encyclopedic work. The first generations of digital (electronic) encyclopedias brought a significant shift in the practicality of their use and enriched them with multimedia and hypertext. Contemporary encyclopedic projects are a step further and contribute new epistemological values as opposed to traditional and digital editions. This paper presents this added epistemological value through the example of a few modern projects, and provides a contribution to detection and systematization of the newly created values.

Key words: encyclopedia's epistemological value, web-based encyclopedia, contemporary encyclopedia.

1. Introduction

As comprehensive projects, encyclopedias hold a great epistemological potential as a high-quality, simple, and quickly available source of information. Since the era of Enlightenment, traditional encyclopedias have put forward a new approach to organization and structuring of knowledge, and thus represent the forerunner of modern information systems. The paradigm of encyclopedic work is transforming as encyclopedia enters a new era of development, of constantly being developed by editors, associates, and users, and as it is now based on ad-

vanced IT and communication technology solutions. For these reasons, a new epistemological assessment is needed to evaluate these new characteristics.

2. Scope and methodology of research

To present the encyclopedia's evolution brought on by technological development and new media, epistemological characteristics of encyclopedic works were analyzed, beginning with traditional printed ones, through early digital works on CD-ROMs and DVD-ROMs, up to contemporary dynamic web-based encyclopedic projects. Some aspects of few well-known general encyclopedias will be concisely analyzed, that were foremost created as printed works to later have their content presented on the internet, and thus made accessible to a larger number of users. Firstly, *Encyclopedia Britannica*[1] will be analyzed, the oldest encyclopedia in English that is still in production. It was first published in 1768 in Edinburgh, Scotland, its last 15th edition was released in 2010, and since then it continues development as a web-based edition. *Brockhaus Enzyklopädie,* the largest printed encyclopedia of present in German was first published 1796–1808. The last printed copies of the 21st edition were sold in 2014, and the content of its 300,000 articles stored digitally. The Croatian Encyclopedia (*Hrvatska enciklopedija*2) is a general encyclopedia available at the web site of The Miroslav Krleža Institute of Lexicography, and it is based on the printed edition published in 11 volumes during 1999–2009. The influence of traditional encyclopedias will be compared with contemporary web-based encyclopedias that function as knowledge portals.

The *Stanford Encyclopedia of Philosophy*[3] project represents the link between web-based encyclopedic editions and peer reviewed original scientific papers in all areas of philosophy. This encyclopedia is exclusively web-based and was set up in 1995. An overview of *Wikipedia*[4], the global multilingual [web-based] edition, based on mass collaboration, initiated in 2001, will be analyzed. Its content is openly editable, meaning anyone is permitted to create and edit its content. The possibility of advanced information retrieval and interconnectivity is demonstrated through the example of Wikipedia's DBpedia[5], ontology database generated from Wikipedia.

This paper analyzes epistemic characteristics of contemporary encyclopedias that include continuity (staying up-to-date), collaboration, boundlessness of scope, information retrieval, and interconnectivity. The continuity, boundless-

[1] https://www.britannica.com/ (accessed Apr 5th 2017)

[2] http://www.enciklopedija.hr/Default.aspx (accessed Apr 5th 2017)

[3] https://plato.stanford.edu/ (accessed Mar 8th 2017)

[4] https://en.wikipedia.org/wiki/Main_Page (accessed Apr 17th 2017)

[5] http://wiki.dbpedia.org/ (accessed May 15th 2017)

ness of scope, and collaborative character of the web-environment will be presented through a quantitative analysis of numerical indicators, demonstrating the epistemological upgrade when compared to traditional encyclopedias. For sake of brevity and keeping in mind the usually strict form of each of the encyclopedia-types (i.e. traditional, digital and web-based encyclopedia), not all aspects of the fore-mentioned encyclopedias were analyzed to be compared to the next.

3. Analysis of encyclopedias' particular epistemological features

3.1. Continuity and collaboration

Epistemic importance of encyclopedias staying up-to-date is reflected in how quickly the information in them out-dates, which reduces the epistemological effect of these works. Data can be outdated even at time of publishing and the time-line between updated editions is regularly several years[6]. The need to continuously keep track of reach and development of human activity is indicated in regular publishing of updated encyclopedic editions. Encyclopedia Britannica, existing for over 240 years, had since 1768 reached 15 editions before it was published online. The Brockhaus Enzyklopädie project began in 1796, and reached 21 editions. The editions differed in scope and number of volumes, thus limiting the number of articles and the content of each. Development of means of digital storage, along with appearance (1970s and 1980s), and later increase in number of personal computers (1990s) permitted encyclopedia's scope growth. In 1986 the Grolier's *Academic American Encyclopedia* was published on a CD-ROM, titled *Electronic Encyclopedia*[7], and became the first digital encyclopedia to be distributed on a digital portable media. The printed version was made up of 21 volumes with 10,000 pages of text, while the CD-ROM was a text-only data base, with over 9 million words in 30,000 articles. The use was enabled by a software for search and retrieval of information, allowing search by article title, or entire text. After digital editions of mostly smaller encyclopedias and encyclopedias for young adults, larger encyclopedias gradually decided to try out in digital publishing. In 1994 Encyclopedia Britannica joined in and started publishing digitally, first as *Britannica online*[8], and in 1995 the same version was published on a CD-ROM[9]. The first digital version

[6] For example, the initial volume of Encyclopedia of Technology, published by The Miroslav Krleža Institute of Lexicography, was printed in 1963, and hasn't been updated since, meaning that for decades there was no encyclopedic update of achievements and reach in technology.

[7] The Electronic encyclopedia, KnowledgeSet Corporation and Grolier Electronic Publishing Inc., 1986.

[8] Then at: http://www.eb.com/eb.htm

[9] Britannica CD, BCD, Version 1.0., Chicago: Encyclopedia Britannica 1994

of the German Brockhaus encyclopedia, *Der Brockhaus Multimedial*[10], was published in 1998 in an abbreviated form[11]. In 2005 the complete digital version of the entire 21st edition of Brockhaus encyclopedia[12] was published on a USB memory card – *Brockhaus Enzyklopädie Digital*[13]. It was to be internet-accessible free of charge since 2008, but that was never realized. The first comparisons between printed and digital encyclopedias highlighted their physical differences, organization of data, (un)limited scope, and (un)changeable content. Digital editions were updated more frequently, usually annually, and along with the advantages of more information retrieval options and added multimedia, a significant reduction in cost of publishing was also achieved.

Contemporary web-based encyclopedias are continuous projects, issued daily by updating existing and adding new entries, making them a more relevant source of knowledge. Further development is aimed at information retrieval capabilities, web design, setting up metadata, and so on. The largest project of such type is Wikipedia. It is based on mass collaboration, allowing anyone to create or modify content, with chronological storage of previous versions. Free access entails a certain risks, such as reduced accuracy in relation to professional works as well as reduced relevance, objectivity, and other components of the encyclopedic concept[14], which are the basis of the encyclopedic work. Nevertheless, such a global, and relatively reliable source of information offers great opportunities for finding information. The Stanford Encyclopedia of Philosophy is a continuous project facilitating constant collaboration between authors and editors working in universities and institutes around the world. The editorial board, made up of experts in each field of philosophy, selects authors, designates articles, and also reviews their work.

The project strives to be up to date, so articles are intended to be revised and updated by the author every 3 to 5 years. Since the content is constantly changing, every update is archived to avoid problems of citing the articles. This allows the project not to out-date as it represents a system for assimilation, processing, and dissemination of new information, and regular implementation in the content of existing articles. The reach of current knowledge in philosophy is thusly monitored, kept up-to-date and tailored to users' needs.

[10] Der Brockhaus Multimedial, Mannheim: F.A. Brockhaus, 1998.

[11] Electronic digital edition Der Brockhaus in fünfzehn Bänden, Mannheim: F.A. Brockhaus, 1997

[12] Brockhaus Enzyklopädie, 21st Vol., Mannheim: F.A. Brockhaus, 2005.

[13] Brockhaus Enzyklopädie Digital, Mannheim: F.A. Brockhaus, 2005.

[14] The encyclopedic concept, along with the stated properties of accuracy, relevance and objectivity, also includes: comprehensiveness, credibility, uniformity, complexity, organization, and keeping up to date. More in: Jecić, Z.: Enciklopedijski koncept u mrežnom okruženju. Studia Lexicographica, 7(2014) 2(13), str. 99–115.

This organization and cooperation of scientific circles around the world in the field of philosophy can be named a scientific encyclopedia[15], because it directly provides results of scientific research or content at the level of review papers. Significant epistemological upgrade of such projects, provided on-line as public service free of charge, stems from their availability to anyone with an internet connection.

The Croatian Encyclopedia[16] is a comprehensive project by The Miroslav Krleža Institute of Lexicography. Though it was primarily produced as a hard copy edition, it is now web-based and shares the advantages of such encyclopedias. It is constantly being updated by editorial staff to prevent falling out of date, and has a growing corpus – expansion of existing and addition of new entries. It is collaborative and open to all users through inquiries and comments to the editorial staff which often leads to new insights and the raise in quality of content. Table 1 shows the number of editorial interventions, article updates and added new entries, which highlights the continuity of work and longevity of the project. The number of user comments in 2015 was 822, declined in 2016 (435 comments), then rose again in 2017, when the number of comments in just the first four months numbered 1231. This also represents a measure of interactivity between users and editorial staff, the user's interest in specific topics, and indicates what users find relevant.

Table 1: Interventions to articles of the Croatian Encyclopedia

Year	Number of updated articles	Number of new articles
2014	14,438	248
2015	5,900	200
2016	4,193	176

3.2. Boundlessness of scope

Due to cost of print and practicality of use, traditional encyclopedic form is limited in scope, having abundance of abbreviations that crowd the text and at times having to omit interesting information. Contemporary works, however, have no such limitations, allowing the use of more natural text flow and mention of any and all relevant information while keeping great information density, characteristic of encyclopedic works. While digital encyclopedia is partly boundless in scope (digital storage is finite), web-based encyclopedia, practically speaking, is not. Boundlessness of scope of web-based encyclopedia is

[15] The term scientific encyclopedia refers to some vocational or special encyclopedias that gather and process the materials of science, art, or specific field. E.g. by The Miroslav Krleža Institute of Lexicography Medical Encyclopedia (1st edition 1957–65, 8 vol.), Encyclopedia of Technology (1963–97, 13 vol.), or personal Krležijana (1993–99, 3 vol.) dedicated to Miroslav Krleža, Institute's founder and first director.

[16] Hrvatska enciklopedija (The Croatian Encyclopedia). 11 vol. Zagreb: The Miroslav Krleža Institute of Lexicography, 1999–2009.

manifested in its dynamic growth and development through expansion of existing articles and addition of new ones. Printed editions are for the better part planned in advance in scope and number of volumes. Scope ceases to be crucial to work organization in digital editions, and work deviates from the strict lexicographical form, but it remains concise and highly informative. The Croatian Encyclopedia's corpus, during 2014–2016 period, grew by 27,577 lines of text[17], including updates to existing articles and addition of new entries. The same, by each year, is shown in Table 2. The overall corpus growth is indicative of the mentioned boundlessness of scope and, if printed, these additions alone would equal around half of one printed volume of a multi-volume edition.

Table 2: Corpus growth of the Croatian Encyclopedia

Year	Total interventions (lines of text)
2014	9,052
2015	7,581
2016	10,944

Another example of an unlimited encyclopedia is Wikipedia. There are currently[18] 295 active editions of Wikipedia in different languages, with a total of 44,092,120 articles. A complete printed edition would require an estimated 16,539 volumes. The largest edition by far is in English, with 5,363,179 articles, making up 12.2 % of Wikipedia, while the Croatian Wikipedia has 172,727 entries, or 0.4 %. The English Wikipedia increases in size each day for as many as 800 new entries.

3.3. Information retrieval and content connectivity

The importance of efficient and simple information retrieval is fully apparent in large, comprehensive encyclopedic works, especially where one article covers a multitude of terms. Tools for thorough, complex information retrieval provide an epistemological upgrade of the work, making it more efficient through better availability of information.

The basis of traditional encyclopedic content organization, along with the standard alphabetical order of articles, is the index - a list of all mentioned terms (titles), concepts, or keywords listed in some order (most often alphabetically), with reference to the article(s) where each is mentioned. Index is especially important when articles cover several terms or names making it less significant in encyclopedic dictionaries and lexicons and a necessity in large general encyclopedias and encyclopedic atlases, allowing a specific concept (term) to be found even in large review articles that can cover several or even dozens of terms. In addition to its organizational function, the index plays a role in comparing dif-

[17] A standard line of text contains 60 characters (with spaces).

[18] https://meta.wikimedia.org/wiki/List_of_Wikipedias (accessed Apr 17th 2017)

ferent encyclopedic works, whereby the number of terms in the index, rather than the number of articles themselves, is often a good indicator of the extent of scope. In printed works, the index system is mundane - branched linear organization of knowledge, such as found in libraries and archives.

Retrieving information in a digital encyclopedia means access to the entire content promptly and efficiently, encompassing article titles and/or entire text. To maximize use and potency, data should be tagged, that is set up in a meta-data network that the computer can process. Such a system's example is DBpedia, an ontological database and a semantic representation of Wikipedia. It represents a knowledge base that comprises of individuals (objects), classes (collections or object types), attributes (related properties, appearances, characteristics, or parameters that an object can have or distribute), and relations (the way objects are related to each other). Ontology is a formal representation of concepts with well-defined relationships between these concepts. Using the info-boxes within Wikipedia articles as structured content, DBpedia extracts meta-data describing a large number of entities (persons, places, music, film, organizations) and it contains RDF triplets[19] extracted from various multilingual editions of Wikipedia. It covers many domains, is updated alongside Wikipedia, is multilingual, and available on the web. Via search interface, the system provides exhaustive information retrieval for a given query as it includes the entire Wikipedia and not just the article titles and text that corresponds to the query's characters (letters).

Traditional encyclopedic content is internally interconnected through referrals and index, but not to outside content (other than with references). Digital works, other than having hypertext links, are no different while the contemporary web-based encyclopedia's content can be connected to any other information on the web. This greater content connection brings about a new epistemological value, making web-based encyclopedia a part of the internet's immense data-base. By structuring content, tagging it and setting up meta-data, connection on multiple levels[20] is enabled. Publishing such content on the web that computers can process facilitates connection to other databases on semantic level.[21]

Further advancement can be made using machine learning algorithms and tools for natural text processing, which have become quite efficient in analyzing raw

[19] The RDF model is based on statements about resources known as triplets. Each is presented in the form of a subject-predicate-object statement, and can be constructed as a graph with two nodes (subject and object) connected with the predicate.

[20] DBpedia extracts metadata, or structured content from Wikipedia articles in form of info-boxes. By setting relationships between objects and associated attributes, a network of terms is created that represents Wikipedia content. Searching such a database provides a deep overview of the entire content.

[21] Such kind of interconnectivity represents the Linking Open Data project: http://lod-cloud.net/ (accessed Mar 8th 2017)

text and allow for the human-created ontologies to be omitted, but are yet to be further implemented in encyclopedistics.

5. Conclusion

This research shows that contemporary web-based encyclopedia is characterized by greater epistemological value regarding its update possibilities (work's continuity), collaboration potential, having unlimited scope, far more possibilities in information retrieval, and content connectivity options.

Increase in human cognizance and knowledge emphasizes the continuous need to organize, catalog, and synthesize knowledge, formatting it into the encyclopedic form, making encyclopedia a highly conductive knowledge-manual. Encyclopedia's overview and capacity for retrieving information provides a number of epistemological benefits: comprehensiveness in a given field, high expertise and relevance for a broad user circle, objectivity, and high accuracy, making it a reliable source of information. Encyclopedia's digitalization, and foremost its new web-based form, changed not only the scope of lexicographer's work, but also its epistemological features. Contemporary encyclopedias, being based on information technology of today, epistemologically surpass traditional encyclopedic form. The way users gather new knowledge has changed as well because web-based encyclopedia is now surrounded with a multitude of data circling the web. Web-based encyclopedia can now take on a role of a digital database. Software tools allow thorough search of content, interconnectivity with other knowledge sources, hypertext links, far more frequent updates, etc., making them an effective public service. Encyclopedia has changed and continues to change, or develop. This phenomenon can surely be named development since digital and especially web-based encyclopedias carry a number of benefits, including the quality of printed editions, thus becoming a place for easy access and exchange of trusted and connected content. Digitalization and globalization lead to networked and interconnected knowledge, eliminating obstacles to fast information retrieval. These features make encyclopedia the epistemological support in digital information. The future of web-based encyclopedia is in effective interconnectivity of its own content and linking to outside sources of knowledge. The responsibility of encyclopedia as an information source in the epistemological sense becomes greater, since they are now available to an even greater extent to anyone with an internet connection.

References

Bizer, Christian; Heath, Tom; Berners-Lee, Tim. Linked Data – The Story So Far. (2009) http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf (May 5th 2017)

Featherstone, Mike; Venn, Couze. Problematizing Global Knowledge and the New Encyclopedia Project. *Theory, Culture & Society*. 23 (2006), 2–3, pg. 1–20.

Jecić, Zdenko; Boras, Damir; Domijan, Darija. Prilog definiranju pojma virtualna enciklopedija. *Studia lexicographica.* 2 (2008), 1(2), pg. 115–126.

Jecić, Zdenko. Enciklopedijski koncept u mrežnom okruženju. *Studia Lexicographica.* 7 (2014), 2(13), pg. 99–115.

Kubelka, Ozren; Šoštarić, Petra. Wikipedija nasuprot Hrvatskoj enciklopediji, kvalitativan odnos slobodno i tradicionalno uređenoga enciklopedijskoga sadržaja a hrvatskome jeziku. *Studia lexicographica.* 5 (2001), 2(9), pg. 119–134.

Thagard, Paul. Internet epistemology: Contributions of new information technologies to scientific research. (1997) http://cogsci.uwaterloo.ca/Articles/Pages/Epistemology.html (April 24th 2017)

Allen, Colin; Jagodzinski, Cecile. *From SEP to SEPIA: How and why Indiana University is helping the Stanford Encyclopedia of Philosophy*. *Against the Grain.* 18 (2006), 4, pg. 42–43.

Allen, Colin; Nodelman, Uri; Zalta, Edward N. The Stanford Encyclopedia of Philosophy: A Developed Dynamic Reference Work. *Metaphilosophy.* 33 (2002), 1–2, pg. 210–228.

Hammer, Eric M..; Zalta, Edward N. A Solution to the Problem of Updating Encyclopedias. *Computers and the Humanities.* 31 (1997), 1, pg. 47–60.

Kane, Gerald C.; Ransbotham, Sam. Collaborative development in Wikipedia. (2012) https://arxiv.org/pdf/1204.3352.pdf (Mar 21st 2017)

Katz, William A. Introduction to Reference Work, Volume I. New York: Mc Grow Hill Book Company, 1978.

Nodelman, Uri; Allen, Colin; Zalta, Edward N. Stanford Encyclopedia of Philosophy: A Dynamic Reference Work. *Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries*. New York: IEEE, 2003, pg. 380.

Pasternack, Jeff; Roth, Dan. The Wikipedia Corpus. (2008) http://cogcomp.cs.illinois.edu/papers/PasternackRo08.pdf (Mar 7th 2017)

Perry, John; Zalta, Edward N. Why Philosophy Needs a 'Dynamic' Encyclopedia. (Nov 1997) https://plato.stanford.edu/pubs/why.html (April 13th 2017)

Tananbaum, Greg. I Hear the Train A Comin'. *Against the Grain.* 18 (2006), 1, pg. 84–85

Torres, Diego; Molli, Pascal; Skaf-Molli, Hala; Diaz, Alicia. Improving Wikipedia with DBpedia. SWCS - Semantic Web Collaborative Spaces Workshop 2012 in 21st WWW Conference. Egyed-Zsigmond, E.; Gripany, Y.; Favre, C.; Largeron, C. (ed.). Lyon: ACM, 2012, pg. 1107–1112

Voss, Jakob. Measuring Wikipedia. *Proceedings of the 10th International Conference on Scientometrics and Informetrics*. Ingwersen, P.; Larsen, B. (ed.). Stockholm: Karolinska University Press, 2005, pg. 221–231

Zalta, Edward N. The Stanford Encyclopedia of Philosophy: A University/Library Partnership in Support of Scholarly Communications and Open Access. *College & Research Libraries News.* 67 (2006), 8, pg. 502–504.

Zalta, Edward N. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/prc-oct99.html (Mar 10th 2017)

Original scientific paper

Encyclopedic Knowledge as a Semantic Resource

Marko Orešković
National and University Library in Zagreb
Hrvatske bratske zajednice 4, Zagreb, Croatia
moreskovic@nsk.hr

Ivana Kurtović Budja
Institute of Croatian Language and Linguistics
Republike Austrije 16, Zagreb, Croatia
ikurtov@ihjj.hr

Mario Essert
Faculty of Mechanical Engineering and Naval Architecture
Ivana Lučića 5, Zagreb, Croatia
messert@fsb.hr

Summary

Extraction of semantic information from unstructured natural language texts is currently a hot topic in the field of computational linguistics. The majority of researchers agree that the core of the problem is in determining "semantic domains", which would give the computer the real meaning of the text. The aim of this research is creation of semantic domains from online encyclopedic texts. This article explains two different approaches to morpho-syntactic analysis that could be used to achieve that.

Key words: knowledge extraction, semantic domains, ontologies, encyclopedia

Introduction

Categorization as seen by Aristotle is the first method of setting the order in a well-organized set of data (information), i.e. creating knowledge about the world around us. Today we can say that Aristotle's categories are the first attempt to establish formal ontology (Bosančić 2016) and the emergence of science of classification – the taxonomic/hierarchical organization of those categories, i.e. the knowledge that is immanent. Over the past three centuries, encyclopedias have kept and maintained that knowledge as an organized set of information. The methods of categorization and classification have been improved for decades, but the core content stayed the same. However, a shift from the printed (paper) media first to the digital one and subsequently by incorporating a network structure in encyclopedic work, essentially changes not only the preparation and processing of encyclopedias (Jecić 2013) but also their content. It is not just about the ease of access to information (by clicking hypertext

links – surfing to the new 'hints' of a virtual encyclopedia) but the possibility of deeper entrance into meaningful aspects of information, both for human and the machine. The machine has the ability to learn and acquire knowledge. That is how an artificial intelligence is developed. As an example, we can use the definition of knowledge from a network encyclopedia (Wikipedia):

Knowledge is a familiarity, awareness, or understanding of someone or something, such as facts, information, descriptions, or skills, which is acquired through experience or education by perceiving, discovering, or learning. (Wikipedia; https://en.wikipedia.org/wiki/Knowledge)

Such definition allows a user to understand the meaning of the main term (knowledge) through its parts, hypertext links (facts, information, description,...), and that new knowledge can further be expanded by opening the links from the definition, so it is just a matter of time and effort which user has to take to acquire knowledge (transfer it to his/her biological data store/brain). The machine already has that information stored in itself, and it does not need any additional collection and processing time. The data connections (in the databases) are exact and strong which enables fast information retrieval.

This kind of 'computer knowledge' can be used (by user) in one of the following ways:

1. In some domains of human activity to solve a problem in steps based on the stored information. An algorithm selects one of the predefined options in each step (Alberico and Micco 1990). These are so called expert systems.
2. By collecting facts that are not specially arranged, categorized or classified, but hidden in the information itself. These are so called inherent ontologies.

This paper will present such research – retrieving semantic information which is neither apparently visible nor explicitly mentioned. An already published, existing knowledge stored in public repositories from the Internet will be maximally used. The aim of the paper is to show how by changing the structure of the language lexicon (Orešković, Brajnović, and Essert 2017), very difficult natural language tasks can be performed (e.g. metaphorical expressions detection, metonymy detection, or detection of any other tropes that people use in everyday writing or speech). In the second chapter, the basic ideas in similar research studies will be presented in order to compare them with a proposed one (in the third chapter). The fourth chapter of the paper will show the available resources and their constraints regarding our main goal. Finally, we will show the realization of the proposed idea – a segment of a network framework that, based on the encyclopedic knowledge, will create the semantic domains which are the basis for the discovery of stylistic figures of the natural language.

Related work

There are numerous studies in a field of an explicit knowledge extraction from online resources, that relied on globally available knowledge databases (most often Wikipedia and WordNet) in order to build a machine readable ontologies. They all aimed to provide a method for information extraction that would have a higher or at least the same quality like those that are manually made. One such ontology is YAGO (Suchanek, Kasneci, and Weikum 2007) which is created by an automated process of information extraction from both Wikipedia and Wordnet. The aim was to create a larger (in terms of the number of facts) and better (by adding additional knowledge) ontology which can later be used as a valuable resource in computational linguistic studies.

Another research project in this filed is Java-based WikiPedia Library (JWPL) and Java-based WiKTionary Library (JWKTL) information extractor developed by (Zesch, Müller, and Gurevych 2008). These libraries use Wikipedia's and Wiktionary's API's to extract explicit knowledge and serve it over API for usage in other projects.

Project Wikify! (Mihalcea and Csomai 2007) also deals with automatic information extraction from textual documents. It parses the text in search of relevant keywords which are then linked to Wikipedia articles.

For Croatian language there were very few researches that deals with semantic. Most significant one is by (Šnajder and Almić 2015).

A new approach to the word tagging

Binary information and associated computer data types (integer, real, string, ...) allow only binary representation of words, but do not provide any grammatical or meaningful features related to them. In order for a computer to 'learn' a natural language, it is necessary that each stored word is tagged with a more appropriate character. The process is similar to coding (at a higher level) of any terms in an encyclopedic dictionary; without the definition, the word itself has no meaning (what is 'Knowledge', from the example above, becomes clear only when the description is read: '… is a familiarity, awareness, or understanding of someone or something, …'). At the level of grammar, mark-up/tagging means that each word must be associated with the part-of-speech category it belongs to (e.g. noun, verb, adjective) and/or other relevant grammatical features (e.g. gender, number, etc…). Several sets of tags were designed for grammatical mark-up, called the tagsets, which are then manually or semi-automaticaly associated to alphabetically ordered lists of words. The manual mark-up process is of course slower, but more accurate than the automated one, especially for languages characterized by rich and complex morpho-syntactic structures (such as a Croatian). Regardless of the way of mark-up, there is another key improvement to that process. As demonstrated at the EURALEX conference (Orešković, Čubrilo, and Essert 2016), instead of a classical vector grammar-semantic features (e.g. MULTEXT-East), it is possible to introduce a new, so-

called T-structure that brings significant possibilities. In the T-structure, there is no difference in the definition and usage of grammatical and semantic features, although they are still being considered separately (as grammatical: WOS - *word of speech* or as semantic: SOW - *semantics of word*). The MULTEXT-East tagset, designed by a unification of different languages in this area, is replaced by a structure that allows the diversity of language features, and highlights the differences that people tend to use in their language. Compared to the linear tagging, the T-structure introduces taxonomic (ontological) characteristics of the word. Instead of making an ontology for a domain (an area of interest) from a dictionary or a word list, each word represents a unique ontology within itself in this case, and its features form branches of a tree to any depth. The information that is stored in the branches of this tree, (except values – string or numeric literals) may contain a link to other branches, a new ontology or a repository of a network information. For the purpose of creating a general linguistic Syntactic and Semantic Framework (SSF), automated links with open network repositories (the Croatian Language Portal (HJP), The Miroslav Krleža Institute of Lexicography's online encyclopedia and Wortverbindungen http://www.lingua-hr.de by Dr. Stefan Rittgasser) are created. In the same way, the "ontology of words" is expanded with other online resources (CroWN, Wikify, ...) or specialized dialects/thesauri in digitized form (Šarić, Google translate, etc.).

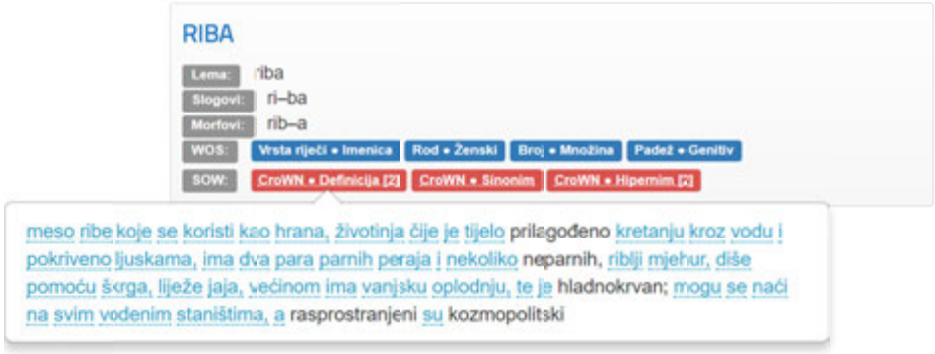


Figure 1. Lexical entry definitions in SSF

A user (or the machine if called via API function) along with grammatical information (WOS; blue tags) also sees a rich semantic information (SOW; red tags) collected locally or from other network resources (respecting the copyrights of their owners) as shown in Figure 1. It is very important for the SSF user to have all the information in the same place, and even more important to have that information interconnected and linked (notice a blue links inside the definition that are direct links to other lexical entries). For automatic information gathering, the existence of already embedded (human) knowledge about

lexical entry is highly important (e.g. different meanings of the same word (homonyms), or the same meaning of different words (synonyms), which is suitable for creating semantic domains, for both literal and metaphoric meaning). This possibility is used concisely in the research presented in Chapter 5 of this paper. Finally, it is worth mentioning that the T-structure can be extended in both (WOS or SOW) directions. Each registered user can make their own tree tags, but of course, they have to make sure that new tags are applied to the words in the dictionary. Such user operations will not interfere with the work and the results obtained by other users.

Open network repositories and dictionaries

By open network repositories, we imply permanent projects that are updated on a daily basis, and are in open (free) access and often allow the user to collaborate in the form of contributions or comments. Certainly the most famous such repository is Wikipedia, which is envisioned as an organized set of encyclopedias written in different languages. Wikipedia content today is provided in about 300 languages, most commonly associated with official and/or national languages. Ten years ago, the famous Miroslav Krleža Institute of Lexicography and its Knowledge Portal http://enciklopedija.lzmk.hr, which encompasses several digitized editions, in line with the Croatian Family Lexicon, began to embark on Wikipedia prestigious footsteps. The SSF uses this information (available online) in agreement with the Institute's disclaimer: "It is permitted to use or quote individual articles in parts or as a whole with the indication of the source". This gives an SSF a new dimension - indexing each word from the definition and creating links to other words from selected network repositories. For CroWN (Croatian version of Wordnet) such indexing means increasing semantic links in the repository itself for at least the order of magnitude. In a similar way, grammatical information from the Croatian Language Portal (http://hjp.znanje.hr) is retrieved, and then automatically compared with those obtained from the morphological generator (Markučič and Govedić 2013) over the "Hrvatska riječ" dictionary (Pinjatela 2001) which is a part of the SSF. The comparison and verification of multiword expressions is achieved with the help of http://www.lingua-hr.de and the algorithm for more accurate decomposition of words into morphs and syllables. In that way the SSF becomes an integrator (hub) of network frameworks that give its dictionary/thesaurus characteristics that have only been theoretically discussed in the models of Generative Lexicon (Pustejovsky 1991) and Meaning-Text Theory (Melcuk 1981). The Figure 2 shows an example of one lexical entry with all the accompanying features (Syllable, Morph, WOS, SOW, MWE etc.)

Figure 2. Lexical entry with all accompanying features in SSF

Formal or computer ontology, defined by (Gruber 1993) as an explicit specification of the conceptualization of a particular domain or a shorter "conceptualization specification", provides the highest level of machine-readability of a particular area. The T-structure has all the features of ontology, at the word level, that is, the lowest level of natural language, which makes it especially strong. In formal and machine-readable (higher) ontologies, developed with the Protégé (or similar) tools, the knowledge is defined by classes and subclasses, their instances, objects, data properties and their interconnections, thus allowing an extremely high degree of formalization. It should be emphasized that the creation of ontology is not the greatest step in knowledge creation, but a process or idea that enables the development of new knowledge from an existing ontology (Antoniou and Harmelen 2004). That is achieved by this kind of lexicon organization. It is easy to notice the two-way activity - the computer in its handling of stored knowledge in an intuitive, understandable, simple and visible way, represents the knowledge to the user and the user controls that knowledge and extends it to new information and/or links to other repositories. In that way the user and the machine create a kind of symbiosis with their complementary work, both of them expanding the basic components of knowledge – the data and their relationships (information). The data and information together makes a knowledge that in the end, philosophically speaking, by understanding and evaluation forms a wisdom (that is known as DIKW hierarchy - which was long represented as a fundamental model: (Ackoff 1989) and (Liew 2007), but also challenged by (Frické 2009). After showing some limitations of word ontology, primarily because of dialectal variation, which is particularly present in Croa-

tian, algorithms that extract the core information from a knowledge written in the sentences of the definition for a given word from the online encyclopedia will be shown. That knowledge is then transcribed in an ordered or unordered sequence of information which is included as a domain of the word in the T-structure and serves for recognition and extraction of the knowledge from other sentences. That leads into one recursive learning and checking cycle, which then reaches a growing precision and accuracy of new vocabulary/thesauri in multiple iterations.

Dialectological extensions of the dictionary base

The Croatian language has three main dialects: Čakavian, Kajkavian and Štokavian, which began as separate language systems. Each of these three is divided into minor dialects, among which, within the same dialect, there are also considerable differences. The Čakavian and Kajkavian dialects differ most from Standard Croatian language. Moreover, the Northern Čakavian dialect is lexically closer to the Slovenian and the Southern Čakavian is lexically closer to the Štokavian dialect. A complete picture of Croatian language can be attained only by combining diachronical and synchronical study. Large amounts of linguistic data can only be processed electronically. Since there is still no single historical corpus of Croatian, one that would cover all three dialects, and since many of the research points (villages) for the Croatian language atlas have not yet been studied, the Croatian language and its history are known to us more or less fragmentarily.

The aim of the project Dialects of Makarska coast – diachrony and synchrony is to study the Štokavian dialect of Makarska coast from a dialectological, textological and sociolinguistic point of view. The data obtained through these efforts will be digitalized and eventually integrated into the framework of the Croatian language. The data obtained through field work will be coordinated with the data obtained from historical texts. For example, in the old texts a noun form *šćeta* has been regularly attested, which form is also spoken today in the Makarska coast (Brela, Podgora), as confirmed by our field data. A computer program will link this form with the standard-language form *šteta*. In order to achieve this, it is necessary to precisely define linguistic characteristics of the Makarska coast dialect and correlate them with the corresponding forms of the standard Croatian. For example, once we define the idiom of Makarska coast as šćakavian (i. e. that psl.*stj and *skj gave šć) and standard Croatian as štakavian (i. e. that psl.*stj and *skj gave št), the computer program will easily link the dialectal forms of gušćerica, šćucat and šćap with standard-language forms gušterica, štucati and štap. By linking these lexems on the phonological and morphological level we are linking them on the semantic level. This is the most used procedure by authors of dialectal dictionaries: in the glossary they usually attach meanings verified in the standard Croatian language to dialectal words. Another way of obtaining the meaning of a particular word is by inferring from

the totality of verified texts of certain area, in this case the Makarska coast: based on the context in which a particular word appears, its meaning can be determined and recorded. As such it completely corresponds to the meaning of the word in the standard language, narrows it, or alters it altogether.
The idea is that a research of the same kind is to be carried out on the whole of the Croatian language. Machine-generated data should present a true picture of the Croatian language, connect the history of the Croatian with its present and show all the changes that have taken place in the language.
In the process, the primary meaning of a word will be determined, the semantic field of each word will be identified, as the overlapping of these fields, their semantical convergence and divergence, all of which would create a respectable encyclopedic knowledge base.

The creation of a semantic domains and their usage

In the context of the SSF a semantic domains represent a set of words that are meaningfully and closely related to a specific word. Building of such domains can be done either manually or automatically. The manual creation of domains is certainly a more difficult process, because it involves a huge human effort. In this article we propose a new way of creating semantic domains based on the definitions of words derived from publicly available online sources. All definitions of words from the available online resources such as the Miroslav Krleža Institute of Lexicography's online encyclopedia or Croatian language portal, are part of the SSF. Beside the fact that all elements of definitions are interconnected, and form a new semantic network, they are also a part of a SSF lexicon, which enables expansion of such domains so long as words with definitions exist in the database. There are two main approaches to a semantic domains creation within the SSF, and both are performed in four steps. In the first step, the definition from an online resource is obtained for a given word, which is then processed with extraction algorithm. The first approach (like shown in Figure 3) extracts the subject, predicate and the object. The result is then lemmatized. The final result is a set of words that makes the semantic domain for a given word. Since every word has a number of properties that have an ontological (T-structure) attached to them, so the name/word itself is also a property, it was easy to make a difference. For example, it is enough to know the accent of the word (which is also one of the properties) or WOS/SOW of words neighbors together with a suitable morphosyntactic pattern to determine what the term is it about.

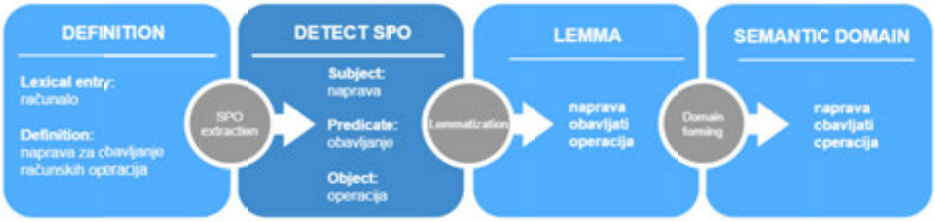


Figure 3. Creation of a semantic domain by extracting subject, predicate and object

The second approach differs only in the second step, where instead of the subject, predicate and the object extraction, definition elements are filtered using WOS marks (Figure 4). A user can define the tags the filter he will use, thereby automatically affecting the size and scope of the newly created semantic domain.

Figure 4. Creation of a semantic domain by WOS mark filtering

Such domains can be further recursively expanded with new elements resulting from the repeated process of extraction.

The main usage of such domains in the SSF is in the process of tropes detection (e.g. metaphors extraction). The algorithm that extracts metaphors from the text consists of three steps (Orešković et al. 2017). The first step is also the simplest because the metaphor is found based on its match within the repository. Stored and tagged multiword expressions are lemmatized and matched with the text. The second step is performed by using a syntactical and semantic patterns inside a defined virtual (semantic) domains, whose formation is described above. The last step is by detailed WOS/SOW analysis of each word in a sentence.

Conclusion

This research showed that creation of "semantic domains" can be done by using online encyclopedic articles. The Miroslav Krleža Institute of Lexicography's online encyclopedia (http://enciklopedija.lzmk.hr/) was used as an example. Also, complex programming support was created to achieve a domain creation. Two approaches are developed: through the open-class words (nouns, verbs, adjectives) and through the word function in the sentence (subject-object-predi-

cate) – dependency grammar. Because of the dialect richness of the Croatian language, program support for standardizing dialect forms to standard language forms was added. The created domains can be used to extract semantic information from unstructured text even at the most semantic level (stylistic figures: metaphor, metonymy, etc.)

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions which were helpful in improving the article.

Funding

This work has been supported in part by the Croatian Science Foundation under the project IP-06-2016.

References

Ackoff, RL. 1989. "From Data to Wisdom." *Journal of Applied Systems Analysis*.

Alberico, R. and M. Micco. 1990. "Expert Systems: For Reference and Information Retrieval."

Antoniou, G. and F. Van Harmelen. 2004. "A Semantic Web Primer."

Bosančić, B. 2016. "Proces Stjecanja Znanja Kao Problem Informacijskih Znanosti." *Libellarium: Journal for the Research of Writing,*.

Frické, M. 2009. "The Knowledge Pyramid: A Critique of the DIKW Hierarchy." *Journal of Information Science*.

Gruber, TR. 1993. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition*.

Jecić, Zdenko. 2013. *Studia Lexicographica.* [s.n.]. Retrieved May 31, 2017 (http://hrcak.srce.hr/114403).

Liew, A. 2007. "Understanding Data, Information, Knowledge and Their Inter-Relationships." *Journal of Knowledge Management Practice*.

Markučič, Joško and Klemen Govedić. 2013. "Morphological Generator of Croatian Language."

Melcuk, I. A. 1981. "Meaning-Text Models: A Recent Trend in Soviet Linguistics." *Annual Review of Anthropology* 10(1):27–62.

Mihalcea, Rada and Andras Csomai. 2007. "Wikify!: Linking Documents to Encyclopedic Knowledge." *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management* 233–42.

Orešković, M., M. Čubrilo, and M. Essert. 2016. "The Development of a Network Thesaurus with Morpho-Semantic Word Markups." *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity* 273–79.

Orešković, Marko, Marta Brajnović, and Mario Essert. 2017. "A Step towards Machine Recognition of Tropes." P. 71 in *Third International Symposium on Figurative Thought and Language*.

Pinjatela, Krešimir. 2001. "Hrvatska Riječ."

Pustejovsky, J. 1991. "The Generative Lexicon." *Computational Linguistics*.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. "Yago." *Proceedings of the 16th International Conference on World Wide Web - WWW '07* 697.

Šnajder, Jan and Petra Almić. 2015. "Modeling Semantic Compositionality of Croatian Multiword Expressions." *Informatica* 39(3):301.

Zesch, Torsten, Christof Müller, and Iryna Gurevych. 2008. "Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary." *Linguistics* 1646–52.

The e-Beliana Project

Tatiana Šrámková
The Society for Open Information Technologies
Piešťanská 2991/2, 010 08 Žilina, Slovakia
tana.sramkova@soit.sk

Miloš Šrámek
The Society for Open Information Technologies
Piešťanská 2991/2, 010 08 Žilina, Slovakia
milos.sramek@soit.sk

Viera Tomová
The Encyklopedic Institute of the Slovak Academy of Sciences
Bradáčova 7, 851 02 Bratislava, Slovakia

Summary

This paper introduces e-Beliana, a project of the Encyclopedic Institute of the Slovak Academy of Sciences and the Society for Open Information Technologies, a non-profit organization aimed at the promotion of free and open-source software, open data and open government in general. The main purpose of the collaboration is to publish any current and future content of the Encyclopædia Beliana, the first general Slovak-language encyclopedia, on the internet under a Creative Commons license. In order to accomplish this goal the whole process of preparing the Beliana articles had to be changed substantially. The new web-based editorial system e-Beliana is the topic of this paper.

Key words: encyclopedia, editorial system, open source software, data analysis and conversion

Introduction

The Encyclopædia Beliana is a Slovak-language general encyclopedia. It was first published in 1999 by the Encyclopedic Institute of the Slovak Academy of Sciences and thus far contains 55,000 articles and 12,000 illustrations in eight printed volumes covering the letters A – K. The plan is to publish the last volume in 2031. The abovementioned facts and the rise in popularity of internet encyclopedias inspired the institute to change the current editorial process and make the content of Beliana available on the internet.

Therefore, the institute started a common project called e-Beliana with the Society for Open Information Technologies, a Slovak non-profit organisation aimed at the promotion of free and open-source software [SOIT]. Both parties

signed a contract according to which SOIT will create an open source based editorial system for free. This system will make the editorial process more efficient and will allow easy transfer of content to the internet. According to the contract, the Encyclopedic Institute will publish any existing and future content under the Creative Commons SA BY licence [CC].

The old and the new workflow

The existing editorial workflow was designed about 20 years ago using the technical means of that era. It was designed for the production of the printed encyclopedia with no intention to publish its content online. The editorial workflow consisted mainly of exchanging of MS Word documents between authors and editors and among editors themselves. The proofreading was done entirely on paper with the document files being subsequently modified (error prone). Articles were proofread in batches of about 200, with each batch containing articles from different categories. A batch could be proofread by only one editor at a time. This approach resulted in a large time gap (even many months) between the preparation of the original text and the proofreading.

The aim of the new web-based editorial system e-Beliana is to overcome the abovementioned drawbacks and to prepare articles simultaneously for the printed and the internet version of Beliana. The new system should enable independent editing of articles which should significantly shorten processing time. The system should be flexible enough to implement the currently used workflow and to easily allow modifications according to future experience and demands.

The basic requirement of such an editorial system is that in a given moment only one author or editor can modify an article. This requires both horizontal and vertical specification of access rights. Horizontally, access rights must be granted to different authors and editors according to the category of the article. Examples of categories are “Mathematics”, “Zoology” or “German literature” (there are currently more than 600 such categories in Beliana). One editor is responsible for several categories, and there may be several authors for a single category. Vertically, an article flows through a sequence of stages: with author, editor, consultant, etc. Therefore, a user (authors, editors, senior editors, etc.) should not only be able to edit articles in a specific stage but also to change this stage to the next one in the workflow.

Editorial system e-Beliana

Based on the abovementioned requirements we analysed various available open-source solutions. From among them we selected the *Drupal* CMS [Drupal] with its modules *Workbench Access* and *Workbench Moderation* which made the implementation of both horizontal and vertical access rights possible. Subsequently, we implemented a first version of the software, imported the published articles and made the system available to users. Even this preliminary

version provided them with incomparably faster search functionality in the texts of published articles as compared to the earlier option – searching in hundreds of MS Word or pdf files or browsing through the printed book. The remaining functionalities of a full editorial system have been implemented gradually since then in close collaboration with the editors. This approach would not have been possible with a proprietary system – in order to try something out, we simply downloaded and installed the software (mainly Drupal modules extending its basic functionality) from the internet, no negotiations and license purchases were necessary.

Several special parts of the software that provided functionality not readily available in Drupal, were later implemented.

Currently the e-Beliana workflow (Figure 1) uses 27 article stages (e.g. proposed, accepted, with author, with editor, etc.) and 14 user roles (e.g. author, editor, consultant, etc.). Users can search and process available articles in the editorial system by using *views* (implemented by means of the Drupal's *Views* module), which were designed according to the demands of their role. Currently, there are about 20 views which take the access rights provided by the Workbench Moderation and Workbench Access modules into account and about 20 general views which either provide an overview of all Beliana articles to any user or provide users with any necessary special rights For example, editors can access a view which enables them to list articles waiting to be edited or a view that allows them to assign articles to authors for editing.

Articles can be opened for modification from a view. Text can be edited in an embedded editor [CKEditor], which in addition to basic formatting enables users to edit mathematical and chemical formulas in LaTeX notation and to track editorial changes by storing information about the date and author of the change. Formulas are edited and displayed in Beliana using the MathJax tool [MathJax]. The change tracking capability is provided by the CKEditor's LITE module [LITE]. Changes can be viewed not only in the editor window, but in all stored revisions of an article.

Exporting articles to web and printing

Articles which have passed the editorial workflow are ready for publication – in our case both publication on the internet and in book form. A Beliana website (not accessible to public yet) has been developed independently of the editorial system. Contrary to the complexity of the editorial system which uses tens of additional Drupal modules, during development of the Beliana site, we mainly focused on simplicity, robustness and speed. Editiorial system content is synchronized to the public website using the REST API.

While the article text is exported to the internet site without changes, it must be modified for the printed version. The major text modifications are the removal of hyperlinks and the abbreviation of common words and article title occurrences in text. The abbreviation tool is based on the stemming and subsequent

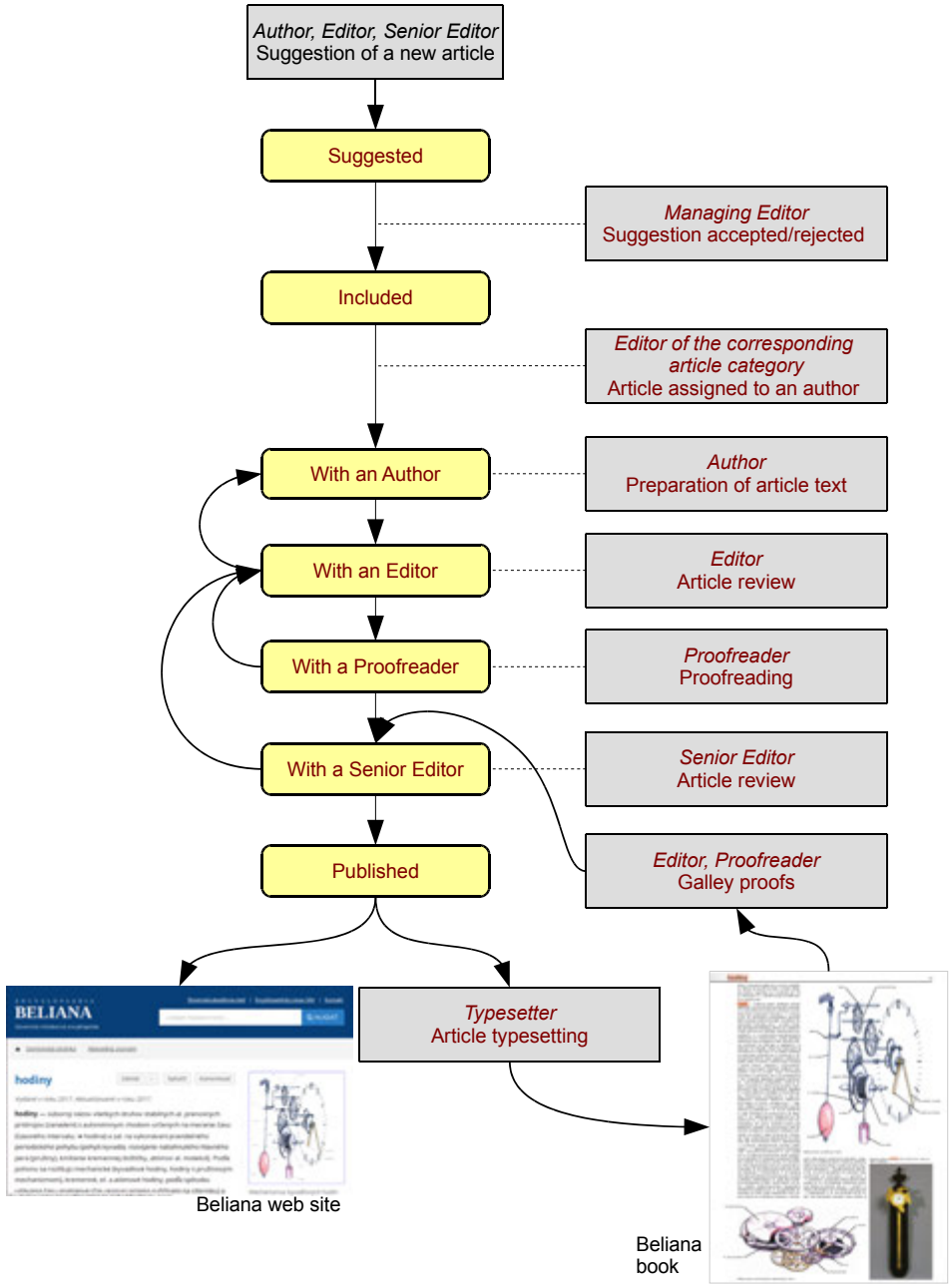


Figure 1. Simplified editorial workflow. Rounded boxes represent editorial stages, sharp boxes represent actions (with user role in italics).

automated inflection of words [Garabík, Radovan] and enables the abbreviation of single words as well as phrases. Beliana books are typeset in InDesign using the MathMagic tool for rendering of LaTeX-based mathematical and chemical equations.

Importing content from existing sources

Each article was imported either from the corresponding text document (published articles) or from a general alphabetic index (future articles). The general alphabetic index is a list of articles to be included in Beliana. It is represented by a set of spreadsheet files containing information such as title and category. The published articles had been created in different versions of the MS Word text editor. Neither the index nor the article document files were edited and maintained with the intention of further machine processing. Both contained numerous random comments and instructions embedded directly in the text. It often became unclear from the point of view of programmatic analysis what was the real text and what was the comment or instruction. The MS Word document files for the first two encyclopedia volumes were not even available; these texts had to be extracted from PageMaker (volume 1) and QuarkExpress files (volume 2).

The development of import tools was not straightforward and required many cycles in which they were refined. This was caused not only by the aforementioned random comments and instructions, but also by the fact that we strived to extract certain information from loosely structured human readable text, as was the case of detection of names and surnames or dates of birth and death.

The import tools were implemented using *bash* as a basic scripting tool, *LibreOffice* for conversion of files in MS Office formats for further processing, the *aspell* spellchecker for detection of misspelled words in the Slovak and Czech languages and the stream editor *sed* for batch replacement of misspelled characters and words. Our own tools were written in *python*.

Among others, the import tools included the following operations:

1. *Correction of incorrectly recognised characters*. The oldest documents were from the 90s and included special and accented characters which were not correctly recognised by today's software.
2. *Text segmentation in articles and detection of the article title*.
3. *Detection of personal names and surnames*. In articles related to persons, surnames were used as an article title (i.e., it was typeset in bold). In the internet version of Beliana we prefer to have article titles with both the name and the surname. In order to detect the first names we implemented a rule based system, taking into consideration different kinds of information (for example, the title was followed by a comma, the article was not in a category related to geography etc.)
4. *Detection of the article category*. Information about the category of the article was available from two sources: the general article index spread-

sheets and one of the early versions of the article texts. A rule-based approach taking advantage of fuzzy matching was used, in order to detect the category of two or more articles with the same name or articles with multiple categories created by merging other articles.

5. *Detection of references to other articles*. A reference to another article was marked in the Beliana text by an arrow which was followed by the title of the referenced article. This title, however, was often inflected and abbreviated and was often not separated from the subsequent text. Therefore, we always considered all words following the arrow till the first non-letter and non-space character (except for a dot) to be title candidates. Words of this string were subsequently lemmatized [Garabík, Radoslav] and an article with the most similar title was looked up in the list of all articles.
6. *Detection of illustration captions*. Illustration captions were included directly in the article text and were separated from it by using the words "Illustration text". Since these were often abbreviated and formatted in different ways, not all captions were detected.

Using the analysis tools, we processed and prepared about 270 document files with the text of the published articles (more than 6 mil. words, 56,000 articles and 12,000 illustrations) and 12 spreadsheet files of the general article index with more than 110.000 entries for import. The analysis tools developed in this process were not perfect, but when compared to purely manual processing they saved a substantial amount of work.

Conclusion

In this paper, we have briefly introduced the e-Beliana editorial system based on the Drupal CMS and its modules. The open-source character of Drupal was the most important feature for a successful realisation of the project. The source code is available in the GitHub repository of the Encyclopedic Institute of the Slovak Academy of Sciences [GitHub]. As the new editorial system has only been used for 3 months, it is too early to summarise its advantages and drawbacks. The public web site will be available in the beginning of 2018.

References

Garabík, Radovan. Slovak morphology analyzer based on Levenshtein edit operations. In: Proceedings of the WIKT'06 conference. Laclavík, M.; Budinská, I.; Hluchý, L. (ed.) Bratislava: Institute of Informatics SAS, 2006, 2 – 5

SOIT. About us. http://soit.sk/sk/in_english/who-we-are (June 30, 2017)

CC. Creative Commons. https://creativecommons.org (June 30, 2017)

Drupal. http://www.dupal.org (June 30, 2017)

CKEditor. http://ckeditor.com/ (June 30, 2017)

MathJax. https://www.mathjax.org/ (June 30, 2017)

LITE. https://www.loopindex.com/ (June 30, 2017)

GitHub. https://github.com/enu-sav (June 30, 2017)

An insight into online encyclopaedias for children and young adults

Cvijeta Kraus
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
cvijeta.kraus@lzmk.hr

Nataša Jermen
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
natasa.jermen@lzmk.hr

Zdenko Jecić
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
zdenko.jecic@lzmk.hr

Summary

Online encyclopaedias and encyclopaedic portals for children and young adults serve as useful tools for acquiring knowledge. This paper presents the results of the analysis of functionalities of three online encyclopaedias for children and young adults, Britannica Kids, Q-files and KidzSearch Encyclopedia in the digital environment. A set of basic criteria for the evaluation of usability of online encyclopaedias intended for young users were established. These criteria could serve as the basis for the definition of a concept for production of online encyclopaedic works for children and young adults.

Key words: children's encyclopaedia, young adult's encyclopaedia, encyclopaedia's evaluation criteria, educational website

1. Introduction

Children and young adults are strongly oriented towards information technologies. The primary sources of information in their world are often (although not exclusively) digital, which is quite different from any generation prior.[1] Therefore, there is a need to ensure that they are furnished with the safe and trusted environments, as well as with the digital content which is trustworthy and suita-

[1] Flanagin Andrew J.; Metzger Miriam J. Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility. // *MacArthur Foundation Series on Digital Media and Learning*. Cambridge, MA: The MIT Press, 2008, p. 6

ble to their needs. As encyclopaedias have always formed an important part of the infrastructure for learning in schools,[2] online encyclopaedias and encyclopaedic portals could represent a very important component of educational e-infrastructure serving as the didactic tools for electronically supported learning and teaching.[3] As such they should be considered as a specific type of educational websites – the sites created with the purpose of educating their users about a certain topic or the sites that are adapted for schools in accordance with curriculum subjects.

General encyclopaedias are intended for a wide range of users, tending to avoid explaining well-known, as well as overly professional terms. Consequently, it is considered that they are adapted for adult users who have the necessary average pre-knowledge for using encyclopaedias. Since this does not cover the needs of the very young audience, specialised encyclopaedic works for children and young adults are being published. One of the first encyclopaedic works for children was *Orbis Pictus (Visible World in Pictures),* which was published in Bratislava in 1658 by a Czech teacher Johann Amos Comenius. Encyclopaedias for children[4] are mainly targeted at primary school children, from 6 or 7 to 12 or 14 years old. These encyclopaedias mostly contain a smaller number of review articles that elaborate a larger number of topics. The articles can be organised alphabetically or thematically. The writing style is simple, as well as the language, and the articles' structure is not strictly determined. Special attention is paid to the illustrations, that are usually more common compared to other encyclopaedias, as well as to the attractive graphic editing. These elements are supposed to attract the attention and interest of younger users. Encyclopaedias for young adults[5] are targeted at users above 12 years old. They encompass a larger

[2] Sundin, Olof; Haider, Jutta. The networked life of professional encyclopaedias: Quantification, tradition, and trustworthiness. // *First Monday*, 8 (2013) 6.

[3] Croatian national educational strategy, for example, points out that e-learning and e-education provide access to updated and current multimedia and interactive teaching materials and enable the use of the repositories of didactic material, digital libraries, archives and museums. The development of educational e-content includes creation of open repositories of knowledge and didactic tools (for example digitised books and lexicons). Strategy of education, science and technology (in Croatian), Official Gazette 124/2014.

[4] Examples of traditional printed encyclopaedias for children: *My First Britannica* (Chicago: Encyclopædia Britannica, 2004; 13 volumes, 525 articles, 1288 pages, 800 illustrations; age group 6 to 10); *Britannica student encyclopedia* (Chicago: Encyclopædia Britannica, 2007; 16 volumes, 2250 articles, 2900 pages, 2700 illustrations; age group 7 to 12.); *Der Kinder Brockhaus in drei Bänden* (5th edition Mannheim: Bibliographisches Institut & F. A. Brockhaus, 2006; 3 volumes, 1000 articles, 676 pages, 2000 illustrations; age group 7 to 12).

[5] Examples of traditional printed encyclopaedias for young adults: *Compton's by Britannica* (Chicago: Encyclopædia Britannica, 2008; 26 volumes, 37 000 articles, 11 000 pages, 23 000 illustrations; age group 10 to 17); *Der Jugend Brockhaus in drei Bänden* (6th edition, Mannheim: Bibliographisches Institut & F. A. Brockhaus, 2005; 3 volumes, 10 000 articles, 1200 pages, 2000 il-

number of articles that elaborate on narrower terms when compared to encyclopaedias for children. They are adapted for school and homework assignments, and often bring references for additional reading.
Information and communication technologies (ICT) have given rise to the specialised internet sources for younger audience, for example specialised search engines, fun sources, sources of knowledge etc. Online encyclopaedias for children and young adults are sources that cover many topics, and as such they represent comprehensive sources of knowledge. One of the first online encyclopaedias for children and young adults was *The World Book Encyclopedia*[6] published on the internet in 1998. New encyclopaedic sources of knowledge have been continuously developed, for example *Q-files, KidzSearch Encyclopedia,* and especially *Fact Monster, Encyclopedia Smithsonian, Encyclopedia.com, Britannica Kids*.[7]

2. Scope of research

This paper aims to explore the functionalities of online encyclopaedic portals for children and young adults in the digital environment. Basic criteria for the evaluation of usability of online encyclopaedias intended for children and young adults will be elaborated. The assumption is that established criteria could serve as the basis for the definition of a concept for production of online encyclopaedias intended for young users. The research was limited to the websites that, despite of their unspecified format, could be considered encyclopaedias, since they have clearly articulated articles that contain a title (article's name) and textual explanation of the subject named in the article's title. This definition meant that the analysis embraced only a small amount of different online sources for children and young adults aimed at acquiring knowledge through fun or at helping to follow curriculum subjects.

3. Analysis of online encyclopaedias for children and young adults

Three online encyclopaedias for children and young adults were chosen for the analysis: *Britannica Kids, Q-files* and *KidzSearch Encyclopedia.*

3.1. Encyclopædia Britannica for children and young adults – Britannica Kids

Encyclopædia Britannica[8] was first published in 1768 in Edinburgh, Scotland. By its eleventh edition released in 1911, it was hailed as the greatest encyclope-

lustrations; age group 12 and above); *Svijet oko nas, enciklopedija za djecu i omladinu* (Zagreb: Školska knjiga, 1960; 2 volumes, 600 pages; age group 10 to 16).

[6] World Book Online, http://www.worldbookonline.com (3.6.2017).

[7] Rain, Ella. Childrens Online Encyclopedia, http://childrens-books.lovetoknow.com/Childrens_Online_Encyclopedia (26.6.2017).

[8] Encyclopaedia Britannica, https://www.britannica.com (2.6.2017).

dia in the world.[9] In 1994 *Encyclopaedia Britannica* published its first CD-ROM and the online edition (*Britannica online*[10]), becoming the first printed encyclopaedia whose digital version appeared on the internet. Having a reputation of being the universally acknowledged outstanding reference work, it soon became the leading publisher of digital encyclopaedic editions, particularly after it stopped publishing printed editions in 2012. At present, the main *Encyclopædia Britannica*'s website (britannica.com) offers access to other specialised websites of the same publisher. One of those websites is the encyclopaedic portal for children and young adults *Britannica Kids.*[11]

The main goal of *Britannica Kids* is to provide easily accessible safe and trusted content to children and teenagers in order to support their homework needs based on a variety of curriculum subjects and standards. The content is gathered from the "great intellects across the globe - including leading educators, Pulitzer Prize winners and Nobel laureates" and edited by a skilled in-house editorial staff.[12]

Targeted for children and young adults from kindergarten to high school and beyond, this website is divided into three reading levels that are adapted to a certain age-group and expected pre-knowledge: Britannica KIDS – up to grade 5 (up to 11 years old); Britannica STUDENTS – grade 6-8 (from 11 to 14 years old); Britannica SCHOLARS – grade 9 and up (above 15 years old). After registration the user selects the desired level (separate websites) on the homepage of the portal.

Certain functionalities are mutual for all three reading levels. The homepage of each of these websites contains a simple search engine, as well as a visually smaller more advanced search engine that provides a drop-down list of possible results. Presentation of search results is the same on all three websites: the list of articles that contain a searched topic, multimedia elements and links to additional sources. Upon opening a selected article the user can easily access the same article on the remaining two reading levels. Furthermore, it is possible to share the article via e-mail, to print, cite and translate it. Audio reproduction of the content is enabled.

3.1.1. Britannica KIDS

Britannica KIDS[13] homepage contains interactive content (a quiz) that enables children to initiate their exploration. Besides searching with the search engine,

[9] Auchter, Dorothy. The evolution of the Encyclopaedia Britannica: from Macropaedia to Britannica Online // *Reference Services Review*, 27 (1999) 3, p. 291.

[10] At the time at: http://www.eb.com/eb.html

[11] Britannica Kids, http://kids.britannica.com/ (2.6.2017).

[12] Britannica Kids, http://kids.britannica.com/about (2.6.2017).

[13] Britannica Kids, Kids, http://kids.britannica.com/kids (2.6.2017).

there is a possibility for browsing of certain sections (Articles, Images&Video, Biographies, Animal Kingdom, World Atlas, Dictionary), that contain thematically or alphabetically organised content.
This encyclopaedic website contains very simple content, adapted for the youngest users. Articles are divided into chapters (the user can decide which chapter to open), accompanied with images and links to the related articles. As well, there is a section *Did you know?* that highlights out article's interesting facts.

3.1.2. Britannica STUDENTS & Britannica SCHOLARS
The homepage of *Britannica STUDENTS*[14] & *Britannica SCHOLARS*[15] brings a lot of additional content (presentation of current events, additional sections such as *Can you guess? Did you know?*), but the search engine is always visible on the page. Once search results have been displayed there is a possibility to carry out additional advanced search. As is the case for *Britannica KIDS*, browsing of main sections, that contain thematically or alphabetically organised content, is enabled. At these two levels, searching of sections' content is more advanced and enables different types of queries.
The content of these two reading levels differs with respect to comprehensiveness and the presentation of multimedia elements. *Britannica STUDENTS* offers a simpler content, adapted to younger users. Content at the *Britannica SCHOLAR* level is actually provided from *Encyclopaedia Britannica* (articles are signed by the contributors), although its presentation is more simplified. Articles at both reading levels are divided into chapters, which are listed in the left panel and are easy to navigate accompanied with multimedia elements and related resources (articles, primary sources and e-books, and websites).

3.2. Q-files – The Great Illustrated Encyclopedia

One of online encyclopedias designed especially for children is *Q-files – The Great Illustrated Encyclopedia*[16] by Orpheus Books Limited, based in Oxford, England. Established in 1993, this publishing house is one of the world's leading producers of children's non-fiction and reference books. The site *Q-files* was launched in 2015. primarily for children aged 8 to 14-years-old. As of June 2017, Q-files is now a subscription service.[17] The content of *Q-files* has been drawn from Orpheus's reference book archive and prepared by specialist children's writers and experienced editorial team, under the expert guidance of the

[14] Britannica Kids, Students, http://kids.britannica.com/students (2.6.2017).

[15] Britannica Kids, Scholars, http://kids.britannica.com/scholars (2.6.2017).

[16] Q-files – The Great Illustrated Encyclopedia, https://www.q-files.com/home/ (3.6.2017).

[17] ClassConnect, http://connect.learnpad.com/content/activity.cfm?id=286650 (8.6.2017).

consultants who worked on the original titles. Users can therefore be assured of the highest standards of accuracy and reliability.[18]
The homepage of this website is visually designed for children and provides access only to the selected content. For the complete access to the content a user needs to register as a home (individual) user or as a school user. The homepage for individual users doesn't provide access to the search by topics and doesn't contain a simple search engine. The individual user can access these functionalities only after opening a selected article on the homepage. School users are equipped with more elaborated homepage that provides search by topics and simple search engine. Every article attaches active links to the similar content on the same site and a Q-facts section with main information about the article. *Q-files* encyclopedia contains a lot of multimedia elements (images, tables, timelines) and provides a print option. The site has a presence on the social networks.

3.3. Wikipedia for children – KidzSearch Encyclopedia

Wikipedia[19] is a web-based, multilingual and free-access encyclopedia developed in 2001. Based on anonymous collaborative editing, Wikipedia soon became the largest and the most popular encyclopaedia in the world. The *Kidz Search Encyclopedia*[20] is based on Simple English Wikipedia, which is similar to regular Wikipedia, but written at a level much more appropriate for children. Editorial staff manually approve all the entries and may also add unique original articles themselves. It is important to point out that unauthorized staff cannot change the pages on KidzSearch site.[21]
As in Wikipedia, content of *KidzSearch Encyclopedia* is free and accessible without registration. The homepage provides a simple search engine and a list of topics to search, accompanied by a few representative articles. The list of topics is also present in the left panel, where each topic is linked to an article that explains it. Search by a search engine takes the user to another website kidzsearch.com, which provides the results from other sources besides *Kidz Search encyclopedia* (other web sites, images, videos, facts, news, games etc.). To access the *KidzSearch encyclopedia* content, it is necessary to select the section *wiki* at the website menu. Articles in *KidzSearch encyclopedia* are organized in the same way as in Wikipedia; they are divided into chapters, contain

[18] Q-files – The Great Illustrated Encyclopedia https://www.q-files.com/about-q-files/ (3.6.2017).

[19] *Wikipedia:* The Free Encyclopedia, Wikimedia Foundation Inc., http://www.wikipedia.org, (29. 7.2008).

[20] KidzSearch Encyclopedia, http://wiki.kidzsearch.com/wiki/Main_Page (3.6.2017).

[21] KidzSearch Encyclopedia, http://wiki.kidzsearch.com/wiki/KidzSearch_Encyclopedia:About (9.6.2017).

images, links to other articles within encyclopaedia as well as to the external sources, references and an information box with main facts.

3.4. Comparative analysis of the content organization and presentation

The most important part of a website is its content that has to be engaging, relevant and appropriate to the audience.[22] In the case of encyclopaedias, it is important to ensure that the information in the articles is structured and presented according to the encyclopaedic principles.[23]

We analysed content organisation and presentation of the encyclopaedic entries in *Britannica Kids, Q-files* and *KidzSearch Encyclopedia.* Assessing the content in terms of its reliability was beyond the scope of this research. Nevertheless, it is important to stress that every encyclopaedic website should provide accurate, up to date and appropriate information for its end-users.

Based on the comparative analysis of several encyclopaedic entries in each of the analysed encyclopaedias, we established that there are certain patterns in the way that content is organised and presented. We will show the main characteristics in content organisation and presentation in each of the analysed encyclopaedia using the example of the article *Croatia:*

- *Britannica KIDS.*[24] Very simple content that doesn't contain a clear definition. The article is divided into 5 chapters, containing 4 images, an audio clip of the Croatian national anthem and a link to related articles.
- *Britannica STUDENTS.*[25] More extensive content without a clear definition in the introduction. The left panel offers a list of chapters, a *Quick Facts* box that contains a compilation of the key information about the state and the links to other related resources. The article is divided into 5 chapters, containing 5 images and an audio clip of the Croatian national anthem.
- *Britannica SCHOLARS.*[26] The most extensive content that contains a definition. It is divided into 8 chapters, with additional subchapters, containing 30 images, 2 video clips and an audio clip of the Croatian national anthem The list of chapters is placed in the left panel together with a detailed *Quick Facts* box and the links to other related resources. As previ-

[22] HHS Web Standards and Usability Guidelines, https://guidelines.usability.gov/guidelines/1 (5.6.2017).

[23] Jecić, Zdenko. Enciklopedički koncept u mrežnom okruženju. // *Studia lexicographica*, 7 (2013), 2 (13), p. 113.

[24] Britannica Kids, Kids, http://kids.britannica.com/kids/article/Croatia/345673 (7.6.2017).

[25] Britannica Kids, Students, http://kids.britannica.com/students/article/Croatia/273859 (7.6.2017).

[26] Britannica Kids, Scholars, http://kids.britannica.com/scholars/article/Croatia/110562 (7.6.2017).

ously mentioned, the article's content in *Britannica SCHOLARS* is the same as the one in britannica.com[27].

- *Q-files*[28]. Simple content with a short definition. The article contains a small information box with main facts about the state. Content is divided into 3 chapters and contains 10 images.
- *KidzSearch*[29]. Simple content with a short definition The information box is organised in the same way as in regular Wikipedia. Content is divided into 3 chapters and contains 2 images.

4. Evaluation of online encyclopaedias for children and young adults

At a time of numerous websites for children and young adults that offer encyclopaedic content, learning through fun or help with learning, parents, teachers and librarians should feel confident about estimating which online encyclopaedic websites suit the needs of a young user, in terms of their age-appropriate content and usability. Therefore, there is a need for a set of criteria for the evaluation of encyclopaedic websites for children and young adults. While evaluation of internet resources has been tackled by a significant number of researches and instructions,[30] online encyclopaedic resources have been mostly overlooked in this respect so far. Jecić and Boras, for example, designed special criteria for reliability testing procedure of virtual encyclopaedias based on lexicographic characteristics.[31]

4.1. Encyclopaedias as educational websites

Being tools for learning and education, online encyclopaedias for children and young adults represent a specific type of educational websites. Nowadays, educational websites are becoming progressively similar to electronic textbooks in terms of both content and presentation. They are becoming more advanced, bringing together rich, scholarly material and employing many of the same design elements as electronic textbooks, such as internal and external hyperlinks

[27] Encyclopaedia Britannica, https://www.britannica.com/place/Croatia (8.6.2017).

[28] Q-files The Great Illustrated Encyclopedia, https://www.q-files.com/geography/europe/croatia/ (7.6.2017).

[29] KidzSearch Encyclopedia, http://wiki.kidzsearch.com/wiki/Croatia (7.6.2017).

[30] Gi-Zen Liu; Zih-Hui Liu; Gwo-JenHwang. Developing multi-dimensional evaluation criteria for English learning websites with university students and professors. // *Computers & Education.* 56 (2011) 1; pp. 65-79; Ping Zhang; Gisela M. von Dran. Satisfiers and dissatisfiers: A two-factor model for website design and evaluation. // *Journal of the Association for Information Science and Technology*, 51 (2000) 14; pp. 1253-1268.

[31] Jecić, Zdenko; Boras, Damir. Fotografija u virtualnim enciklopedijama – razrada kriterija evaluacije internetskih sadržaja. // *Proceedings – 10. međunarodno savjetovanje tiskarstva, dizajna i grafičkih komunikacija Blaž Baromić* / Bolanča, Zdenka; Mikota, Miroslav (ur.). Zagreb, Senj: Grafički fakultet Zagreb; Matica hrvatska Senj, 2006, pp. 87-92.

and multimedia.[32] Since the internet has become an important feature of the learning environment, it is necessary to develop functional, reliable and trustworthy educational websites with appropriate content presentation, which would be adjusted to the young user's needs. According to Loranger and Nilsen, teenagers expect websites to be easy to use and to let them accomplish their tasks. They also prefer short text with words that they understand, short sentences and paragraphs. The main point of a website for teenagers is to teach them something new and keep them focused on a goal. The authors conclude that a good website for children must balance many elements like amount of text, multimedia elements, length of pages.[33] Wing-Shui showed that, from the teachers' point of view, during the process of developing a high-quality educational website, web designers should put a high emphasis on ease of users' browsing experience by providing a good web navigation system. Secondly, the educational website should be appealing by appropriately manipulating multimedia elements including color, graphics, fonts and typography, which should enhance the effectiveness of achieving educational purposes of the website.[34] In their study into the usability of e-encyclopaedias, Wilson et al. showed that school group users were shown to be less successful at completing tasks involving cognitive skills and at extracting relevant information from highly interactive, information abundant web sites, and there was some indication that they had a lower tolerance for unfamiliar design.[35] By applying educational multimedia materials for training purposes the user should be able to understand and memorise given content. The goal is to develop efficient and high-quality multimedia content that encourages active cognitive processing and leads to the meaningful learning through creative problem solving.[36]

Besides setting the primary goal of a website (which determines its audience, content, function and look), every website should contain certain elements that are necessary for its successful functioning. One of the most comprehensive compilation of these elements is collected in *The Research-Based Web Design*

[32] Wilson, Ruth; Shorteed, Julie; Landoni, Monica. A Study into the Usability of E-encyclopaedias. // *Proceedings of the 2004 ACM symposium on Applied computing.* Nicosia, Cyprus, 2004, pp. 1688-1692.

[33] Loranger, Hoa; Nilsen Jakob. Teenage Usability: Designing Teen-Target Websites, Nielsen Norman Group, 4 February 2013, https://www.nngroup.com/articles/usability-of-websites-for-teenagers/ (10.5.2017).

[34] Wing-Shui, Ng. Critical design factors of developing a high-quality educational website: Perspectives of preservice teachers. // *Issues in Informing Science and Information Technology*, 11 (2014), p. 105.

[35] Wilson, Ruth et al. (2004), ibid., p. 1691.

[36] Mateljan, Vladimir; Širanović, Željko; Šimović, Vladimir. Prijedlog modela za oblikovanje multimedijskih web nastavnih sadržaja prema pedagoškoj praksi u RH. // *Informatologia,* 42 (2009) 1, p. 39.

& Usability Guidelines (Guidelines), that was developed by the U.S. Department of Health and Human Services (HHS) in partnership with the U.S. General Services Administration and published in 2004.[37] This manual, containing 209 guidelines, represents innovative, research-based approach that should result in highly responsive and easy-to-use websites for the public.

4.2. Criteria for evaluation of usability of online encyclopaedias for children and young adults

Based on the results of the researches explained in the previous chapter and the analysis of online encyclopaedias *Britannica Kids, Q-files* and *KidzSearch Encyclopedia*, we found the main elements that should be taken into consideration when developing a well-designed encyclopaedic website. Accordingly, we established a set of basic criteria for evaluation of usability of encyclopaedic websites intended for children and young adults.

When evaluating the usability of mentioned encyclopaedic websites, the following elements should be considered:

- **Content.** Apart from being reliable and adapted to the needs of end-users, in this case children and young adults, the content of an encyclopaedic website should be well-organised and presented in a way to facilitate understanding of information.
- **Search engine.** The search engine should be easy to use and should allow users to search the entire site and to be successful when searching. More advanced search engine options provide a drop-down list of possible results that should facilitate finding the relevant one, as children are often not sure how to construct a search query. In addition, encyclopaedic websites should provide a possibility to search the content by topics as well as alphabetically. The search engine should be present on each page, including the homepage.
- **Homepage.** The homepage should be simple, of limited length, and should contain all the major options available at the site along with the information about the site. Furthermore, it is important to enable access to the homepage from every page in the site.
- **Links.** Links are very important element in the construction of online encyclopaedic content. They enable users to access other pages with related content on the same site (internal links) or on a different site (external links). In both cases they should be active, providing a useful and reliable

[37] Research-Based Web Design & Usability Guidelines, Official U.S. Government Edition. 2004. (https://www.usability.gov/sites/default/files/documents/guidelines_book.pdf). Updated version of Guidelines was published in 2006. U.S. Department of Health and Human Services (HHS) put them into a database for easier access in 2012. During 2013. HHS presented a new design of their web site and started the process of updating the Guidelines. (https://guidelines.usability.gov/).

content. They should be well-marked, underlined and easy to notice. External links should be isolated from the main content of the site.

- **Web site navigation.** The website should provide users useful navigational options, such as feedback on their location and link to destination pages. It is important that users can locate where they are in the site, return to the homepage, see the query history and easily proceed to the next activity.
- **Multimedia elements.** In order to facilitate learning and understanding of the content, images, tables, graphs and multimedia elements should be used appropriately. They should supplement the site's content and shouldn't unnecessarily distract users. As, well they should be accompanied with accurate labelling.
- **Content sharing.** The site should provide a print option and sharing of the content by email or, in case of teenagers, through social networks. In addition, it is important that an encyclopaedic website provides a possibility to leave a comment to editors about the presented content.
- **Online working space.** Specialised websites, in this case encycloapedic websites for children and young adults, should provide to users the possibility to create their own working space which would include saving the links to the content, creating their own content and communicating with editors and other users.

5. Evaluation of Britannica Kids, Q-files and KidzSearch Encyclopedia

We analysed the compatibility of *Britannica Kids, Q-files* and *KidzSearch Encyclopedia* to the established criteria (Table 1).

The analysis showed that all three encyclopaedias, with their simple and shorten content, are adapted for their end users, that is kids and teenagers. In some cases the lack of definition in the articles aimed at younger population is supposed to facilitate the didactic approach, while in the case of *Britannica SCHOLARS* the content, as expected, is longer and more detailed. *Britannica Kids* and *Q-files* provide well-organised content, while *KidzSearch* provides a very elaborated information box, similar to the one in regular Wikipedia, which is not simplified and adjusted to young users' needs. All the analysed encyclopaedic websites have a simple search engine, present on each page. In addition, *Britannica Kids* and *KidzSearch* contain well-designed search engine, with *Britannica Kids* being the only encyclopaedia that enables alphabetical searching. As for the homepage, *Britannica Kids* has developed a homepage at all three levels of reading, with search engine and additional interactive content. *KidzSearch* homepage is simple but without any interactive content, while *Q-files* homepage contains small amount of randomly selected articles and doesn't contain the search engine. Active and well-marked links are present in all three encyclopaedias, with *Q-files* being the only one that doesn't provide any external links.

KidzSearch external links did not prove to be a reliable source. Multimedia elements are well presented and elaborated in *Britannica Kids* and *Q-files*.

Table 1: Compatibility of *Britannica Kids, Q-files* and *KidzSearch Encyclopedia* to the criteria for evaluation of online encyclopaedias for children and young adults

Criterion	Britannica Kids	Q-files	KidzSearch
Content			
Well-organised content	YES	YES	NO
Search engine			
Well-designed search engine	YES	NO	YES
Enabled search by topics	YES	YES	YES
Enabled alphabetical search	YES	NO	NO
Search options on each page	YES	YES	YES
Homepage			
Simple Homepage	YES	NO	YES
Enabled access to the Homepage	YES	YES	YES
Links			
Active links	YES	YES	YES
Well-marked links	YES	YES	YES
External links	YES	NO	YES
Web site navigation	NO	NO	NO
Multimedia elements	YES	YES	NO
Content sharing			
Enabled content sharing	NO	NO	NO
Enabled print option	YES	YES	NO
Online working space	NO	NO	NO

A deficiency that all three encyclopaedias showed was a poor web site navigation, which did not always allow users to find and access information effectively and efficiently. Content sharing is enabled only by print options (except for *KidzSearch*), and in case of *Britannica Kids* by e-mail, but not through social networks or by leaving comments to the editors. Furthermore, none of the analysed websites provide the option for creating a personal online working space, which could facilitate and encourage young users' learning activities.

6. Conclusion

Encyclopaedias have always played an important role in the process of acquiring knowledge and as such have strived to be adapted to the needs of their users. This research showed that well-established specificities regarding content organisation and presentation in traditional printed encyclopaedic works for children and young adults, which differentiate those encyclopaedias from ones targeted at adult population, exist also in the era of digital media, when the majority of encyclopaedic works are published online. Since young population tend to search for information through different online and internet services, greater attention should be given to creating of functional online encyclopaedic

content intended for young users. A set of basic criteria for the evaluation of usability of online encyclopaedias for children and young adults was established, which could also serve as the basis for the development of a concept for future online encyclopaedias. Although analysed encyclopaedic websites satisfied a great deal of criteria for the successful adaptation to young users, there are still many elements that should be taken into account in the process of upgrading their usability, since all the possibilities of digital media has not yet been fully exploited. To tackle this issue there is a need for a more systemic approach to further investigations of the poorly developed field of encyclopaedistics for children and young adults.

References

Auchter, Dorothy. The evolution of the Encyclopaedia Britannica: from Macropaedia to Britannica Online. // *Reference Services Review*, 27 (1999) 3, pp. 291-299.

ClassConnect, http://connect.learnpad.com/content/activity.cfm?id=286650 (8.6.2017)

Flanagin Andrew J.; Metzger Miriam J. Digital Media and Youth: Unparalleled Opportunity and Unprecedented Responsibility. // *MacArthur Foundation Series on Digital Media and Learning*. Cambridge, MA: The MIT Press, 2008. pp. 5–28.

Gi-Zen Liu; Zih-Hui Liu; Gwo-JenHwang. Developing multi-dimensional evaluation criteria for English learning websites with university students and professors. // *Computers & Education,* 56 (2011) 1, pp. 65-79.

Jecić, Zdenko. Enciklopedički koncept u mrežnom okruženju. // *Studia lexicographica*, 7 (2013), 2 (13), pp. 99-115.

Jecić, Zdenko; Boras, Damir. Fotografija u virtualnim enciklopedijama – razrada kriterija evaluacije internetskih sadržaja. // *Proceedings – 10. međunarodno savjetovanje tiskarstva, dizajna i grafičkih komunikacija Blaž Baromić* / Bolanča, Zdenka; Mikota, Miroslav (ur.). Zagreb, Senj: Grafički fakultet Zagreb; Matica hrvatska Senj, 2006, pp. 87-92.

Loranger, Hoa; Nilsen Jakob. Teenage Usability: Designing Teen-Target Websites, Nielsen Norman Group, 4 February 2013, https://www.nngroup.com/articles/usability-of-websites-for-teenagers/ (10.5.2017)

Mateljan, Vladimir; Širanović, Željko; Šimović, Vladimir. Prijedlog modela za oblikovanje multimedijskih web nastavnih sadržaja prema pedagoškoj praksi u RH. // *Informatologia* 42 (2009) 1, pp. 38-44.

Ping Zhang; Gisela M. von Dran. Satisfiers and dissatisfiers: A two-factor model for website design and evaluation. // *Journal of the Association for Information Science and Technology*, 51 (2000) 14, pp. 1253-1268.

Rain, Ella. Childrens Online Encyclopedia, http://childrens-books.lovetoknow.com/Childrens_Online_Encyclopedia (26.6.2017)

Research-Based Web Design & Usability Guidelines, Official U.S. Government Edition. 2004. https://www.usability.gov/sites/default/files/documents/guidelines_book.pdf (1.6.2017)

Strategy of education, science and technology, Official Gazette 124/2014.

Sundin, Olof; Haider, Jutta. The networked life of professional encyclopaedias: Quantification, tradition, and trustworthiness. // *First Monday*, 8 (2013) 6.

Wilson, Ruth; Shorteed, Julie; Landoni, Monica. A Study into the Usability of E-encyclopaedias. // *Proceedings of the 2004 ACM symposium on Applied computing.* Nicosia, Cyprus, 2004, pp. 1688-1692.

Wing-Shui, Ng. Critical design factors of developing a high-quality educational website: Perspectives of preservice teachers. // *Issues in Informing Science and Information Technology*, 11 (2014), pp. 101-113.

Links

Encyclopedia Britannica, https://www.britannica.com/
Encyclopedia Britannica for kids – Britannica Kids, http://kids.britannica.com
KidzSearch Encyclopedia, http://wiki.kidzsearch.com/wiki/Main_Page
Q-files – The Great Illustrated Encyclopedia, https://www.q-files.com/home/
Wikipedia, http://www.wikipedia.org
World Book Online, http://www.worldbookonline.com

Thematic portal Znameniti.hr

Kristijan Crnković
ArhivPRO Ltd.
Proljetna 45, Koprivnica, Croatia
kristijan.crnkovic@arhivpro.hr

Vedrana Juričić
Croatian Academy of Sciences and Arts
Trg Nikole Šubića Zrinskog 11, Zagreb, Croatia
vea@hazu.hr

Irina Starčević Stančić
The Miroslav Krleža Institute of Lexicography
Frankopanska 26, Zagreb, Croatia
irinas@lzmk.hr

Summary

The digital infrastructure in Croatia is still waiting for a stronger investment cycle. Its development needs standardization and closer cooperation of all digital environment stakeholders. The work on the development of the thematic portal Znameniti.hr showed that these assumptions indeed are necessary preconditions. One-off financial support was provided by the Adris Foundation in 2016 within the project that was registered under the title Distinguished and Worthy Croats. The issue of standardization of data collection processing has finally started and an agreement has been reached that the name authority records from the National and University Library in Zagreb's digital repository should be used for building of the portal. Three unique set of metadata that will be applied for the bibliographic description of manuscripts, books, and collections of materials on the portal are also developed. Up until today, it has not been the case in Croatia that the materials are processed by standardized sets of metadata. Instead, institutions are applying their own descriptive elements. For all that, we believe that work on this project has spontaneously shifted the value of digital infrastructure in Croatia.

Key words: portal, distinguished and worthy Croats, Library of Croatian Academy of Sciences and Arts, collecting digital records, aggregation

About thematic portal Znameniti.hr

First of all, one-off financial support was provided by the Adris Foundation in 2016 within the project that was registered under the title Distinguished and

Worthy Croats[1]. The portal Znameniti.hr was launched in 2016 by the Croatian Academy of Sciences and Arts, its Library as a coordinator, together with the National and University Library in Zagreb, the Zagreb City Libraries, and the State Archive in Varaždin. In the second phase of building the portal, other institutions, which have online collections and/or repositories with valuable materials on distinguished and worthy persons from the Croatian history and present, joined the project. Those are the Miroslav Krleža Institute of Lexicography, the Institute of Ethnology and Folklore Research, and the Museum of Arts and Crafts in Zagreb.

The project's objective is to establish a basis for collecting and consolidated search of metadata of Croatian cultural, artistic and scientific institutions' digital materials by building the portal Znameniti.hr containing digital materials on the champions of Croatian culture, science, arts, and public life from different collections and/or repositories. The purpose of the project is to enable better accessibility and use of non-commercial scientific, cultural and artistic digital content to researchers and wider public.

One of the main project tasks was to define the content framework and scope of the portal. The first step was to define the distinguished and worthy Croats. They are persons who worked in Croatia or outside Croatia, regardless of their origin or birth, or persons who contributed to the recognition, definition and affirmation of the Croatian identity by their actions and whose works became an important part of the Croatian heritage. According to this broad definition it was decided that the portal's title will be Znameniti.hr (eng. Distinguished.hr) (Figure 1). What was achieved by this? First of all, the name was shortened and made more memorable. The diacritical signs were avoided since they can cause different renderings on the computer screen, i.e. on the Internet. And finally, the national affiliation was removed from the portal's title itself allowing for future internationalization of the content.

Namely, thematic portal Znameniti.hr should be the beginning of the "little Croatian Europeana". Europeana, the largest digital platform in Europe, has been developing for ten years on the principles and standards that are related and applied to the portal Znameniti.hr. At the end of 2015 the Europeana portal, with its subtitle Think Culture, has been redesigned and renamed to Europeana Collections. The internet address has also been changed from http://www.europeana.eu to https://www.europeana.eu/portal/hr. At that time the first two collections appeared at the homepage – Art and Music. Today there are also collections Fashion, Photography, 1914-1918, Maps and Geography, Natural

[1] Inspiration for the project title was found in the renowned book Znameniti i zaslužni Hrvati te pomena vrijedna lica u hrvatskoj povijesti od 925-1925. (eng. The distinguished and worthy Croats and the persons worth mentioning from the Croatian history from 925 to 1925) published in Zagreb in 1925. (reprinted in Zagreb in 1990).

History and Sport. One day, we hope, there will be a Distinguished Europeans portal.

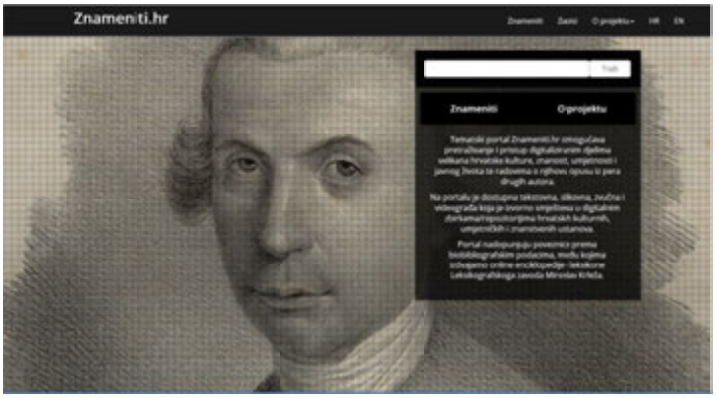


Figure 1. Homepage of thematic portal Znameniti.hr

Selection criteria

In this section, the main criteria for determining the process of publishing records of the distinguished persons will be discussed. It is important to determine the criteria and the possibilities for regular publication, e.g. one of the criteria may be a person's anniversary. During this process the rules for collection and presentation of digital data are determined.

Criteria for selecting the distinguished persons were: a) all persons who have a milestone anniversary of birth or death in the current year; b) all persons who already have a notable collection of digitised materials; c) special attention will be paid to the distinguished women; d) all persons with 10 or more digitised objects will be considered. Two data bases, the Croatian Encyclopaedia by The Miroslav Krleža Institute of Lexicography and the Zagreb City Libraries' Calendar of annual events were chosen as basis for collecting information on the distinguished persons. After collection, in the next step, distinguished persons who celebrate an anniversary between 2017 and 2020 were selected by applying the provenance criteria. Based on the current records a spreadsheet was created containing list of all distinguished persons and number of digital objects (texts, images, audio and video materials) available in the digital collections and/or repositories (Figure 2). It showed that some persons does not have a digital presence at all.

For example, it came as a surprise that only a few digital documents were created about Faust Vrančić whose 400th death anniversary is celebrating in 2017. In the same anniversary category for 2017 are Dobriša Cesarić, Gustav Krklec, Vesna Parun, Ivan Lukačić, Franjo Tuđman, Maksimilijan Vrhovac, Cardinal

Franjo Kuharić, etc. On the other hand, there are many digitised materials about Vinko Žganec[2], a well-known Croatian ethnomusicologist.

Figure 2. Spreadsheet containing list of all distinguished persons

Metadata collection

During the first project meeting in January 2017 it was decided that the process of collecting metadata will be performed as follows: in the repository of a participating institution, the digitized objects having a particular person as the author or the subject, will be selected, with all metadata included. Those records, when filtered, will be entered in the Znameniti.hr portal. Filtered metadata are merged at the OAI-PMH level, and are displayed at the portal interface.

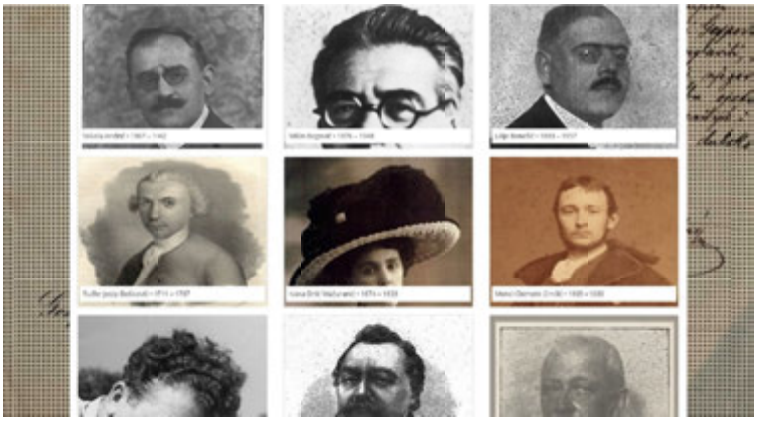


Figure 3. Photographs of the distinguished persons

[2] Around 60 CDs and around 80 collections were digitized. The collection of 40 photographs with dr. Žganec and 342 photograph mentioning Žganec in the title still remain to be digitized.

An agreement was made to enrich the name authority records with ISNI and VIAF numbers – international standard identifiers. This is the criterion for aligning the identifiers of the name authority records. Local identifiers from each repository have been replaced by a global standard identifier that is unique in all repositories from which the data are selected. This solution has enabled the aggregation of heterogeneous data from different repositories and their integration into the portal Znameniti.hr. Consequently, the portal became the first aggregator portal to combine records created according to different metadata profiles from different institutions whose repository records are downloaded by the OAI-PMH protocol.

We have used the knowledge and technology that ArhivPRO Ltd. implemented in the aggregation system launched by the Ministry of Culture in 2013. This portal showed how important it is to develop an aggregate system that collects data from different digital sources.

Search and retrieval

The Znameniti.hr portal is created for wide-range of users (students, scientists, researchers) so it should be simple, incorporating well-designed search engine and instructions on how to navigate it. The photographs of distinguished persons, which follow immediately after the home page and offer the ability to browse through the content of the portal (Figure 3), are put forward. The portal provides search by different sets of metadata. The search results are displayed in two columns. In the left column, they are arranged in two groups of facets: *Izvor zapisa* (source record, i.e. institutions in whose repositories the requested data are stored), and *Znameniti* (list of distinguished persons whose works are included in the portal). The digitized objects are displayed in the right column. The structure of the portal Znameniti.hr and instructions for its use are shown in Figure 4.

URI (Uniform Resource Identifier), defined as a unique identifier of various sources on the Internet, has been identified as a prerequisite for acquiring the Croatian digital resources (URI HR). By using such identifier each resource can be uniquely identified and accessed.

The development of the portal Znameniti.hr also showed that the existing aggregating system, developed four years ago for aggregation of data for the Europeana, needs to be upgraded. Besides the name authority records data base, there is a lack of services for integration and unified search of other authoritative records. The idea of developing the Croatian Europeana, initially formulated in parallel with the development of the aggregating system but for different reasons still unrealised, now seems within reach by the development of the cooperation portal.

In the beginning, all the data and the portal application were hosted at the server of the company responsible for the application development (http://znameniti.eindigo.net). Later, the portal has been successfully moved to the virtual private

server (VPS) allocated by the University Computing Centre (SRCE) to the Institute of Ethnology and Folklore Research. The allocated VPS space is shared with the DARIAH-HR consortium of institutions coordinated by the Institute.
The Ministry of Culture of the Republic of Croatia has approved the use of the domain Znameniti.hr without fee. Thanks to that, http://znameniti.hr will be the permanent URL of the portal.

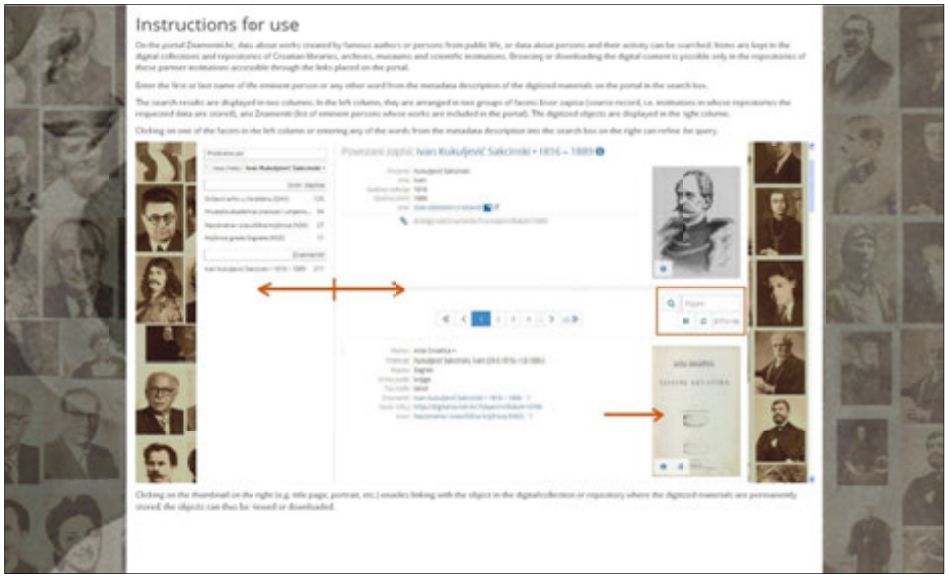


Figure 4. Instructions for use

Conclusion

The project Znameniti.hr, defined as a pilot project, aspires to become a unique access point for all digitized and/or digital objects on the distinguished persons originally placed in the digital collections and/or repositories of the MLA, scientific, and other institutions. The main precondition for developing this kind of portal is the existence of digital resources with accompanying metadata. This can facilitate the retrieval and provide a unified approach to all digital resources.
Currently in Croatia there is relatively modest number of repositories and they are not interconnected. There are also digital materials that are not preserved in an adequate manner and are not published on Internet, therefore not publically available. Croatia does not have its own, national, digital library. Because of that, it may be believed that the portal Znameniti.hr will draw attention to the need for development of the digital infrastructure in culture and science and initiate systematic metadata collection of the related records. Collecting the related records will allow their organization on a unique portal (e.g. Znameniti.hr) where they can be mutually connected and searchable.

The main goals of the project Distinguished and Worthy Croats, as it was originally called, were significantly exceeded. The planned number of collected records, initially set to cca. 1,500, was by the end of the first phase of project exceeded by 4½ times (the total of 35 distinguished persons with 6,511 records were collected). It is expected that the number of participating cultural and scientific institutions will grow as well as the number of included digital records on the distinguished persons. In the beginning of the project, four institutions were involved, but now four more institutions are showing their interest in collaboration and systematic collecting, publishing and interlinking of digital resources. That is why it was agreed that the cooperation of these eight institutions will continue, that the selected representatives from those institutions will form a coordination committee, and to define the cooperation through a special agreement.

Other institutions, having similar digital records, are also showing interest to contribute to the Znameniti.hr portal. In the initial phase, the cooperation may be realized by connecting the thematically organized metadata, e.g. music, film, art, history etc. This might be the foundation for development of the national cultural digital infrastructure, e.g. a national digital library. This portal is a great example in the still developing digital humanities area showing how digital materials, when processed by the cultural institutions, can become an important resource for education and science usable in the curriculum.

Resources

ArhivPRO Ltd., http://www.arhivpro.hr
Croatian Academy of Sciences and Arts, http://www.hazu.hr
Europeana portal, https://www.europeana.eu/portal/hr
Indigo – hybrid repository platform based on semantic technologies, http://www.eindigo.net/
National and University Library in Zagreb, http://www.nsk.hr
The Institute of Ethnology and Folklore Research, http://www.ief.hr
The Miroslav Krleža Institute of Lexicography, http://www.lzmk.hr
The Museum of Arts and Crafts in Zagreb, https://www.muo.hr/
The State Archive in Varaždin, http://www.dav.hr
The Zagreb City Libraries, http://www.kgz.hr

Original scientific paper

Proposing an Instrument for Evaluation of Online Dictionaries of Sign Languages

Klara Majetić
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
kmajetic@ffzg.hr

Petra Bago
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
pbago@ffzg.hr

Summary

Dictionary criticism has been criticized for only being concerned with the description of design features of dictionaries, while not giving much attention to the evaluation of dictionary features, and for lack of objective standards. In this paper we describe the features of Croatian sign language (HZJ), investigate characteristics of dictionaries of signed languages (with special attention to dictionaries of Croatian sign language), and examine evaluation criteria for different types of printed and online dictionaries. We propose a set of evaluation criteria relevant for a sign language dictionary and point out why sign language dictionaries should transcend the traditional printed formats. The evaluation criteria described in this paper will in further research be used as an evaluation instrument of some existing online sign language dictionaries to assess the instrument. We will apply that as a foundation for the construction of a model of an online HZJ dictionary. It is our hope the evaluation instrument provided in this paper, and the model that will be suggested afterwards might help make a professional, well thought out, and beneficial online dictionary of HZJ.

Key words: Croatian sign language (HZJ), sign language lexicography, online dictionary, dictionary evaluation

Introduction

As expected of a paper in the field of lexicography, we begin with a definition of dictionary criticism by *Dictionary of Lexicography:* "a branch of dictionary research concerned with the description and evaluation of dictionaries and other reference works" (Hartmann and James 2002: 32). Dictionary criticism has been criticized for only being concerned with the description of design features of dictionaries, while not giving much attention to the evaluation of dictionary

features. Moreover, most of those reviews were unfavourably judged by other lexicographers, mainly for the lack of objective standards (Swanepoel 2008: 209). In this paper we describe the features of Croatian sign language, investigate characteristics of dictionaries of signed languages, and examine evaluation criteria for different types of printed and online dictionaries. We propose a set of evaluation criteria as an instrument for evaluation of online dictionaries of sign languages, thereby hoping to contribute to an improvement in the quality of dictionary criticism in the field of sign language lexicography.

About sign languages

In this section we bring a brief overview of features of a sign language[1]. A sign language is a primary or first language of the Deaf[2] and Deafblind[3]. The media of a sign language are signs made with hands, head, and body. A sign language is "not a crude approximation to a spoken language" (Trask 1999: 184) – it is genuine and natural, it has an extensive vocabulary, complex grammar, and is as flexible and as expressive as a spoken language. There are many sign languages – for example Croatian Sign Language (HZJ), American Sign Language (ASL), Australian Sign Language (Auslan), British Sign Language (BSL), Brazilian Sign Language (BLS or Libras) and so forth. Even though those sign languages share names with some spoken languages, they are not equally spread – sign languages don't depend on national borders since they develop in any group of deaf people. For example, BSL and ASL are mutually unintelligible, even though British and American hearing people share the same spoken language. Grammars of sign languages don't resemble those of languages spoken in the same geographical area – "ASL shares more with spoken Japanese than with English" (Nakamura 2008). Unlike the oral-auditory modality of spoken languages, sign languages have visual-spatial modality. From this difference in modality stem many issues in writing any sign language down and creating a dictionary of a signed language, as will be shown in the next section.

[1] Due to the space constraints, some parts of the overview may seem as an oversimplification of some sign languages. While organizing this section, we had the characteristics of the Croatian Sign Language in mind. However, while compiling the evaluation instrument, we took into account characteristics of other sign languages.

[2] The term “Deaf” refers to the members of the linguistic community of sign language users, while the term “deaf” describes the audiological state of deafness. (Morgan, Woll 2002: 20)

[3] Deafblindness is a specific and unique double sensory damage of sight and hearing in various possible combinations of intensity: hearing and visual impairment, deafness and visual impairment, blindness and hearing impairment, and practical deafblindness. A person with progressive visual impairment or with constant visual impairment with deafness forecast can also be considered as Deafblind. Also, this classification covers the Deafblind with characteristic syndromes (Usher syndrome, Charge syndrome). (translated by the author, Tarczay 2001: 146-147)

Sign languages were first systematically described in the middle of the 20th century. The basic unit in any sign language is a sign. The five elements of a sign are handshape, location, movement, orientation, and non-manual markers. Croatian sign language (HZJ) distinguishes 44 different handshapes and 17 main locations on the body. Movement is the most complex element of a sign – it describes how the hand moves through the articulation space: its direction (up, down, left, right, forward, backward), repetition, speed, and coordination (when using both hands to sign). Orientation element is the distinctive degree of rotation of the hand in relation to the signer. Non-manual markers are facial expressions, and movements of the body and head. These elements of a sign, or sign's parameters, are phonemes of a signed language. They are different from phonemes of a spoken language in that they often appear simultaneously.

Sign languages cannot be written down as spoken languages can. "ASL (…) does not have a written form. A sign conveys a concept, not an English word, and the production of a sign involves five elements that need to be described" (Tennant, Gluszak Brown 1998: 26). There are notation systems[4] – Stokoe's notation system being the first – which can be used to write sign languages down so that the analysis of the language structure could be possible. Those notation systems use symbols and abstract pictures to describe a sign and all its elements, and are not written in a single line but also use the vertical plane to add information. Course books by the Croatian Alliance of the Deafblind *Dodir* "Znak po znak" (ZPZ)[5] offers a notation system that is a combination of a translation of the signs with added symbols for extra information about movement, repetition, etc.

It is interesting that HZJ, unlike its spoken counterpart, does not use grammatical cases nor verb conjugations in the sense the spoken Croatian language does. There are some modifications to the signs depending on the quantity of subjects and/or objects, but signers often use the canonical form of nouns and adjectives in combination with pronouns, adverbs, and prepositions (which can often be modified, e.g. in numeral incorporation), while verbs of motion and location use other kinds of grammatical markers (e.g. classifier predicates).

[4] For example:
Stokoe notation (Stokoe, W. "C.(1960). Sign language structure." Studies in Linguistics: Occasional Paper 8.)
The Hamburg Notation System (HamNoSys) (Prillwitz, Siegmund, and Hamburg Zentrum für Deutsche Gebärdensprache und Kommunikation Gehörloser. HamNoSys: Version 2.0; Hamburg Notation System for Sign Languages; An Introductory Guide. Signum-Verlag, 1989.)
SignWriting (Sutton, Valerie. Lessons in Sign Writing: Textbook. SignWriting, 1995.)

[5] Tarczay, Sanja et. al. Znak po znak 1, 2, 3 : udžbenik za učenje hrvatskog znakovnog jezika. Hrvatska udruga gluhoslijepih osoba “Dodir”. 2006-2007.

Dictionaries of signed languages

Since the first systematic descriptions of sign languages, the need for a good sign language dictionary was obvious. The first printed sign dictionaries in the 20th century were monodirectional alphabetical lists of words of a spoken language – that could also be divided thematically – which was then translated into a sign language via a picture and/or a description of the sign's elements. Another, but rarer kind of dictionaries that developed later in the 20th century are dictionaries that listed signs by one of the sign elements (e.g. handshape) and gave a translation to a spoken language, e.g. *A dictionary of American Sign Language on linguistic principles* by Stokoe, William C., Dorothy C. Casterline, and Carl G. Croneberg from 1976.

Those first dictionaries usually chose their entries on a basis of a dictionary of a spoken language. Creating a corpus for a dictionary of a signed language mostly takes several deaf people who are given topics to talk about while being recorded. Today several larger corpora exist, e.g. "[…] the sign language of deaf communities in the Netherlands, the UK, Ireland, Sweden, Greece, Australia and the US but also e.g. Mali" (Crasborn 2010). In such way lexicographers can find out which signs are the more frequent ones, how the signs are used and what meanings they carry.

Since sign languages have a different modality from spoken ones, it is difficult to print a dictionary that will provide enough visual-spatial information that could easily be understood. Some dictionaries use one of the existing notation systems to try to convey more information. "The problem is that, in order to profit from the encoded information, the dictionary user has to invest time and effort either in learning the (not very transparent) codes or in constantly consulting the key to the codes. It is not by any means evident that most dictionary users are prepared to make this kind of investment" (Singleton 2000: 205).

Therefore it is much easier to use dictionaries that rely on pictures with descriptions of signs. Modern technology brings new media into play and enables the creation of e-dictionaries of sign languages that can use a variety of media – text, picture, notation systems, and, most importantly, video. When using both online dictionaries of sign language and those on a CD-ROM, it is impossible not to agree with Singleton (2000: 200) when he writes that "many of these are quite disappointing in terms of their failure to use the extraordinary possibilities offered by the technology, some being little more than rather crude glossaries with very limited search facilities." Of course, there are also some great examples of dictionaries that have taken advantage of technology and put the traditional formats aside[6].

[6] For example: an online dictionary of Dutch - Flemish Sign Language "Woordenboek Nederlands — Vlaamse Gebarentaal, Vlaamse Gebarentaal — Nederlands" URL: http://gebaren.ugent.be/ (10.5.2017.) and an "Online Dictionary of New Zealand Sign Language" URL: http://nzsl.vuw.ac.nz/ (8.5.2017.)

HZJ dictionaries

At the time being, Croatian Sign Language or HZJ doesn't have a high quality dictionary. ZPZ as a learner's handbook (2006 and 2007) offers "znakovnica"[7] – a thematical alphabetical list of signs with their images and translations into Croatian at the end of every theme. There is a traditional printed dictionary "Hrvatski znakovni jezik"[8] by a group of editors which was published in 2015[9]. The dictionary is a short alphabetical list of pictures of HZJ signs and their translations into Croatian. There are not many signs and the ones that were included into the dictionary are not always the most frequent ones – one can find *žuboriti* (to murmur, to babble), but not *žvakati* (to chew). Another tends to be a specialised dictionary: "Gluhi i znakovno medicinsko nazivlje: kako komunicirati s gluhim pacijentom"[10] published in 2010[11]. It consists of many texts on the Deaf and HZJ, a small section of general signs (e.g. *ići* (to go)), and a slightly bigger section on medical terminology, e.g. *alergija* (allergy), *Alzeimerova bolest* (Alzheimer's disease). While the entries in the general dictionary section only consist of a gloss and a picture, the medical ones include a short definition. This dictionary consists of around 250 entries. The only online HZJ dictionary called CroDeafWeb[12] is on a good track – it has an alphabetical list of words, each entry has a video or a gif showing how a sign is signed, and a short written description of the movements in signing. Unfortunately, it only has around 500 signs, a lot of which are liturgical. Comparison of HZJ dictionaries by the approximate number of entries is shown in Table 1.

Table 1: Approximate number of entries in dictionaries of HZJ

Dictionary	Number of entries
Znak po znak	4500
Hrvatski znakovni jezik	1200
CroDeafWeb	500
Gluhi i znakovno medicinsko nazivlje	250

[7] A rough translation is “a sign book”, a coined word based on “slikovnica” (“a picture book”).

[8] Translation: “Croatian sign language”.

[9] Hrvatski znakovni jezik . Ristić, Milan.; Baštijan, Zdravka.; Biškupić Andolšek, Tajana. (Eds.) Zagreb : Hrvatski savez gluhih i nagluhih, 2015.

[10] Translation: "The deaf and signs for medical terminology: how to communicate with a deaf patient"

[11] Gluhi i znakovno medicinsko nazivlje : kako komunicirati s gluhim pacijentom. Šegota, Ivan; Šendula-Jengić, Vesna; Herega, Damir; Petaros, Anja; Conar, Jevgenij. Zagreb. Medicinska naklada. 2010.

[12] http://www.crodeafweb.org/rjecnik/index.html

Dictionary evaluation

Evaluations of traditional printed dictionaries were often written but have, as Hartmann notes, "(…) been beset by personal prejudice rather than noted for the application of objective criteria" (1996: 241). To make it easier to apply objective criteria, a kind of a measuring system had to be made. Researchers made lists of points one should mention in a review, or evaluation criteria.

Haas lists 12 desiderata (as cited in Landau 2011: 11) any bilingual dictionary should contain, while Landau (2011: 11) notes some limits which those needed elements set for each other:

1. "It provides a translation for each word in the source language.
2. Its coverage of the source language lexicon is complete.
3. Grammatical, syntactic, and semantic information is provided.
4. Usage guidance is given.
5. Names are included.
6. It includes special vocabulary items, such as scientific terms.
7. Spelling aids and alternative spellings are indicated.
8. Pronunciation is included.
9. It is compact in size – which obviously limits its coverage of items 1-8."

A traditional printed dictionary can't be compact in size and provide all the above mentioned information. The first of quoted Haas' desiderata – providing translation for each word in the source language – might be possible only for a dead language because only a dead language has a finite number of texts and "no new sentences are produced in a dead language" (Zgusta 1971: 217). This can be done with an e-dictionary – "Electronic dictionaries are not subject to such constraints, and, with their capacity to offer links to other entries and to other sources of information, may indeed be virtually limitless in respect of the quantity of information they can make available" (Singleton 2000: 199-200), thereby having the capacity to continuously improve the coverage of the dictionary. Jackson (1996: 7-11) proposes a range of vocabulary, word formation, homographs, defining, sense division, lexical relations, collocations, connotations, pronunciation, grammar, usage, examples, etymology, and special features as the main criteria for evaluating a dictionary. More recent criteria for dictionary criticism could be divided into categories like the ones Svensén (2009: 483) lists: dictionary functions, dictionary users, advice given to the users, price, layout / web design, the compiler(s), comparison with other dictionaries, prehistory of the dictionary, reference to other reviews, the reviewer, dictionary basis, outside matter, lemma selection, establishment of lemmas, search and access options, entry structure, the normative/descriptive dimension, equivalents, grammar, orthography, pronunciation, semantic and encyclopaedic information, diasystematic information, etymology, examples, collocations, idioms, illustrations, synonymy/antonymy, cross-references, entertainment value, and unified concluding evaluation. This list already mentions some elements which can only be applied to e-dictionaries, such as

web design, and search and access options. For a sign language there are some additional elements that should be evaluated. In the recent years, the field of sign language lexicography tackles issues such as lemmatization, lemma information, and ordering and searching in e-dictionaries (Zwitserlood, Kristoffersen and Troelsgård 2013: 259-283). Among information on sign languages and lexicography, Zwitserlood (2010) writes about situation in the Netherlands and reviews their online dictionary of Dutch Sign Language (NGT). Capovilla et al. (2003) present how Libras went from a printed dictionary to a digital encyclopedia, and even made it possible for deaf quadriplegic users to compose Libras-based sign messages that can be converted to ASL, printed, spoken with digitized speech both in Portuguese and English. Hanke (2004) and Hanke and Storz (2008) describe the more technical side of creating a corpus of technical terms in German Sing Language (DGS) and of writing signs down using the Hamburg Sign Language Notation System (HamNoSys), an alphabetic system describing signs on a mostly phonetic level, first published in 1987. Kristoffersen and Troelsgård (2012) point out hyperlinks and multimedia, search facilities, flexibility, and the ability to meet diverse user needs as particularly useful characteristics of a sign language dictionary.

In our opinion, the contribution of the paper to the sign language lexicography is twofold. First, as already mentioned, the evaluation instrument contributes to the improvement in the quality of dictionary criticism by introducing a framework for description and evaluation of online sign language dictionaries. Second, due to the lack of comprehensive and extensive online sign language dictionaries for many languages, the instrument can be implemented in the initial stages as a tool for the development of a model for these type of dictionaries. By researching relevant literature, the chosen criteria for the instrument are deemed the most relevant.

An instrument for evaluation of online dictionaries of sign languages

1. Intended users of the dictionary:

a) experts
b) native signers (the deaf users)
c) learners of a sign language as a foreign language (for the hearing users)

Rationale: As with all dictionaries, it must be known who the majority of users will be. Will it be the deaf population to whom a sign language is their first language, or an expert in special education or in linguistics, or someone who started a course in a sign language and would like to become fluent. This categorization is based on Varantola's (2002) rough division of dictionary users into professional users, non-professional users, and language learners. To the best of our knowledge, most of the sign language dictionaries have been aimed at all the intended user groups mentioned above, due to the lack of resources needed to create dictionaries for specific users.

2. Type of the dictionary according to the subject field covered:
a) general
b) specialized

Rationale: Since there generally aren't too may sign language dictionaries to begin with, most of them are considered general type of dictionaries. However, there are examples like the Institute of German Sign Language and Communication of the Deaf (IDGS) at the University of Hamburg that has been working on the development of dictionaries for specialized areas of sign language use in fields such as computer technology, psychology, joinery, domestic sciences, social work, health and nursing care, and landscape and horticulture (König, Konrad, and Langer 2004).

3. Type of the dictionary according to the norm:
a) normative
b) descriptive

Rationale: It is important to identify the function of the dictionary, whether the dictionary is meant to prescribe the correct form(s) of a sign and proscribe others, or whether its objective is to document and describe the usage of a language.

4. Number of languages in the dictionary:
a) monolingual
b) bilingual
c) multilingual

Rationale: Today, the majority of the official spoken languages have their monolingual dictionaries where a word from the language is described using words from that same language. With a sign language, a monolingual dictionary would only be possible and usable in a video format or with the use of notation systems. To the best of our knowledge, there are no monolingual sign language dictionaries (where sign language would be used as the metalanguage). They are usually bilingual – translating from a spoken language into the corresponding national sign language (e.g. from Croatian into HZJ). Multilingual dictionaries can have translations of a sign into several spoken languages or translations of e.g. an English word into signs in different sign languages.

5. Scope:
a) monoscopal
b) biscopal

Rationale: According to Hausmann and Werner (1991:2740), a monoscopal bilingual dictionary contains dictionary entries in one language and their translations into another (e.g. from HZJ into Croatian). A biscopal bilingual dictionary contains translations to and from both languages (e.g. from Croatian into HZJ and from HZJ into Croatian), with majority of dictionaries having the two parts separated.

6. Function
a) for text/sign production
b) for text/sign reception

Rationale: Another characteristic of bilingual dictionaries is its function (Hausmann and Werner 1991:2741) that indicates whether the purpose of the dictionary is to aid in text production or text reception. The majority of dictionaries try to cover both functions. For bilingual sign dictionaries it is important to identify whether the dictionary can be used for sign production or sign reception.

7. Direction:
a) monodirectional
b) bidirectional

Rationale: According to Hausmann and Werner (1991:2742), a monodirectional bilingual dictionary indicates whether the mother tongue of the user is the source or the target language. A bidirectional bilingual dictionary indicates that it is intended for mother tongue speakers of both languages covered. Bidirectional dictionaries usually either don't meet the objectives or have an extensive and complex structure, making the monodirectional dictionaries more user-friendly.

8. User interface design:
a) overview of the page
b) simplicity of the design
c) intuitivity, the ease of use
d) adaptivity for users with visual impairment (font size, colour contrast)
e) simple navigation between connected information
f) user's guide (instructions for use, key to symbols, key to notation system, and abbreviations and content of entries)

Rationale: The design and layout of an electronic dictionary must always keep the needs of their user in mind. Since both the Deaf and the Deafblind use sign languages, the design should be clean with strong contrast and easily accessible possibility to magnify the images and text for visually impaired people. The dictionary options should be easy to use and intuitive to users. Some directions for usage and legends of used symbols should appear on the home page. Since we expect an online dictionary to connect multiple pieces of information and link relevant data together, navigating those links should be fast, clean, and simple.

9. Searching:
a) direction of the search function:
    i) monodirectional (from a spoken to a signed language)
    ii) monodirectional (from a signed to a spoken language)
    iii) bidirectional
b) spellchecker
c) searching through headwords or through whole entries

Rationale: Online dictionaries of spoken languages are almost always bidirectionally searchable. That is rarely the case with sign languages since they are only written down with special notation systems. Sign language dictionaries can make searching by sign possible by choosing elements related to sign form or sign usage, for example handshape, place of articulation, orientation, movement, handedness, mouth movement, region- or age-specific use (Zwitserlood et al. 2013). With that option they become bidirectional. Searching by defined topics could further speed up the process of finding a term particular to a subject field. Spellchecker within the text search option is always an advantage, especially when the user of the dictionary is using his second/foreign language in the search. Searching only by headwords and not searching through whole entries can give fewer results.

10. Search results:

a) type of information included in the result list
b) ability to narrow the result list
c) search relevance marker

Rationale: As with every search result, it is important what type of information is included in the result list. The presentation of search results of online sign language dictionaries can include the following information: photograph, drawing, video, gif, formal notation, ID number, word class, gloss, equivalent(s), mouth action, text description, and topic (Kristoffersen and Troelsgård 2012, Zwitserlood et al. 2013). Another useful feature of the search function is the ability to narrow down a result list (e.g. by elements of a sign or word class). Additional useful information to help the user is the search relevance marker that indicates what results are more relevant to the search criteria (e.g. by displaying three stars for the most relevant search results or one star for the least relevant ones).

11. Entries:

a) number of entries
b) can users add entries
c) amount of free content
d) can the content be downloaded (image or video of signing a sign)
e) criteria for ordering the entries:
    i) alphabetically by the translation into the spoken language
    ii) by an element of the sign
    iii) thematically
f) how the entries were chosen:
    i) from a corpus
        (1) are the entries signs of high-frequency
    ii) from another source
    iii) are proper names included

g) content of the entry:
   i) translation of a sign to a spoken language
   ii) translation of a sign to several spoken languages
   iii) detailed video of signing
   iv) detailed image (photograph, drawing, or gif) of signing
   v) description of signing (elements of the sign) written in a spoken language or as an audio file
   vi) description of a sign written in a notation system
   vii) information about mouthing
   viii) context, the sign in use
   ix) grammatical information (e.g. modifications, numeral incorporation, classifier predicates, etc.)
   x) sense division (e.g. polysemy, homonymy)
   xi) geographical information about a sign
   xii) examples
   xiii) collocations
   xiv) idioms
   xv) lexical relations (i.e. synonyms, antonyms, hypernyms, etc.)
   xvi) etymology of a sign
   xvii) semantic and encyclopedic information
      (1) images (photograph, drawing, or gif) of sign meaning
   xviii) ID number
   xix) topic

Rationale: Does the dictionary have a big or a small entries list? All sign language dictionaries at present time have a much smaller number of entries than dictionaries of spoken languages – many meanings are crowded in the same entry because they can be seen as a minor modification of a sign or not mentioned at all. How is the sense division handled, i.e. are the meanings a sign can have found in the same entry? Are selected entries used frequently by the Deaf? Since sign languages are often not standardized and not seen nor used in the media[13], the (regional) Deaf community could contribute to such dictionaries by broadening the number of entries and linking synonyms, but there needs to be a (manual or automatic) validation process in place. Is all content free? For users of a sign language it is useful to be able to download and/or print out some content, especially images and videos. If there is a list of entries, they are ordered alphabetically by the translation into a spoken language, by a sign's elements, or by topics, although ordering is less of an issue in e-dictionaries than in traditional printed ones. Unless corpus studies have been carried out, sign frequency information is not usually available. The

[13] HZJ can only publicly and officialy be seen in public television on channel 4 in "Vijesti uz hrvatski znakovni jezik" (News with Croatian sign language) and "Dnevnik 2" (the main news program), where an interpreter translates what the speaker says.

frequency of the spoken language equivalent is not the same as the frequency of the sign. Besides a sign's translation into at least one spoken language, an entry in an online sign language dictionary should have a video of the most frequent ways of producing each sign, which includes the most frequent modifications and variations of the sign that depend on the context. A description of the sign's movements written in a spoken language or in an audio recording can be of much help to the partially blind. Signs don't have pronunciation but they do have a *mouthing* segment – the signer shapes his lips as if silently pronouncing some vowels, syllables, or words[14]. Mouthing is not obligatory for all signs and can depend on context. One sign produced in isolation can have extensive mouthing, while produced in context may not show any mouthing at all (e.g. Schermer 2001). Content such as examples, collocations, idioms, and lexical information (i.e. synonyms and antonyms) enrich dictionaries, and are especially valuable for language learners. An additional piece of information that can be useful since sign languages develop naturally in any group of deaf people is the information about the sign's geographic spread.

12. Evaluating extra content:
a) word/sign of the day
b) thesaurus
c) grammatical description of the language
d) orthography description
e) word/sign games
f) related entertainment
g) links to other dictionaries or sources

Rationale: Dictionaries can be enriched by some extra content such as word/sign of the day, thesaurus, word/sign games, some related entertainment and links to other dictionaries which is especially helpful for language learners, but also engages frequent users in a fun way (i.e. entertainment value).

13. Evaluating quality of the content:
a) who are the compilers
b) are there enough entries considering who the intended user is
c) accuracy of entry information (e.g. translations, sign variants, description of production of the sign, grammatical information, usage information)
d) if the dictionary is corpus based: is there information about the corpus and how it was used
e) are there hyperlinks to the corpus

Rationale: Due to the authoritative role that the dictionaries often play, the content of entries must be of high quality – a dictionary is a guide for its users and therefore its translations and additional information must be accurate, and

[14] Schermer, Trude. The role of mouthings in Sign Language of the Netherlands: Some implications for the production of sign language dictionaries. The Hands are the head of the mouth: The mouth as articulator in sign languages (2001): 273-284.

the members of the editorial team (i.e. the compilers) must be known. Furthermore, it is essential that the dictionary has enough vocabulary coverage for the intended user. If the dictionary is corpus-based, information about the corpus must be given (e.g. when the data for the corpus was collected, what topics the data covers, how many signs are in the corpus). In addition to information about corpus data, it must be clearly indicated how was the corpus used (e.g. for lemma selection, for usage information, for example sentences, for collocations selection, for division into senses). Information about the corpus and connections of entries to the corpus make the dictionary more reliable and trustworthy.

Conclusions and future work

The goal of this paper was to examine evaluation criteria for printed and online dictionaries, and to extract the ones relevant for a dictionary of a sign language, thereby proposing an instrument for evaluation of such dictionaries. By introducing a framework for description and evaluation of online sign language dictionaries, the instrument contributes to the improvement in the quality of dictionary criticism for this type of dictionaries. Another goal was to point out why sign language dictionaries should transcend the traditional printed formats. The evaluation criteria described in this paper will be further used as an evaluation instrument in evaluations of some existing online sign language dictionaries to assess the instrument. The instrument can be implemented in the initial stages as a tool for the development of a model of online sign language dictionaries, and therefore will be the foundation for the construction of an online HZJ dictionary. Considering the lack of HZJ dictionaries which could be useful to a broader audience – and not just to hearing people who have begun learning HZJ – it is our hope the evaluation instrument provided in this paper, and the model that will be developed based on this instrument, might help make a professional, well thought out, and beneficial dictionary of HZJ.

References

Capovilla, Fernando, Duduchi, Marcelo, Raphael, Walkiria Duarte, Luz, Renato, Rozados, Daniela, Capovilla, Aleassandra, Macedo, Elizeu. Brazilian sign language lexicography and technology: Dictionary, digital encyclopedia, chereme-based sign retrieval, and quadriplegic deaf communication systems. Sign Language Studies, 3(4), 393-430. 2003.

Crasborn, Onno. The Sign Linguistics Corpora Network: towards standards for signed language resources. 2010.

CroDeafWeb: rječnik hrvatskog znakovnog jezika. http://www.crodeafweb.org/rjecnik/index.html (3.4.2017.)

Hanke, Thomas. HamNoSys-representing sign language data in language resources and language processing contexts. In LREC (Vol. 4). 2004.

Hanke, Thomas, Storz, Jakob. iLex–A database tool for integrating sign language corpus linguistics and sign language lexicography. In LREC 2008 Workshop Proceedings. W 25: 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. Paris: ELRA (pp. 64-67). 2008.

Hartmann, Reinhard Rudolf Karl, ed. Solving language problems: from general to applied linguistics. Vol. 20. University of Exeter Press, 1996.

Hartmann, Reinhard Rudolf Karl, and Gregory James. Dictionary of lexicography. Routledge, 2002.

Hausmann, Franz Josef; Werner, Reinhold Otto. Spezifische Bauteile und Strukturen zweisprachiger Wörterbücher: eine Übersicht, Art. 286. In: Hausmann, Franz Josef; Reichmann, O.; Wiegand, H.E.; Zgusta, L. (Eds.) Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. Dritter Teilband. Berlin : /, 1991, p. 2729-2769.

Hrvatski znakovni jezik . Ristić, Milan.; Baštijan, Zdravka.; Biškupić Andolšek, Tajana. (Eds.). Zagreb : Hrvatski savez gluhih i nagluhih, 2015.

Jackson, H. 1996. Dictionary Criticism. Unpublished manuscript. Birmingham City University, Faculty of Computing and Information Studies, Research Papers.

König, Susanne, Reiner Konrad, and Gabriele Langer. What's in a sign? Theoretical lessons from practical sign language lexicography. Signs of the time. Selected papers from TISLR. 379-404. 2004.

Kristoffersen, Jette Hedegaard, Troelsgård, Thomas. The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. Oxford University Press. 2012.

Landau, Sidney I. Dictionaries - The Art and Craft of Lexicography, second edition. Cambridge University Press. 2001.

Morgan, Gary, and Bencie Woll, eds. Directions in sign language acquisition. Vol. 2. John Benjamins Publishing, 2002.

Nakamura, Karen. 28.3.2008. Deaf Resource Library: About American Sign Language. http://www.deaflibrary.org/asl.html (6.5.2017.)

McKee, David, McKee, Rachel, Pivac Alexander, Sara, Pivac, Lynette, and Vale, Mireille. Online Dictionary of NZSL. 2011. Wellington: Deaf Studies Research Unit, Victoria University of Wellington. Retrieved from http://nzsl.vuw.ac.nz/ (8.5.2017.)

Prillwitz, Siegmund, and Hamburg Zentrum für Deutsche Gebärdensprache und Kommunikation Gehörloser. HamNoSys: Version 2.0; Hamburg Notation System for Sign Languages; An Introductory Guide. Signum-Verlag, 1989.

Schermer, Trude. The role of mouthings in Sign Language of the Netherlands: Some implications for the production of sign language dictionaries. The Hands are the head of the mouth: The mouth as articulator in sign languages. 273-284. 2001.

Šegota, Ivan; Šendula-Jengić, Vesna; Herega, Damir; Petaros, Anja; Conar, Jevgenij. Gluhi i znakovno medicinsko nazivlje: kako komunicirati s gluhim pacijentom Zagreb. Medicinska naklada. 2010.

Singleton, David. Language and the lexicon. An introduction. Arnold, 2000.

Sutton, Valerie. Lessons in Sign Writing: Textbook. SignWriting, 1995.
Stokoe, William C. Studies in Linguistics: Occasional Papers 8. Sign Language Structure: An Outline of the Visual Communication System of the American Deaf. Linstock Press, 1960.Svensén, Bo. A handbook of lexicography: the theory and practice of dictionary-making. Cambridge University Press, 2009.
Swanepoel, Piet. Towards a framework for the description and evaluation of dictionary evaluation criteria. Lexikos 18.1 2008.
Tarczay, Sanja et al. Znak po znak 1, 2, 3: udžbenik za učenje hrvatskog znakovnog jezika Zagreb: Hrvatska udruga gluhoslijepih osoba „Dodir". 2006–2007.
Tarczay, Sanja. Gluhosljepoća-jedinstveno oštećenje. // *Ljetopis socijalnog rada*. 14, 1 (2001), str. 143-153.
Tennant, Richard A., and Marianne Gluszak Brown. The American sign language handshape dictionary. Gallaudet University Press, 1998.
Trask, Robert Lawrence. Temeljni lingvistički pojmovi. Zagreb: Školska knjiga, 2005.
Van Herreweghe, Mieke., Slembrouck, Stefaan. &Vermeerbergen, Myriam (Eds.) Woordenboek Nederlands — Vlaamse Gebarentaal, Vlaamse Gebarentaal — Nederlands. Retrieved from http://gebaren.ugent.be/ (10.5.2017.)
Varantola, Krista. "Use and usability of dictionaries: common sense and context sensibility?." Corréard, Marie-Hélène (ed.) (2002): 30-44.
Zgusta, Ladislav. Manual of Lexicography. The Hague/Paris: Mouton, 1971.
Zwitserlood, Inge. "Sign language lexicography in the early 21st century and a recently published dictionary of Sign Language of the Netherlands." International Journal of Lexicography 23.4. 443-476. 2010.
Zwitserlood, Inge, et al. Issues in sign language lexicography. *The Bloomsbury Companion to Lexicography. London: Bloomsbury*. 259-283. 2013.

A New Project – Croatian Web Dictionary MREŽNIK

Lana Hudeček
Institute of Croatian Language and Linguistics
Republike Austrije 16, 1000 Zagreb, Croatia
lhudecek@ihjj.hr

Milica Mihaljević
Institute of Croatian Language and Linguistics
Republike Austrije 16, 1000 Zagreb, Croatia
mmihalj@ihjj.hr

Summary

A new project Croatian Web Dictionary financed by the Croatian Science Foundation is presented. The main goal of the project is to create a monolingual corpus-based dictionary of Standard Croatian compiled in accordance with contemporary findings of computational linguistics. The authors explain the importance of such a project, its methodology, and expected results. As entries in Mrežnik will be connected with many other databases from the Institute of Croatian Language and Linguistics an overview of these databases and other resources is given.

Key words: corpus-based dictionary, Croatian language, e-lexicography, web dictionary

Introduction

Contemporary lexicography is primarily e-lexicography1 and that will undoubtedly be its future as the range of possibilities of new digital technologies cannot be compared with the limited possibilities of printed dictionaries. Since 1990s e-lexicography has developed rapidly and this led to the appearance of the first online dictionaries (before that, dictionaries were stored on CDs and DVDs). This development was followed by research in e-lexicography and corpus linguistics. There are many online dictionaries of many languages in Europe, e.g. elexiko (http://www.owid.de/wb/elexiko/start.html) of the Institute of German Language, Wielki słownik języka polskiego (http://www.wsjp.pl/) of the Institute of Polish Language, Swedish online dictionary (http://spraakbanken.gu.se/karp), etc. It is important to note that in some Slavic countries (e.g. Poland) the

[1] More about this see in the paper Jermen, Kraus, Starčević Stančić (2015) and Štrkalj Despot, Möhrs (2015).

development of an online dictionary of the national language is considered a project of national importance.

In Poland, for instance, such a dictionary has been created in several phases starting in 2006, from 2008 with the support of the Polish Ministry of Science and Higher Education, and from 2013 to 2018 with the support of the National Programme for the Development of Humanities (Narodowy Program Razwoju Humanistyki). Among Slavic languages, Croatian is one of the few languages without a scientifically compiled online dictionary of the national language.

In Croatia, the Croatian Language Portal (HJP) exists, with an online dictionary which is the result of collaboration between Novi Liber and Srce (http://hjp.novi-liber.hr/). However, this dictionary was not compiled as an online dictionary, but an online version of the already printed dictionary published by the publishing house Novi Liber, and sold in the printed version for the last 15 years. There are some other monolingual dictionaries of Croatian language available digitally, such as Prvi školski rječnik hrvatskoga jezika (The First Dictionary of the Croatian Language) by Ankica Čilaš Šimpraga, Ljiljana Jojić, and Kristian Lewis (2008) and Veliki rječnik hrvatskoga jezik (The Big Dictionary of the Croatian Language) by Ljiljana Jojić et al. (2015), but they are not corpus-based, based on the principles of e-lexicography, normative, and publically available. Apart from the aforementioned dictionaries, the following Croatian language lexical resources are available: Wječnik (https://hr.wiktionary.org/wiki/)2, CroWN, the Croatian Wordnet (http://meta-share.ffzg.hr/repository/browse/croatian-wordnet), Meta-Net.HR (http://ihjj.hr/metafore), Struna (Croatian Special Field Terminology of Croatian Repository) (http://struna.ihjj.hr), Croatian Lexicographic Heritage Portal (http://crodip.ffzg.hr/), and Croatian Terminology Portal (http://nazivlje.hr/), which searches Struna, dictionaries, glossaries, and lexicons from the Lexicographic Institute Miroslav Krleža3. There are also many multilingual terminological databases including Croatian as one of the languages (e.g. EMITEL – e-Encyclopaedia of Medical Physics and Multilingual Dictionary of Terms, http://www.emitel2.eu, Multilingual Archival Terminology, http://www.ciscra.org/mat/mat/termlist/l/Croatian, Meta-Share, http://meta-share.ffzg.hr/repository/search/, Microsoft Terminology Collection https://www.microsoft.com/Language/en-US/Search.aspx, etc.

From this short overview, we can conclude that although there are many language resources for Croatian available online there is no monolingual corpus-based dictionary of Standard Croatian compiled in accordance with contempo-

[2] *Wječnik* is the Croatian version of *Wiktionary*. *Wiktionary* is a collaborative international project to produce a free-content multilingual dictionary. It aims to describe all words of all languages. It is designed as the lexical companion to *Wikipedia*. *Wiktionary* is a wiki, which means that everybody can edit it.

[3] More about this see in the paper Jermen, Kraus, Starčević Stančić (2015).

rary standards of computational linguistics. For the purpose of creating such a dictionary, detailed research in e-lexicography is required. The final and primary goal of this project is to develop a corpus-based Croatian Web Dictionary. The logo of the project and of the dictionary is:

Figure 1. Logo of the project

Methodology

The Croatian Web Dictionary will include dictionary entries which have accentuated entry words, grammatical definitions, and grammatical blocks, together with accentuated word forms, detailed definitions with appropriate examples for individual meanings, the most frequent collocations and idioms, synonyms and antonyms. The dictionary will be normative in nature, which is indicated by that fact that the project will include the writing of three hundred short linguistic advice entries (up to several sentences long). Entry words and collocations which are not recommended for use will be linked from the basic text of the e-dictionary to linguistics advice entries. Some very common anglicisms will also be connected with the portal *Bolje je hrvatski* (Better in Croatian) compiled also by team members in which Croatian words and phrases are suggested for some commonly used anglicisms. Three thousand dictionary entries will include definitions for elementary school children, and one thousand dictionary entries will include definitions for learners of Croatian as a foreign language. Dictionary for schoolchildren will be illustrated and definitions from that dictionary will also be used on the website *Hrvatski u školi* (Figure 2).

Figure 2. A crossword puzzle based on *Prvi školski pravopis* on the website *Hrvatski u školi*

Conjunction dictionary entries will include descriptions of conjunction groups and modifiers. Ktetic and ethnic dictionary entries will include links to the ethnic and ktetic repository, and some most common terms will be linked to the terminological database Struna (struna.ihjj.hr) also compiled at the Institute of Croatian Language and Linguistics. During the compilation of the online dictionary, two types of activities will take place: 1. activities connected with the development of computer and computational linguistic prerequisites for the compiling of an online dictionary and 2. activities connected with lexicographic data processing. The dictionary will be written using the TLex software package, a professional software application for compiling dictionaries adapted to the needs of the project (designing the entry fields according to the dictionary entry model developed by the editors of the dictionary). SketchEngine will be used to search the corpora. The dictionary will be based on these two corpora: the Croatian Web Corpus hrWaC (http://nlp.ffzg.hr/resources/corpora/hrwac/) and Croatian Language Repository (riznica.ihjj.hr). The obvious problem with the methodology of MREŽNIK is that MREŽNIK will be based on the large unbalanced corpus the Croatian Web Corpus and on Riznica (which is a smaller corpus containing many texts from literature and from older periods). However, MREŽNIK will be only corpus-based and not corpus-driven and the lexicographers will select freely data from the corpus as well as from other Croatian dictionaries. The compilation of the dictionary will be based on designing word "sketches" (WordSketches) for each corpus separately, the prerequisite of which is a developed grammar sketch (SketchGrammar), the application of the GDEx module for finding appropriate examples in the corpus, checking individual entries using a morphological lexicon (https://www.clarin.si/repository/xmlui/handle/11356/1056) and exporting data from TLex in order to use it in Web applications and repositories clarin.si and GitHub.

Project results

Thus, the result of the MREŽNIK (Croatian Web Dictionary) project will be a free, monolingual, hypertext, searchable, online dictionary of Standard Croatian with ten thousand dictionary entries compiled during the four-year period. The dictionary entries will contain links to repositories which will be created as a part of this project and compiled simultaneously with the dictionary as well as with repositories which have already been compiled within other projects conducted at the Institute of Croatian Language and Linguistics. The project will include:

1. Compiling a dictionary of ten thousand dictionary entries (with accentuated entry words and accentuated word forms in the grammatical block, with detailed definitions (also definitions for schoolchildren and definitions for foreigners), examples, antonyms, synonyms, collocations and idioms, male-female relations, pragmalinguistic explanations, etc. At the end of the project, the dictionary will be available online on the domain rjecnik.hr.

2. Developing repositories and connecting them with the basic dictionary: The Linguistic Advice Repository of (300 linguistic advice entries), The Conjunction Repository (for all conjunctions in the dictionary), The Idiom Repository (50 entries), The Ethnics and Ktetics Repository (300 inhabitant names and adjectives).

biciklista > biciklist

U hrvatskome standardnom jeziku imenice koje završavaju na *-ist* muškoga su roda i pogrešno ih je izgovarati i zapisivati s *-a* na kraju kao imenice ženskoga roda. Takve su, primjerice, imenice *aktivist, alpinist, biciklist, daltonist, harfist, idealist, iluzionist, kroatist, okulist, optimist, perfekcionist, pesimist, pijanist, šahist.* Pogrešno je *ići okulisti , vidjeti harfistu, razgovarati o šahisti,* a pravilno je *ići okulistu, vidjeti harfista, razgovarati o šahistu.*

	N	G	D	A	L	I
jednina	biciklist	biciklista	biciklistu	biciklista	biciklistu	biciklistom
množina	biciklisti	biciklista	biciklistima	bicikliste	biciklistima	biciklistima

NE	aktivista	alpinista	biciklista	daltonista	harfista	idealista	okulista
DA	aktivist	alpinist	biciklist	daltonist	harfist	idealist	okulist

Figure 3. An entry from the database of language advice *Jezični savjeti* (http://jezicni-savjetnik.hr/)

3. Connecting the basic dictionary with other online resources which are currently being developed at the Institute of Croatian Language and Linguistics: *The Verb Valence Repository*, *The Collocation Repository*, *The Croatian Terminology Repository* (Struna) (Figure 4), *The Croatian Metaphor Repository*, website *Bolje je hrvatski* (Figure 5).

4. Compiling a reverse dictionary. Although the reverse dictionary is planned for the last year of the project as it has to contain the complete word list of ten thousand words, a pilot reversed dictionary has already been compiled by Josip Mihaljević using a test word list. This dictionary will be a working tool for all lexicographers on the project and it will become available online at the end of the project (Figure 6).

Figure 4. An entry from the terminological database *Hrvatsko strukovno nazivlje STRUNA* (http://struna.ihjj.hr/naziv/zubna-proteza/13383/#naziv)

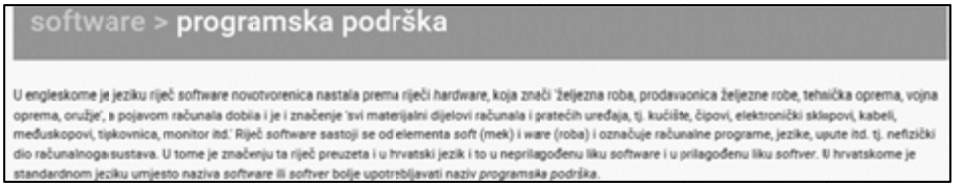


Figure 5. An entry from the database *Bolje je hrvatski* (http://bolje.hr/)

Figure 6. Pilot version of the reverse Croatian dictionary

5. Extensive research on e-lexicography, training, and dissemination of acquired knowledge as well as a contribution to the area of e-lexicography, which did not receive the attention it deserves in the Croatian scientific community so far, is needed.

Acknowledgement

This paper is written within the research project Croatian Web Dictionary – Mrežnik (IP-2016-06-2141), financed by the Croatian science foundation.

Literature

Blagus Bartolec, G.; Hudeček, Lana; Jojić, Ljiljana; Kovačević, Barbara; Lewis, Kristian; Matas Ivanković, Ivana; Mihaljević, Milica; Miloš, Irena; Ramadanović, Ermina; Vidović, Domagoj. (2012). *Školski rječnik hrvatskoga jezika*. Institut za hrvatski jezik i jezikoslovlje – Školska knjiga. Zagreb.

Hudeček, L. Jozić, Ž., Lewis, K., Mihaljević, M. (2016). *Prvi školski pravopis*. Institut za hrvatski jezik i jezikoslovlje.

Jermen, N., Kraus, C., Starčević Stančić, I. (2015). Lexicography and Encyclopaedistics in the Digital Environment. Infuture 2015. The Future of Information Sciences. E-Institutions Openness, Accessibility, and Preservation, Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, 65-75.

Štrkalj Despot, K., Möhrs, C. (2015). Pogled u e-leksikografiju. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 41(2), 329-353.

E-dictionary for Asian Languages

Marijana Janjić, Marko Požega, Dario Poljak, Sara Librenjak, Kristina Kocijan
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučić 3, Zagreb, Croatia
marijanajanjic@yahoo.com, mpozega1@ffzg.hr, dpoljak1@ffzg.hr,
sara.librenjak@gmail.com, krkocijan@ffzg.hr

Summary

In this paper, authors present an e-dictionary of Asian languages designed for students whose primary language is Croatian. The project started with six Asian languages (Hindi, Japanese, Chinese, Korean, Persian and Sanskrit) all of which are using Croatian as a meta-language. Entries in Asian languages include several attributes (translation, Latin letters, grammatical notations, sentence with examples) that mainly depend on the language chosen for translation. Croatian entries are linked to the Croatian language portal where the meaning, word forms and categories are provided. The growth of the dictionary is largely dependent on students' participation.

Key words: e-dictionary, student's dictionary, Asian languages, Croatian language

Introduction

In order to get any work done, a student of Asian languages in Croatia must carry around several heavy dictionaries to each of the classes s/he takes. The survey we conducted among Croatian students of Asian languages at the beginning of this project revealed that they mostly use dictionaries with English-Asian language combinations. Apart from English-Croatian dictionary, not one of the used dictionaries includes Croatian. The reason behind the usage of English oriented dictionaries is (next to the fact that there are no Croatian-Asian languages dictionaries[1]) that most of the study materials are also available in English rather than Croatian language, although we are talking about Croatian students studying in Croatia.

This situation inspired us to create an online dictionary for beginner learners of Asian languages, primarily including those Asian languages that are being studied at the University of Zagreb such as Hindi, Japanese, Chinese, Korean, Persian and Sanskrit.[2] We have opted for the e-version rather than the printed

[1] Apart from the Croatian-Japanese dictionary (2006) and Croatian-Turkish dictionary (2014).

[2] Turkish is also available at the University, but, so far, we do not have any resources for it.

version of such a dictionary for several reasons. The first reason was the time needed to produce such a book. The time needed to produce the e-version of a dictionary was relatively shorter compared to the printed dictionary. In addition, the e-version of a dictionary allows us to add, edit and delete as many entries as needed and to add new languages as well, all of which cannot be done with the printed version. Its on-line availability and free access allows students to 'carry it around' and have it available wherever and whenever they need, either via the website or as an android application.

In the following sections, we will shortly describe the content of this dictionary from the user's and administrator's perspectives.

The e-Dictionary Content

User's perspective

We have opened the dictionary with 5.953 Croatian entries. However, this number is not evenly distributed among Asian languages. The languages with the most entries at the moment are Hindi (2 132) and Japanese (1 324) while Persian has the least (156). In between them are Sanskrit (1.028), Chinese (762) and Korean (668). At the moment, some words have been translated into only one and some into more Asian languages. It is possible to compare translations across different Asian language. This can be especially useful for learners of similar languages like Hindi and Sanskrit (Figure 1).

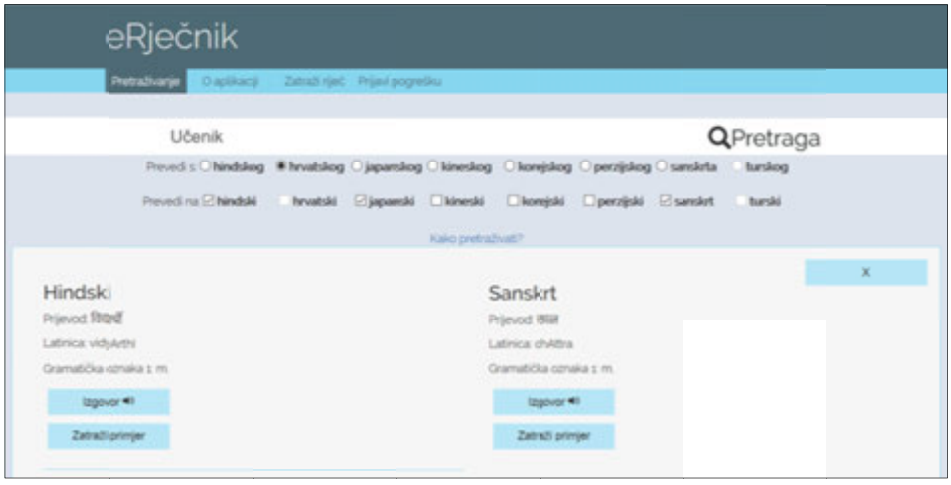


Figure 1. Comparing the Croatian word 'student' (hr. *učenik*) in Hindi and Sanskrit

Envisioned as a learner's dictionary (beginner's stage), it consists of elementary information that a student would need: (1) entry in one of Asian languages, (2)

grammatical information, if needed or available, (3) Croatian equivalent, (4) example(s) of use for each meaning.
The first three elements are what any reader would expect to find in a dictionary. The last element is, what we believe, the key element that separates this e-dictionary from others. Among many e-dictionaries available for the Asian languages that we came across, examples are not included in their structure. Our survey, however, showed that examples are of great importance to students. Even as experienced users of these languages, we are often at a loss when learning a new word from a dictionary trying to actually use it in a sentence.
The dictionary also provides a translation to Croatian if the searched item is in one of the offered Asian languages. However, we did not provide any additional information for Croatian words. With the permission of Croatian language portal[3] that has an online database of Croatian, we have linked the words to their page where all the existing information for that word can be found.
The important role in building a learner's dictionary must also be given to learners themselves. For that reason, we have enabled to our users two-way communication with language administrators. This way they can ask for a new entry, new example for particular entry or suggest their own example for an existing entry.

Administrator's perspective
Administration of our e-dictionary allows us several possibilities: adding, editing and learning about the dictionary usage. The new words can be added either as single entries, mainly upon the user's request, or as a batch prepared by language administrators. All the entries that have already been stored in our database can be accessed and edited (i.e. fixing the errors, adding additional meanings, attributes, examples). The third section of the administration portal enables us an insight into the dictionary usage statistics, making sure that no personal data is collected or stored. We are thus familiar with the data on how many searches were made each day, what words were searched for, in what languages, if the search was successful or not. Additionally, we are using Google Analytics to check user's demographic as well as how many of them are new and how many returning users.

The Technology behind the e-Dictionary?

Developing a web application that is useful, intuitive and visually compelling is a big undertaking and is therefore commonly divided up in frontend and backend development. Our process was no different while developing e-dictionary as it encompasses a lot of information that has to be displayed in a suc-

[3] http://hjp.znanje.hr/

cinct and readable manner whilst also implementing a number of features for both users and administrators.

Frontend development is commonly done with HTML, CSS and JavaScript. By adhering to some of the newest standards and technologies, we have been able to speed up the development and debugging time. In that regard, we have used CSS pre-processor SASS (Syntactically Awesome Stylesheet) which extends CSS functionality by providing many features from the simplest ability in creating reusable variables to creating whole snippets of extendable code.

To make the e-dictionary as fluid as possible, we have also opted for asynchronous bi-directional calls to the server. In this way, the user's page does not need to refresh each time the search is made. We have accomplished this by using AJAX (Asynchronous JavaScript) calls that were handled with a precompiled JavaScript library, jQuery. It has simplified the convoluted approach of writing reusable AJAX functions and has allowed us to efficiently request and receive data from the server in real time. With jQuery, we are able to send the searched item, source and target languages to the server. The results are returned in the form of an encoded JSON.

The entire e-dictionary system is built on the top of the Model–View–Controller (MVC) software architecture. The Model is the application object usually related with database entity, the View is its screen presentation, and the Controller defines the way the user interface reacts to the user input. Because e-dictionary uses pivot language, database must contain transfer tables between every language and pivot language. This kind of relational model requires smooth data manipulation using models and MVC.

Open source PHP framework, Laravel, offers full stack development environment following MVC architectural pattern, which we used to design data access layer (backend) and presentation layer (frontend) of our application. Furthermore, by using Laravel router we easily implemented AJAX web development technique resulting in zero page reloadings. Such approach increases user experience and leaves prepared setup for light JSON application programming interface (API) implementation. Our API is currently being used by android mobile application which is still in a development stage. Also, our API could be used by other applications which want to implement results of our dictionary search into their own products.

Conclusion

The development of this small e-dictionary has showed how cooperation of experts in different fields (linguists, lexicographers, information scientists) can be fruitful. As a result of scientific excellence, a new free on-line tool had been devised for Croatian students while the research team will gain further insight into students' habits and needs. The project has also showed how students can be and should be included in the development of tools oriented towards their needs.

HER.IT.AGE

The new Information Technologies at the Service of Historical and Cultural Heritage and Tourism Promotion

Basma Makhlouf-Shabou
Information Science Department, Geneva School of Business Administration,
University of Applied Sciences and Arts Western Switzerland
Rue de la Tambourine 17, Bât. B, 1227 Carouge, Switzerland
basma.makhlouf-shabou@hesge.ch

Maria Sokhn
University of Applied Sciences and Arts Western Switzerland,
HES-SO Valais-Wallis
Technopôle 3, 3960 Sierre, Switzerland
maria.sokhn@hes-so.ch

Summary

City-Zen is an interactive spatio-temporal knowledge-browsing platform that aims to valorise cultural heritages. Considering the explosive growth of information, data and knowledge sharing can ensure valuable interdisciplinary applications. While many organizations propose relevant data sets, they are hardly accessed, analysed and reused because of the formats inconsistency and the inappropriate information browsing and visualization. The goal of the project is to valorise the existing cultural heritage through a citizen centric design platform. Based on information qualities, the use case of this project involves a user willing to discover the history of a region and to embark in a cultural journey in the past. This paper exposes the main functionalities City-zen application and shows how those latest should reach the different user needs. It illustrates how historical and cultural heritage valorisation can take advantage from the advancement of new technologies, multimedia & mobile and how those technologies should promote tourism services.

Key words: Open linked data, cultural heritage, information quality, citizen, tourism services, information services, spatio-temporal browsing, knowledge sharing

Introduction

Given the complexity of modern life, interest in long-term vacations is a trend that seems to be weakened. Unlike conventional tourist stays and tourist attractions, short-term stays are no longer limited to religious buildings, political buildings or military buildings. Rather, the entire neighbourhoods of tourist

destinations are now the subject of beautification strategies whose objective is to increase their attractiveness.
Additionally, the cultural heritage is nowadays more and more influenced by the access economy. The access economy is an approach where accessibility is more important that owning. This approach uses is mostly based on mobile technologies and avoid third parties. This terminology arose with the sharing economy advent. Several works were based on this new paradigm to apply it on cultural heritage marketing (Cucchiara, 2017, Rialti, 2016, Wroblewski, 2017).
In our work we believe that increasing the availability of relevant information about cultural and historical heritage could be one of those strategies. We aim to take advantage of digital technologies to promote the cultural heritage. For example, during his stay, or even before arriving to destination, tourist needs to know a lot and every think about the area or the city he wants to visit. Even though citizens (residents) may have valuable information for other citizens who visit their region as tourists, it is barely usable with today's information systems because of the inevitable information resources dispersed. This paper presents the functionalities of City-zen including how this platform proposes to assume the challenge of information quality assessment.

Concrete needs & user scenario

In order to describe the goal, we first present a use case scenario. The scenario is seen from three different points of view according to two different potential roles of a citizen within our platforms: 1) tourists and 2) residents.

1) Citizen as a tourist

Anne, living in Geneva, is visiting the city centre of Sion. When she organized her sightseeing itinerary at home, she searched the web in advance for information about the city. She browsed existing tools and web sites such as tripadvisor, social networks, etc. in order to gather some information about the city. After her arrival at the destination, she decided to have a walk in the city and discover its story and culture. She is standing in front of the white facade of the town hall. She might be interested in a variety of questions regarding the building and the local environment:

- What building is this, when was it built?
- What was the role of this building 50 years / 200 years / 500 years ago (at any other date in the past); did something of interest happen at this place?
- Who is the architect of the building?
- Where are other relevant sights located? How can I get there?
- Are there photos (e.g., from other perspectives or with different scenes in the foreground, etc.) of the gate on the web?

Anne holds a smart phone, but in order to get information along the lines of the questions above, she has to manually query different sources on the web (and she has to identify these sources first). For some questions, partial support already exists (e.g., spatio-temporal services from Google showing restaurants and hotels on a map), but in most cases, no existing and in particular no integrated solution can be consulted.

2) Citizen as a resident

Patricia is a resident of Sion since 1954. She knows many anecdotes and stories about her city and she has a lot to tell about places, buildings, people, practices (such as the famous Sion carnival), etc. Patricia has also some inherited age-old objects and objects related to Sion traditions (her first drums when she was 10 years old). With the City-Zen platform, Patricia can be an actor in valorising the cultural and intangible heritage by participating in the knowledge sharing. Patricia can virtually situate her age-old objects on the map using the City-Zen platform and describe them with photos and texts. She can also propose a time slot where she would offer to a citizen (tourist or not) to pass by her house and have a look on these objects.

City-Zen Answering User Needs

With the City-Zen platform, Anne will be able to directly browse and/or submit queries of different types:

Simple location and spatial queries

Using the GPS coordinates of her current location, Anne will be able to identify the building she is currently looking at and get access to basic information regarding this building, combined from several data sources on the web. She will search for similar buildings (or buildings that take a similar role) in the vicinity of the current location – e.g., where are the other city gates of Sion located. This query type will also show other relevant information on a map like hotels, restaurants, cinemas, and their most relevant and most recent ratings.

Temporal queries

On the basis of information from various sources that have been integrated beforehand into the City-Zen platform, Anne will be able to query details of the building's history, if available together with photos from different stages of the building (in case it has been incrementally extended over the years). Moreover, she will also get information on the building's environment at different points in time, and on historic events that took place there.

Profile-based push notification

Anne may personalize the City-Zen platform by describing her interests and she may activate the "radar-mode" of City-Zen. This mode pushes notification to Anne whenever she is close to such a point of interest.

The City-Zen platform offers Patricia the object linking service which, in a transparent way for her, links the objects she described to existing information related to that object. As an example, Patricia will be able to link the drum object to the concept of drums explaining that it is an instrument, giving details about its story, and she will also be able to link this object to the traditional art of drumming in the Sion carnival. Two facilities have been used then:

- **Heterogeneous multimedia integration**: Patricia can access the platform and choose the type of information she wants to share: she may upload an old photo, she could also write an anecdote, she may also upload a video, etc.
- **Linked data integration**: Patricia pushes the information related to her drums, and the City-Zen platform proposes her to link this description to the concept of drums and to the page web describing the traditions of the drums in Sion. Patricia can ignore this proposition or accept it. In the latter case, the object will be linked to other data already existing objects on the web.

Considering the needs below, City-Zen platform was built on three main modules:

1. Data integrator module responsible of gathering distributed and heterogeneous data. City-Zen makes use of existing approaches to crawl and link accessible or user-generated content;
2. Data analyser module responsible of linking, mapping and cleaning data offering advanced spatio-temporal and personalized queries. City-Zen takes advantage of spatiotemporal information and the web of data approaches.
3. Data visualization module responsible of the adaptive and profile aware knowledge visualization and navigation interfaces. City-Zen takes advantages of existing approaches of Knowledge visualization and valorisation.

Methodological considerations

A rigorous and systematic literature was realized in order to identify relevant approaches for implementing the previous architecture. A knowledge sharing approach was explored through: Data integration (Heterogeneous databases, Knowledge linking and Data analyser), Temporal multimedia browser (Data and knowledge based queries and profile-context based push notification), and Knowledge visualization (Personalized based knowledge visualization and mobile adapted visualization and navigation).

Even if each of the described components may require a distinct research and development approach, for the overall methodology, the "Design Science Research Methodology" is used. This process includes the following steps: identifying issues and motivation in order to accurately plan the implementation of our proposed solution.

Defining the goals and objectives: given the issues and motivations defined earlier we infer the objectives of the solution in order to answer the needs and solve the problems. The objective of this step is to define the criteria against which we will be able to evaluate the solution once it is implemented.

Designing & implementing: designing the system architecture and its components concerns this step. Some elements of the architecture can be implemented just by reusing existing knowledge and technologies, but others, and given the research questions, will need to be developed.

Testing: this step is directly related to the previous and it starts sometimes with the step of design. It concerns the design and development of the proof of concept that will demonstrate the defined use cases.

Evaluating and analysing: this step is concerned with assessing whether the proposed solution meets its defined objectives. Objectives of this step are:

- Technical: to evaluate the accuracy of the results as an objective measure, the flexibility to browse the information space and the performance of the queries.
- Human: to evaluate the usability of the system and the accuracy of the results as a human evaluation given the human mental model for the decision.
- Disseminating: the objectives of this step are to disseminate knowledge through scholarly publications. It will take the form of regular reports and deliverables as well as peer reviewed articles in conferences related to diverse domains: web of data, knowledge visualization, data integration, tourism management and innovation.

City-zen module configuration

As mentioned, the goal of the project is to valorise the existing cultural heritage through a citizen centric design platform. The use case of this project involves a user willing to discover the history of a region and to embark in a cultural journey in the past. As shown in Figure 1, City-zen addresses (1) the data integration by making use of existing approaches to crawl and link accessible or user-generated content, (2) a novel approach of data analysis of both assessment methods of data quality and spatio-temporal information, (3) the data visualization by taking advantages of existing approaches of knowledge visualization.

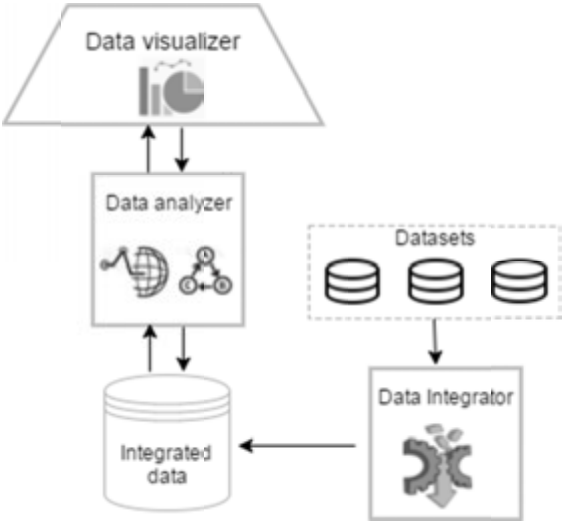


Figure 1. City-zen Architecture

The first part consisted in building the core component of the architecture presented in Figure 1. The platform takes advantages of existing data coming from existing resources. Heterogeneous data collected from different resources have been transformed to a RDF triple store (Figure 2), following a specific data model designed and developed for the needs of City-Zen project. The data model is depicted in Figure 3. For implementing this, different tools have been combined. This aspect is combined with the dimensions of quality applied to historical archives. It focuses mainly on the definition of quality dimensions and the methods that assure their measurement based on specific indicators and variables in the context of historical archives (Makhlouf Shabou, 2011, 2014).
The platform is oriented towards citizens and integrates their data, as they will become available. The platform offers adequate methods and appropriate visual interfaces in order to answer the goal of our project.

Information quality assessment challenge: criteria and metrics

To enable City-zen platform assessing information quality and information sources, a set of criteria has been identified by using different documentation related to the measurement of quality dimensions applied to electronic records and archives. Two types of information quality criteria were identified: 1) the general information qualities applied on information sources and 2) the specific qualities that could be applied on information content.
The identification of information quality criteria is based on two recent researches: the first is the a doctoral study *on the archival appraisal criateria* and archival quality metrics (Makhlouf Shabou, 2011) and the second is QADEPs: a

study on the assessment of digital archives quality metrics (Makhlouf Shabou, Mellifluo, Rey, 2013; Makhlouf Shabou, 2014). Based on those researches, a selection has been established. Two raisons motivate the use of those of information quality criteria in this City-Zen's mobile application: first, for the mobile application, we are going to use electronical data and we need to applied a type of measurement of the information quality; second, this typology of measurement has been already tested in the Canton of Valais (QADEPs, 2013).

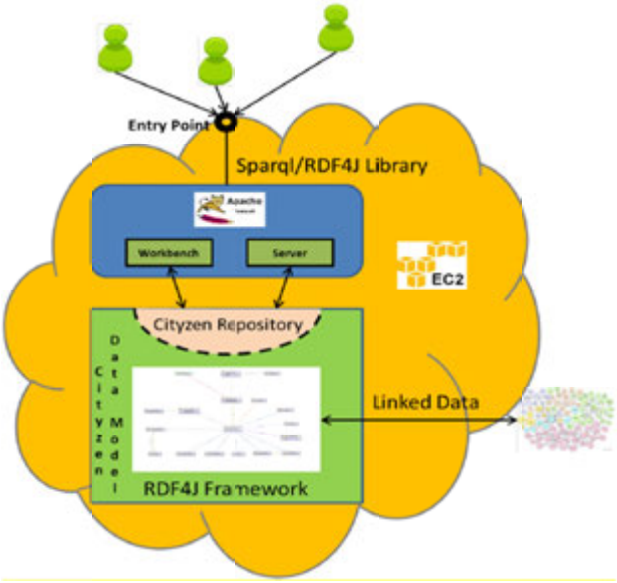


Figure 2. City-Zen RDF triple store design

Information qualities criteria

City-Zen's application is based on several information quality criteria. Those latest represent the generals and specifics information quality criteria that determine his usability for a tourist when he visits the town of Sion in Valais. Each of the criteria is assessed by a scoring from one to five stars: more the information has a high score, more the information will be trustworthy, exploitable and representative for the tourist.

General information qualities

Trustworthiness

The tourist will trust this City-Zen's application if this one gathered all the conditions to win his confidence, this means an authentic and reliable information that will make the tourist exploit it by using it. The Trustworthiness "refers to the ability of a document to gain the trust of the user as the preferred supporting facts source. This quality depends on the authenticity and the reliability and the

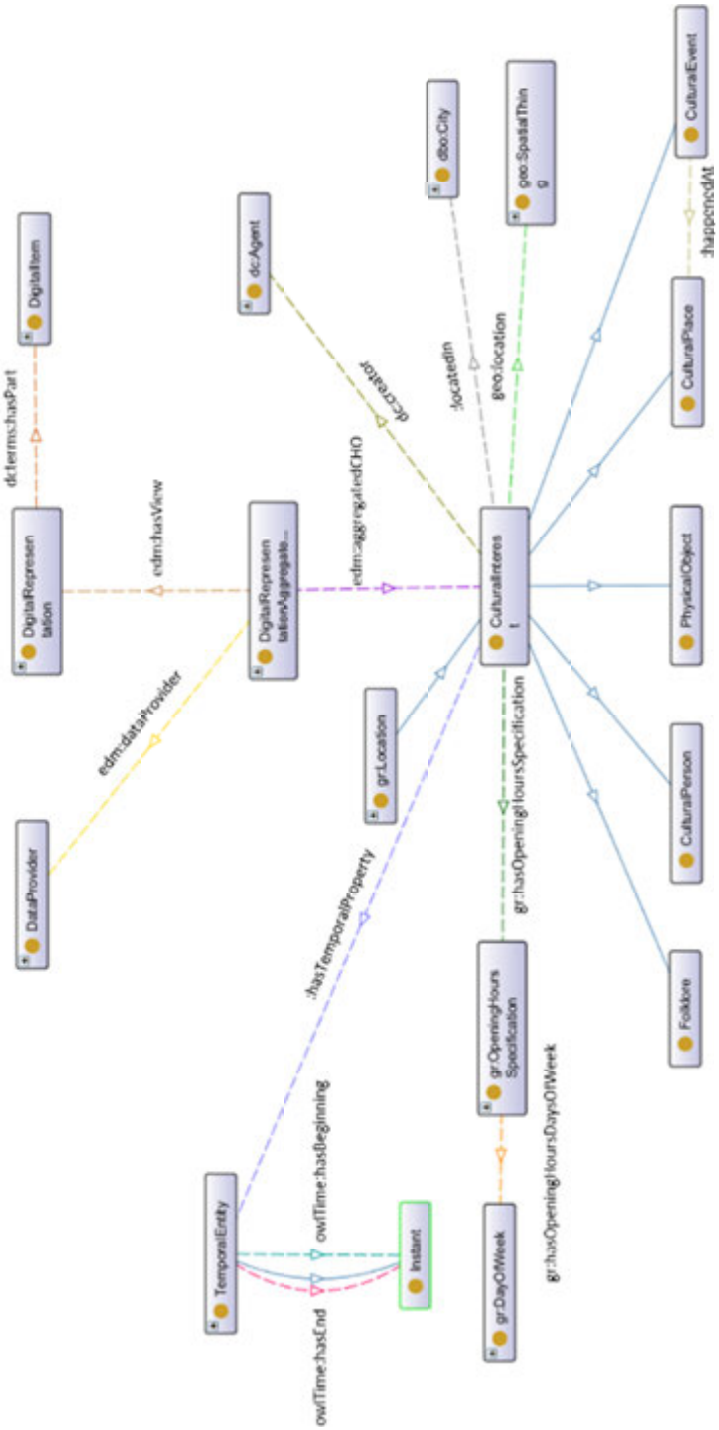


Figure 3. City-Zen data model

durability of these qualities over time" (Makhlouf Shabou, 2011, p. 115; InterPARES 2, 2013). The tourist will take a document with authentic supporting fact sources that gain his trust over a document with no identifiable source that is difficult to know who the creator is.

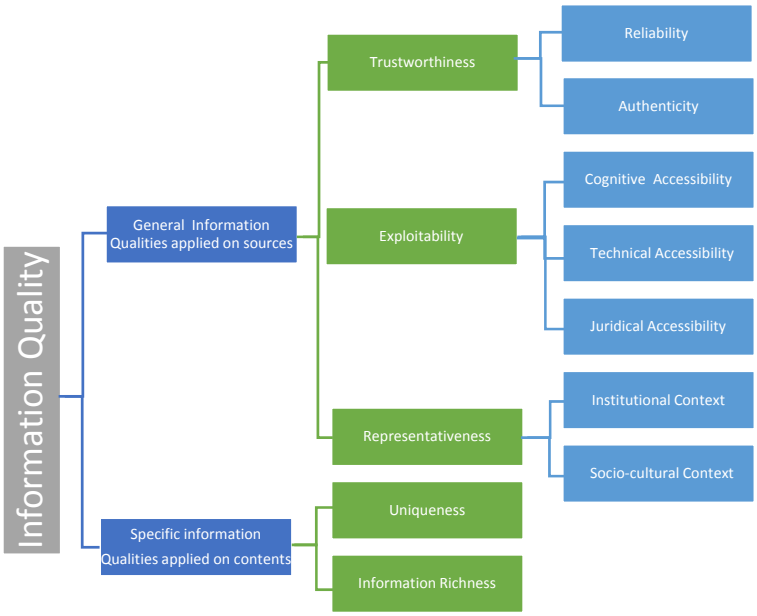


Figure 4. Information Quality Criteria

Exploitability

The Exploitability is the dimension that "refers to the ease of use of a document, thanks to its location, retrievability, diffusion and interpretability. The exploitability depends on three types of document accessibilities: technical accessibility, including physical and material needed for reading; legal accessibility, including regulatory and administrative environments required for the diffusion of document; and cognitive accessibility that guarantees an adequate comprehension and interpretation of document contents" (ISO 15489-1 :2001; Makhlouf Shabou, 2011)

The information diffused is legally accessible because is author's wright free (only this legal open-access type of information will be put in place for City-Zen's application). The information will be easy to comprehend for the tourist but technical information can be also available for the history researchers that wanted to get to know more about Sion.

Representativeness

This criterion "refers to the capacity of the documents to provide a significant testimony of the institutional context in which they were created. This quality depends on two essential elements: the completeness of testimony; and the representativeness of the socio-cultural context in which these documents were created" (Makhlouf Shabou, 2011).

Specific information qualities

The specific content quality criteria are the Uniqueness and the Information Richness that City-Zen's application have to provide to give a rare, complete and precise information for the tourist. In this part, we will describe each specific content quality criteria and a specific example of a tourist practice for each criterion. A scoring is going to be set for each criterion.

Uniqueness

The Uniqueness data quality criterion "describes the fact that each document is related to the others within and outside the fond of which it is a part, and to the creator of the fond by a special relationship, which makes it unique." (AAS Glossary, p.55). In short, this information quality criterion bring attention that there should be no data duplicates reported in the City-Zen's application. Each data will be unique or else the tourist will receive common and a several package of information instead of up-to-date and exclusive information. Asserting uniqueness of the entities within a data set implies that no entity exists more than once within the data set and that there is a key that can be used to uniquely access each entity within the data set. For example, in the City-Zen's application, each information diffused must appear once and be assigned a unique identifier that represents that information across the client applications. The dimension of Uniqueness is characterized by stating that no entity exist more than once within the data set. When there is an expectation of uniqueness, data instances should not be created if there is an existing record of that entity (Loshin, 2006).

Information richness

With the City-Zen's project, we discuss about the topic of information richness applied to the lowest type of rich information that is the numerical documentation (Kurstedt, 2000). This criteria will be assessed on the basis of the method used by Daft and Lengel (1984), which proposes 5 metrics to consider: 1) the medium (by distance, or face to face); 2) the speed of feedback; 3) Channel (audio, visual; multimedia); 4) source; 5) language.

The highest source in Information Richness is when there is a combination of an official source (impersonal) like information of the Tourism Office in Sion and a citizen's testimony (personal) that's not qualified as an official source.

Information qualities metrics

The scoring is based on 3 principals: 1) gradual logic; 2) accumulation of conditions and 3) the lowest level is 1 not 0. Considering those principals, we identified mainly 5 levels. The Figure 5 presents an illustrative example of such method applied on information reliability which is a part of trustworthiness.
The detailed description with specification of different information quality levels is available in the table in Annex 1.

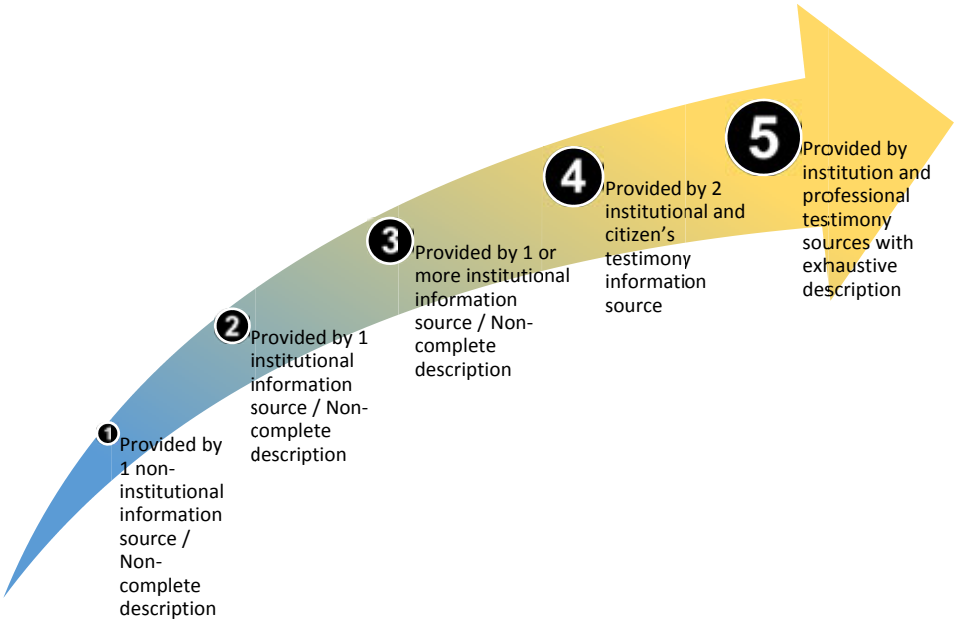


Figure 5. Information reliability: example of applied metrics

Function Display and Visualization

The City-zen radar mode is depicted in Figure 6. The prototype application uses the user's location and pre-setup filters such as radius and user's interests to visualize in a radar map the interesting places or events around him a video in the annexe show the use of this prototype.
We have designed and implemented à prototype of the interface targeting users with mobile devices. The mobile app provides basic searching and visualization of integrated data.
It allows users to run temporal search for historical multimedia data and navigate in more details about the cultural interests (CI) by reading text information, browse images about the CI over time and if available play videos and audios, and read digitized documents. CIs that have geolocation data, they are visualized on the map (Figure 7).

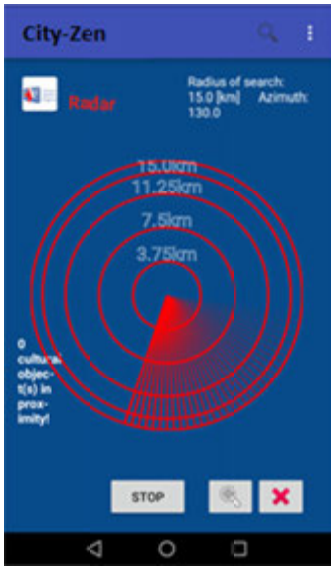


Figure 6. City-Zen radar mode

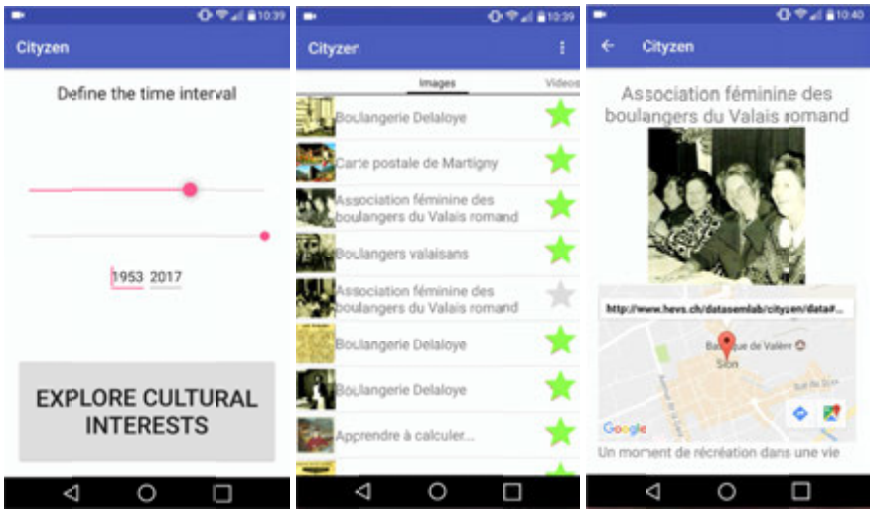


Figure 7. City-Zen mobile app user interfaces: visualization & navigation

Display of information quality assessment and scoring

As shown, the requested information is displayed in this app with a specific rank of quality. At this stage of testing, in order to simplify screen view of this service, we reduced the 5 quality levels in 3 levels: the green star corresponds to fourth and fifth, the blue star indicates a medium quality and the grey star refers to the first and second level (Figure 7).

Conclusion

As explained in this paper, the City-Zen project has proposed an application in a smartphone that offers a resident or a visitor quality information which on one side vulgarize the material heritage (architecture, museum object, historical site, etc.) and immaterial patrimony (culture, oral, culinary traditions). This stage of the project enable the testing of this innovative approach based on open liked data. City-Zen offers through its functionalities an access economy approach that is helpful both for visitors (as they can improve their experience by having full access to aggregated content) and policy makers (as they may promote their cultural heritage through this platform and their city more visible and accessible). On the other hand, several aspects need to be improved such as the section on quality and its metrics to increase the automation of its application. We will also integrate a gamification dimension to motivate citizens to upload their data and share their knowledge. This module will be integrated to the platform in a second phase.

References

Aberer, Karl. *Peer-to-Peer Data Management*. Morgan & Claypool Publishers, 2011.

Barrow, Time. Information Richness Theory. 7th November 2010. http://blog.timebarrow.com/2010/11/information-richness/ (19th May 2017)

Cazes, G. L'émergence d'un nouveau système vacancier : temporalités et territorialités en mutation. *// Hommes et Terres du Nord*. N°2 (2001), pp. 63-70.

Cucchiara, R., & Del Bimbo, A. (2016, September). Bridging the experiential gap in cultural visits with computer vision. In Research and Technologies for Society and Industry Leveraging a better tomorrow (RTSI), 2016 IEEE 2nd International Forum on (pp. 1-4). IEEE.

Cudré-Mauroux, Philippe. Emergent semantics: rethinking interoperability for large scale decentralized information systems. Lausanne : EPFL, 2006.

Cuvelier, Pascal. Anciennes et nouvelles formes de tourisme. Une approche socio-économique. Paris : L'Harmattan, 1998.

Daft, R. L.; Lengel, R. H. Information Richness – a New Approach to Managerial Behavior and Organizational Design. *Research in Organizational Behavior*. Vol. 6 (April 1983), pp. 191-233.

Instagram. https://instagram.com/ (19th May 2017)

International Organization for Standardization. *Information et documentation : Records Management.* Geneva: ISO 15489-1: 2016.

INTERPARES TRUST. *The InterPARES Glossary*. Vancouver: International Research on Permanent Authentic Record in Electronic System (InterPARES), Decembre 2001. http://www.interpares.org/book/interpares_book_q_gloss.pdf (19th May 2017)

Kilany R., Sokhn M., Hellani H., Shabani S. (2016) Towards Flexible K-Anonymity. // Knowledge Engineering and Semantic Web. 7th International Conference, KESW 2016 / Ngonga Ngomo Axel-Cyrille; Křemen Petr (eds). Cham: Springer, 2016.

Lee, Brent. *Authenticity, Accuracy and Reliability: Reconciling Arts-related and Archival Littérature*. InterPARES 2 Project, septembre 2005. http://www.interpares.org/display_file.cfm?doc=ip2_aar_arts_lee.pdf (19th May 2017)

Loshin, David. *Monitoring Data Quality Performance Using Data Quality Metrics.* Redwood City : Informatica, novembre 2006. https://it.ojp.gov/documents/informatica_whitepaper_monitoring_dq_using_metrics.pdf (19th May 2017)

Makhlouf Shabou, Basma. *Measuring the Quality of Records to Improve Institutional Memory*. Cincinnati, IEEE International Professional Communication Conference, 17-19th October 2011.

Makhlouf Shabou, Basma. Le projet QADEPs : un outil au service de la pérennisation des archives publiques. // *De la préservation à la conservation : stratégies pratiques d'archivage* / Hiraux, Françoise; Mirguet, Françoise. Louvain-la-Neuve : Academia l'Harmattan, 2014, pp. 87-98.

Makhlouf Shabou, Basma. *Étude sur la définition et la mesure des qualités des archives définitives issues d'une évaluation*. Montréal : École de bibliothéconomie et des sciences de l'information. Thèse de doctorat, 2011. http://hdl.handle.net/1866/4955 (19th May 2017)

Makhlouf Shabou, Basma. Digital diplomatics and measurement of electronic public data qualities: What lessons should be learned? // *Records Management Journal*. Vol. 25 (2015) Issue: 1, pp. 56-77. http://dx.doi.org/10.1108/RMJ-01-2015-0006 (19th May 2017)

Makhlouf Shabou, Basma; Mellifluo, Laure & Rey, Raphaël (2013). *QADEPs: définition et mesure des qualités des archives et documents électroniques*. Genève : Haute école de gestion, rapport de recherche, 2013.

Notrehistoire. http://www.notrehistoire.ch/ (19th May 2017)

Pearce-Moses, Richard. *A Glossary of Archival and Records Terminology*. Chicago: Society of American Archivists, 2005. http://files.archivists.org/pubs/free/SAA-Glossary-2005.pdf (19th May 2017)

Olivieri, Alex Carmine. Improving Automated Fact-Checking Through the Semantic Web. // *Web Engineering. ICWE 2016.* / Bozzon Alessandro; Cudre-Maroux Philippe; Pautasso Cesare. Cham: Springer, 2016.

Olivieri, Alex Carmine; Sokhn, Maria; Schegg, Roland. Cityzen: a social platform for cultural heritage focused tourism. // *Proceedings of the 8th International Conference on Management of Digital EcoSystems*. Hendaye, 2016, pp. 129-136.

Rialti, R., Zollo, L., Boccardi, A., & Marzi, G. (2016). The impact of technologies on visitors' experience personalization: a case study. Micro & Macro Marketing, 25(2), 251-280.

Rialti, R., Zollo, L., Ciappei, C., & Laudano, M. (2016, July). Digital cultural heritage marketing: The role of digital technologies in cultural heritage valorization. In 2016 Global Marketing Conference at Hong Kong (pp. 1062-1063)

Sokhn, Maria. *Ontology driven framework for multimedia information retrieval in P2P network*. Computational Engineering, Finance, and Science [cs.CE]. Télécom ParisTech, 2011.

Sokhn, Maria; Makhlouf-Shabou, Basma; Olivieri Alex. *Citizens' Platform: an Interdisciplinary Approach for Cultural Heritage Valorisation and Visualization*. Tunis, Colloque international sur les bibliothèques et archives à l'ère des Humanités numériques (CIBAHN), 2016.

Tărâță, Cristina. Data quality dimensions – from Accuracy to Uniqueness. 3 March 2015. http://www.performancemagazine.org/data-quality-dimensions-from-accuracy-to-uniqueness/ (19th May 2017)

Tièche, Julien. La mesure des dimensions de la qualité des archives électroniques : apport des textes normatifs en matière d'archivage électronique à long terme. Genève : Haute Ecole de Gestion, travail de Bachelor, 2015. http://doc.rero.ch/record/258018/files/TDB_Tieche_Julien.pdf (19th May 2017)

Valais – Service de la Culture. Vallesiana, le patrimoine numérique du valais. 2017. http://www.vallesiana.ch/ (19th May 2017)

Wroblewski, L. (2017). Culture Management: Strategy and marketing aspects. Logos Verlag Berlin GmbH.

Annex 1: Information Quality Levels

Information Quality Levels					
Score / Criteria	*	**	***	****	*****
General information qualities applied on sources					
Trustworthiness					
Reliability	Provided by 1 non-institutional information source Non-complete description	Provided by 1 institutional information source Non-complete description	Provided by 1 or more institutional information source Non-complete description	Provided by 2 institutional and citizen's testimony information source	Provided by institution and professional testimony sources with exhaustive description
Authenticity	Provided by source non-identified	Provided by Citizen's authentic testimony	Provided by Authentic professional testimony	Provided by multi-source information	Provided by authentic multi-source information
Exploitability					
Cognitive Accessibility	French	French/ German/Italian	French/ German/Italian/ English/Spanish	Multi-language (7 languages)	Multi-language (more than 10 languages)
Technical Accessibility	Easy to understand for the French tourist or citizen	Easy to understand for the French/ German/ Italian tourist or citizen	Easy to understand for the tourist or citizen lambda	All tourists and citizens can understand easily	All tourist and citizen can understand easily
Juridical Accessibility	Information is accessible but non-consultable physically	Information is accessible but non-consultable physically	Information is accessible but non-consultable physically	Information is Open access and can be consulted with a request to the institution with an identifier (download if check-in as user)	Information is Open access and can be consulted with a request to the institution with an identifier (download if check-in as user)
Representativeness					
Institutional Context	Recommended by Office of Tourism but information not up to date (of last year)	Recommended by Office of Tourism but information not up to date (of last year)	Recommended by Office of Tourism but information not up to date (of last year)	Recommended by Office of Tourism and information is up to date	Recommended by Office of Tourism and information is up to date

Socio-cultural Context	Information is not scarce, in a context summary (abstract)	Information is not scarce, in a context summary (abstract)	Information is not scarce, in a context summary (abstract)	Information scarce by the exhaustiveness of the context	Information scarce by the exhaustiveness of the context
Specific information qualities applied on contents					
Uniqueness	- Density of information, lacks of precision - Abundance of format, media, information typology, subject, period/context	- Density of information, lacks of precision - Abundance of format, media, information typology, subject, period/context	- Selection of information, lacks of precision - A selection of format, media, information typology, subject, period / context	- Exclusive Information, precise and indexed - Scarcity of format, media information typology subject, period/context.	- Exclusive Information, precise and indexed - Scarcity of format, media information typology subject, period/context.

Potentials of Digital Archives: Topotheque of Smart Novel Vilijun – Case Study

Vlatka Lemić
Croatian State Archives
Marulićev trg 21, Zagreb, Hrvatska
vlemic@arhiv.hr

Josipa Mijoč
Faculty of Economics, University of Osijek
Trg Ljudevita Gaja 7, Osijek, Croatia
jmijoc@efos.hr

Nikolina Filipović
Faculty of Economics, University of Osijek
Trg Ljudevita Gaja 7, Osijek, Croatia
nfilipovic@efos.hr

Summary

Observed from the "memory economy" perspective, digital archives of cultural heritage, besides the reliable memory function, have economic role in mapping local and global "cultural geography" as well. The mentioned economic role of digital archives is manifested in the long-term promotion of cultural heritage, thus turning archived content into the promoter of the cultural property itself, a place where it is located or with which the cultural property is connected with its origin, while a local community has economical, socials or historical links with this cultural property. The visibility of digital archives is closely related to the platform where the archival material is presented. This paper analyses archive collection of novel Vilijun as an example of open digital archives dedicated to one contemporary novel with heritage content. Topotheque digital platform uses interactive IT tools, it is based on a collaboration and engagement of heritage professionals, users and visitors, and descriptions of presented material go beyond the standardized rules of archival description. Topotheque Vilijun is private collection related to interactive novel Vilijun whose content promotes heritage. Promotion of the novel Vilijun is a new form of "memory economy" whose visibility, and thus promotional reach, transcends authors 'and publishers' activities. The presented content (the book - product of the publishing industry) through the reach and visibility of the Topotheque platform becomes a product of the cultural and creative industry whose promotion is happening and documenting on the long-term. It allows for long-time preservation and access to digital content and ensures it's promotion and is thus a "real" ar-

chive of the information society of the 21st century, mapping the spaces of "cultural geography".

Key words: cultural geography, digital archives, Topotheque, heritage, economy, novel *Vilijun*

Introduction

Memory is at the subject of interest of various sciences, ranging from humanities (history, art, philosophy), social sciences (sociology, anthropology, economy, information and communication), nature (medicine, chemistry, biology) till technical sciences (computer, robotics). Digital archives[1] are connected with memory and technology and many global processes, as such centre of interest of archives, information managers, heritage professions and many other disciplines. One of comprehensive overview of relevant authors and topics regarding interconnections of archival science archives principles, ICT and digital era is given in Eric Ketelaar article *Archives, memories and identities* where authors concludes:

> "Nevertheless archive(s) "as it is" have a unique quality and it is the archivist's calling to advocate that uniqueness benefiting many if not most processes of "meaning making" leading to identification and categorization; self-understanding and social location; commonality, connectedness, groupness. These identities are rooted in memories and these memories need inscription and need a space. Both inscription and space will increasingly be "located" "in the cloud" and maintained (in distributed custody) by individuals, groups, and memory institutions. Together they are actors in an ecology which comprises archives/records and other memory texts in a societal context…"[2]

The memory of cultural heritage in the digital age is realized twice: on the Internet itself and in cultural monuments whose presentation is (not) realized on the Internet. Key role for this lays in archives, libraries and museums as memory institutions: they organise the [...] cultural and intellectual record. Their collections contain the memory of peoples, communities, institutions and individuals, the scientific and cultural heritage, and the products throughout time of our imagination, craft and learning. They join us to our ancestors and are our legacy to future generations.[3]

[1] Digital archives, digital libraries and digital collections are on the Internet often differed only by name in terms of digital repositories.

[2] Ketelaar, Eric. Archives, memories and identities (…), pp. 69

[3] Dempsey, Lorcan. Scientific, Industrial, and Cultural Heritage: A Shared Approach // *Ariadne* 22 (Tuesday, 21 December 1999). http://www.ariadne.ac.uk/issue22/dempsey

The emergence of transnational public digital archival platforms (like Monasterium, Mapire, Topotheque) and digital archives such is Archives Portal Europe transcends some of the major controversies regarding trust in digitized memory, their content, and their source. At the same time, public digital archives prevail the network entropy where democracies in the advertising of data/information are crucial to phenomena such are: the difficult finding of relevant content, insufficient verifiability of relevant content and insufficient maintenance of advertised content. This paper will consider the assumptions and opportunities that arise from the presenting of cultural content in the public digital archive as well as its cost-effectiveness in the promotion of heritage and cultural memory.

Archives in digital age

Contemporary information society has influenced archives towards outreach, enhancing public knowledge on archival sources and encourages easy access to archives on the international level, while archival programs are connected with information society development and cultural heritage policy in general. Great number of projects under the "culture and history" framework are focused on programs and activities related to digital heritage, democratization of access to cultural heritage, social inclusion, information use and re-use, cultural industries and similar topics, including digital platforms, cultural networks and e-services.

The best picture of 21st century archives provides the Universal Declaration on Archives made by International Council on Archives, stating importance and necessity of archives, their diversity in recording every area of human activity, multiplicity of formats in which archives are created, the role of archivists in serving their societies, as well as collective responsibility of all society members in management of archives[4]. Declaration recognizes archives as unique and authentic whiteness of administrative, cultural and intellectual activities and as reflection of society evolution. As such, they are of vital importance for supporting business efficiency, accountability and transparency, for protecting citizens' rights, for establishing individual and collective memory, for understanding the past, and for documenting the present to guide future actions.

Contemporary archives are expected to be a public administration service regarding document management and protection and also to be providers of new services which would ensure better availability, visibility and presentation of archives and archival sources in the public by using new technologies. The international archival community through its documents and activities also emphasizes interaction and cooperation between archives, public administration and other professions and various public and private sectors. Accordingly, the activities of European archival community for decades are focused on co-operation and networking of archival institutions at all levels, development and im-

[4] ICA Universal declaration on Archives. http://www.ica.org/sites/default//files/UDA_ June%20 2012_web_EN.pdf

plementation of professional standards, transfer of knowledge and creation of a common information infrastructure.[5]

Cultural memory and memory economy

The global era is characterized by information overwhelming and hence entropy as a result of semantic and informational controversy in finding content that is presenting online. Already Escarpit notes that the book as a product of the cultural and creative industries intended for mass selling differs books whose life on the book market primarily appears as a category of "short-term" creative product (best-sellers) to continue under certain circumstances its course towards a product that on the market lives as a "long-term book."[6]

One of the prerequisites for a book, and its content, to live "in the long run" is the public visibility of the book – product of the creative industry. In terms of entropy of data published on the Internet, public digital archives can also contribute to the promotion of creative products in the long term, and thus the longevity of content that a given product represents. Although the purpose of the public digital archives is not primarily aimed at promoting the creative industries, their role can also be observed from this perspective, aligned with re-use of public sector information directions and outreach initiatives. Namely, the public archiving of the life span of a book that deals with heritage content, apart from publicly promoting publishing, at the same time builds up a map of cultural links that literary texts achieve – whether it is the representations of the literary text itself as part of the publishing promotion activities, or introducing heritage content described in the text. This opens the possibility to consider the public digital archives as a platform where "archival certified memory" contribute to the economy and social benefit of mapped areas, themes and archive material. Memory economy thus becomes a platform for long-term promotion of (cultural and/or creative) products, but for the first time in the history of literary engagement, it is possible to build a long-term memory at the time of its creation.

Topotheque digital platform

Topotheque is a digital platform – a collaborative online archives – providing public and free access to digitized historic sources from various community public and private collections. It is created by ICARUS[7] in the framework of EU founded project co:op – "Community as Opportunity – the Creative Users' and Archives' Network" as a new opportunity of safeguarding and presenting less known, marginalised, and often not easily accessible historic documents.

[5] Lemić, V. Archives and society – what archives are, can and should be – Croatia case study (….), pp.128

[6] Escarpit, Robert. The Book Revolution. London : Harrap, UNESCO, 1966, pp. 147

[7] ICARUS - International Centre for Archival Research. http://icar-us.eu/en/

International project co:op is financed through the Creative Europe program and it brings together 17 archival and academic institutions with more than 40 associated partners from all around Europe aiming at strengthening transnational cooperation between institutions and user groups.[8] Following former project ENArC – "European Network on Archival Cooperation", co:op is going wider and deeper in strengthening and promoting the cooperation between archives and other institutions preserving our common cultural heritage, as well as, encouraging the active involvement of the general public. A variety of creative, pedagogical and didactic activities planned inside a four year schedule (including Topoteque, "Adventure in the archives" and "Bring your history days" programs, educational material for schools, historical workshops, scientific research etc.) are dedicated to the promotion of archival activities to the wider community, to fostering collaboration between the public and archives and to facilitating access to archival material by using the possibilities of the digital age.

Topotheque digital platform[9] provides description, presentation and search of archival material by using interactive IT tools and description scheme compliant with ISAD (G) standard which enables data transfer in other archival information systems. The administrational work within every Topotheque collection done by a registered topothequers, while visitors and users can also be engaged through answering questions online and, as guest-topothequers, uploading and indexing data (crowd work).

During the last two years more than 120 Topotheque collections all across Europe were published online and they helped visibility of its local communities on regional and national level, encouraged local programs and events (history and memory days) and helped the promotion of cultural and other manifestations and history specifics of local areas. Through them one can meet private family documents and photographs (like Bischoff family), monitor the changes in life and landscape of some small places (all around Europe) or famous sights (like Viennese Prater). All material and data on common Topotheque platform are delivering further to Europeana, thus building individual and local stories in shared European history[10].

Smart novel Vilijun Topotheque

Smart novel Vilijun Topotheque was open to the public on 20 June 2017. It is private collection made by author of novel, consisting of various materials (archival records) connected with the novel: parts of the original text, illustrations in the novel, recordings made on book promotions, photos made on novels

[8] Project partners list is available at: https://coop.hypotheses.org/category/project-partners.

[9] Topotheque is available at http://www.topotheque.eu/.

[10] Lemić, V. Mogućnosti suradnje arhiva i zajednice – co:op projekt // *Glasnik arhiva i Arhivističkog udruženja BiH.* 46 (2016), pp. 107-109

presentations, media texts, literary theoretical reception, guest presentations at book fairs (Peking, Zagreb), footage of rehearsals the premiere performances of the novel *Vilijun,* recordings of the premiere performances of the novel *Vilijun* and interviews with the author.

Vilijun is novel by author Jasna Horvat published in 2016 by Ljevak publishing company, labelled from critics as the first QR i.e. "smart" novel whose reading requires the use of a smartphone. In the annotation of the novel is the following description:

> "The protagonists of Vilijun are Marco Polo and Kublai Khan in the year of their farewell. Marko Polo tells Kublai Khan about the cities on the Silk Road, and Khan is interested in Marko Polo stories to decide whether to allow him to return to his homeland. It is bond of two nomads and two cultures within which Marko Polo also describes numerous other cultures he met and got to know on the Silk Road. It is a novel about nomadism – thought and traveller, but also about trust, friendship and loyalty."[11]

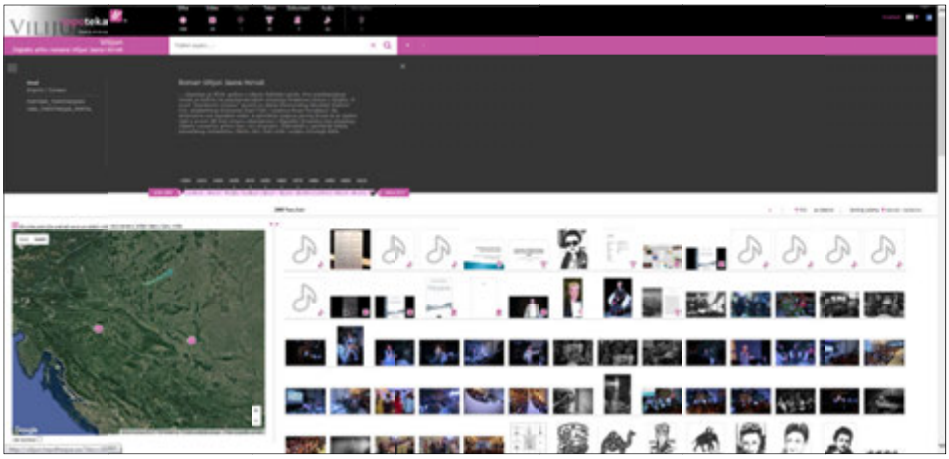


Figure 1. Topothetheque Vilijun (http://vilijun.topotheque.eu/)

Topotheque Vilijun is a new step in the interactivity of this text, which allows everyone paper and web travel by the Silk Road stations – it is a digital archive of one novel, who is also a holder of heritage memories. Topotheque Vilijun in many ways is unique archive – it keeps various records (presentations, promotions, newspaper articles, theatrical plays, performances at book fairs, thematic talks, published cultural and literary criticisms, scientific papers) which document the life of this novel; it consists of all types of media (documents, photos,

[11] Knjižara Ljevak, http://www.ljevak.hr/knjige/knjiga-20935

AV records etc) and it encompasses existing, as well as yet non created documents that we do not know when will be make, how they will look like and what kind of ideas will trigger them. As live, timeless public archives that links the story and the characters, readers, scientists, artists, users and all those whose paths are connected with them at a particular moment –through the theme and life of the novel *Vilijun* – Topotheque Vilijun is an archive of the information global society of the 21st century and at the same time a new form of promotion of heritage themes and the very novel itself as a product of the creative industry.

Multimedia novel

Novel *Vilijun* is an example of contemporary literary text that links the media of printed books with the Internet – a global digital media. It is the first published QR novel in Croatia and also the first novel that presents heritage themes in an innovative way. In her Afterwords in the print novel, Dubravka Oraić-Tolić states the following:

> " … A new shift in the work of Jasna Horvat occurred in the book *Vilijun* (2016). It is a multifaceted novel-toy. This is, on the one hand, conscious, planned, organized and thematic re-conceived permutation of the *Vilikon* novel. On the other hand, it is a "smart book" (author's self-concept) that is symbolically and truly linked to new technologies, mobile phones and the world of the Internet. In the first, textual layer of a novel, the author is playing with her own novel *Vilikon* and his reconceptualization. In the second, para-textual layer (QR-codes scattered in the novel) she offers the reader an opportunity for endless games and thus creates an interactive hypertext – a book toy."[12]

In her critical review of the novel, Oraić-Tolić notes that this innovative way of literary expression give the possibilities of multiple readings, and that Vilijun is "first Croatian interactive hypertext novel with a million of possibilities of textual, visual and network nomadism" and illustrates this with two QR codes from *Vilijun* which are results of the author's work, i.e. the first of them came about one year before the publication of the novel, and the second came about at the first presentation of the novel *Vilijun*.

Both QR codes perform (musical and acting) poem *Million*, which is part of the novel *Vilijun*. In addition, both contents encrypted with these QR codes are in some way archival documents because they are part of the author's private collection and are directly related to the creation and the presentation of the novel of which they are an integral part. In this way, novel *Vilijun* can be seen as an

[12] Oraić-Tolić, Dubravka. Ars Horvatiana. U: Horvat, Jasna. Vilijun. Zagreb: Naklada Ljevak, 2016, pp. 211.

own archive collection, thus opening up the question of the documentary capacity of the novel itself.[13]

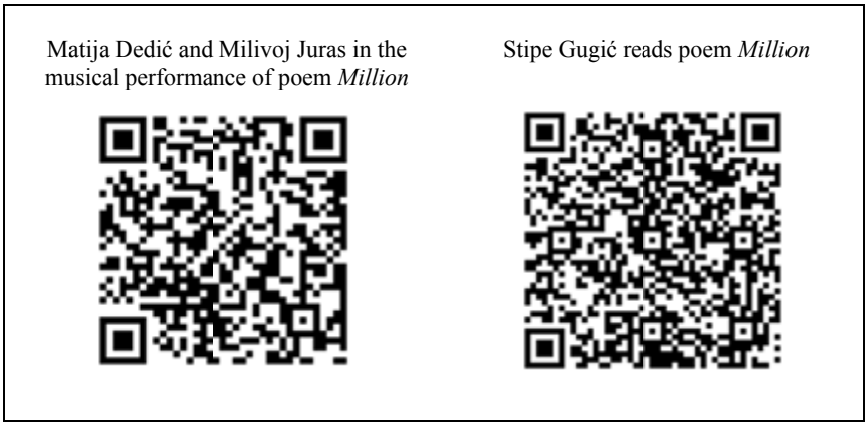


Figure 2. QR codes from *Vilijun* novel

Novel - digital archive of heritage themes

If we accept the possibility that contemporary literary text with the mediation of QR codes has capacity of digital archiving of its own content, the question is whether such a record is able to promote topics whose information is relevant as a basis for non-fictional considerations. Since in *Vilijun* novel there is a "non-fictive, lexical part" that "fully functions in accordance with the principles of lexicon as a lexicographic type of text"[14], Table 1 lists the lexicographic sections that the novel brings forth.

As it can be seen from the allegation of the lemma shown in Table 1„ the novel *Vilijun* offers readers information about cities on the Silk Road, cities of importance to the life of Marco Polo (Field 2), the land and sea route of Silk Road, selected cities on the Silk Road, the products traded in the 13th century (Field 6)and about the four symbols of the identity of Croatian culture (Field 4). The lexicographic approach to the formulation of literary text and the additional presentation of documents with the help of QR codes allows concluding that the *Vilijun* printed novel implied the digital archiving of heritage themes.

[13] Ibid, pp. 217

[14] Kos-Lajtman, Andrijana. Poetika oblika. Zagreb: Naklada Ljevak, 2016, pp. 199.

Table 1: Novel *Vilijun* lemmas

First row of magic square of number 12	Second row of magic square of number 12	Third row of magic square of number 12
FIELD 3	**FIELD 8**	**FIELD 1**
(1) Korčula (2) Šibenik – City of Krešimir (3) Venetia	(1) Jerusalem (2) Mosul (3) Bagdad (4) Samarkand (5) Baktra (6) Kashgar (7) Lanzhou (8) Karakorum	(1) The song about the names of cities on the Silk Road
FIELD 2	**FIELD 4**	**FIELD 6**
(1) About land route of the Silk Road – the way to the East (2) About sea route of the Silk Road – the way to the West	(2) Croatian coat of arm (Kockovlje) (3) Early-Croatian three-strand patern (Troplet) (4) Name Croat (Hrvat) (5) Glagolitic script (Glagoljica)	(1) Tea and spices (2) Silk (3) Porcelain (4) Cashmere (5) Paper (6) Compass
FIELD 7	**FIELD 0**	**FIELD 5**
(1) Bagan (2) Chengdu (3) Camblau (4) Hormuz (5) Arbil (6) Trabzon (7) Carigrad	About not mentioning Marko's return to the Kingdom of Croatia.	(1) Golden plate (2) Salt (3) Fairies notifications (4) World map (5) Million

Double archiving in Topotheque Vilijun

Archival activities, mostly classified as part of the cultural and creative industries sector and often encountered within the GLAM acronym, were greatly influenced by the digital age, like other cultural and creative industries. Thus, it becomes part of an "open society" and affects the understanding and use of terms such as "access" and "re-use". Ideas of openness, networking and integration in building of common information Internet infrastructure are part of many EU strategies, reports, summaries, programming documents and initiatives, like in the following words: "Especially the aspects 'access' and 'reuse' of digital resources are strongly connected with the sustainability of the digital resources, because if the digital resources are not preserved, this naturally means an end to all access and reuse. This problem begins with the well-known phenomenon of a broken link if a website is no longer maintained and no more of use to verify information. The common denominator 'digital sustainability' (also long-term preservation or digital curation) describes a span of activities that more or less encompass the whole research (data) lifecycle and exceeds the narrow sense of archiving in general linguistic usage. The term archiving means to archives, museums, and libraries more than permanent storage on a medium, it encom-

passes the notion of ensuring long-term access and therefore includes the need to preserve modes of reuse and retaining the interpretability of the digital resources[15]. This is a collective task, which includes many stakeholders, from researchers to digital preservation specialists."[16]

Apart from the above, it can be seen that the contents presented in public digital archives (such is Topotheque) also have other advantages, including public visibility of the presented content and overcoming the problem of semantic web. Considering the basic role of Topotetheque in linking the places of creation of archival records with the main theme of Topotheque collection, it is possible to conclude that Topotheque realizes the mapping of "cultural geography" and that the "archive map", along with the effects of long-term memory, also promotes archived content on long time.

According to the example of Topotheque Vilijun, it is noticed that double archiving of heritage content was achieved: a) primarily in the creative industry product itself (novel Vilijun) and based on the author's research of historical and cultural sources, and then b) in the digital archive of Topotetheque Vilijun by advertising scientific and professional studies of the novel Vilijun as well as documents that certify the market and social life of this text.

Conclusion

Topotheque Vilijun is real example how archives can "come out of a box". It is open and borderless in all senses – by its scope (type and quantity of records), content, use and opportunities. It also shows how digital archives can actively link heritage, education, community, creative industries and other potentials, be resource and inspiration for creation of new information and cultural products and services, speed up and facilitate sharing, gathering, presentation, research, publishing and documenting sources and making it accessible to the whole world. This is what archives should be – link between past and future.

[15] See: Neuroth et al. (nestor Handbuch), 2010, Kap. (title) 1:3

[16] Wuttke et al. (…), pp. 22

References

Dempsey, Lorcan. Scientific, Industrial, and Cultural Heritage: A Shared Approach // Ariadne 22 (Tuesday, 21 December 1999). http://www.ariadne.ac.uk/issue22/dempsey

Escarpit, Robert. The Book Revolution. London : Harrap, UNESCO, 1966.

Horvat, Jasna. Vilijun. Zagreb : Naklada Ljevak, 2016

H. Neuroth et al., Hrsg. 2010. NESTOR-Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung. (Version 2.3.) Göttingen. Kap.4:13, Kap.4:15- Kap.4:16. Online: http://nbn-resolving.de/urn/ resolver.pl?urn:nbn:de:0008-2010071949

ICA Universal declaration on Archives. http://www.ica.org/sites/default//files/UDA_ June%20 2012_web_EN.pdf

ICARUS - International Centre for Archival Research. http://icar-us.eu/en/

Ketelaar, Eric. Archives, memories and identities // R*ecords, archives and memory : selected papers from the Conference and School on records, archives and memory studies, University of Zadar, Croatia, May 2013* / Willer, M.; Gilliland, A.J.; Tomić, M. (ed.). Zadar : Sveučilište u Zadru, 2015, str 47-76

Kos-Lajtman, Andrijana. Poetika oblika. Zagreb : Naklada Ljevak, 2016

Knjižara Ljevak. http://www.ljevak.hr/knjige/knjiga-20935

Leathem, Camilla; Adrian, Dominik,. Survey and Analysis of Basic Social Science and Humanities Research at the Science Academies and Related Research Organisations of Europe. Berlin: Union of the German Academies of Sciences and Humanities, 2015. http://www.akademienunion.de/fileadmin/redaktion/user_upload/Publikationen/Proj ectReport_SASSH_2015.pdf.

Lemić, V. Mogućnosti suradnje arhiva i zajednice – co:op projekt // *Glasnik arhiva i Arhivističkog udruženja BiH.* 46 (2016), str. 99-110

Lemić, V. Archives and society – what archives are, can and should be – Croatia case study // *Proceeding book with peer review / Symposium Archives in the Service of People – People in the Service of archives in conjuction with 5th International Scientific Conference All About People: Interdisciplinarity, Transnationality and Building Bridges, Maribor, 10.-11.3.2017.* / Filej, B.; Klasinc, P.P. (ur.). Maribor: Alma Mater Europea – ECM, 2017, str. 128-136

Oraić-Tolić, Dubravka. Ars Horvatiana. U: Horvat, Jasna. Vilijun. Zagreb: Naklada Ljevak, 2016, str. 206-223

Topotheque. http://www.topotheque.eu/

Topothetheque Vilijun. http://vilijun.topotheque.eu/

Wuttke, Ulrike; Ott, Carolin; Adrian, Dominik; Worthington, Simon. AGATE: Concept for a European Academies Internet Gateway for the Humanities and Social Sciences, 2017, http://doi.org/10.5281/zenodo.815916

GOVERNMENTAL AND BUSINESS SECTOR INFORMATICS

Original scientific paper

Confronting Internet Security Threats

Radovan Vrana
Department of Information Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb,
Ivana Lučića 3, Zagreb, Croatia
rvrana@ffzg.hr

Summary

The paper presents the results from the research study of students at the Faculty of Humanities and Social Sciences, University of Zagreb, Croatia about their awareness and exposure to well-known internet security threats as well as their awareness about security threats countermeasures. The results of the research study indicate high level of awareness of students about internet security threats as well as the fact that they are exposed to widely internet security known threats to a certain degree. While students demonstrated knowledge about application of security threats countermeasures, that part of their activities could be improved further. Generally, students need additional and updated knowledge to raise the level of their readiness in order to be able to respond to known and emerging internet security threats. The results of the research will be applied in planning of university courses related to the internet security.

Key words: internet security, security threats, students

Introduction

In today's world, care for computer and other networked devices security has become almost as equally important as the development of information systems themselves. Recent cyber-attacks showed the importance of intrusion prevention by monitoring vulnerabilities and reducing security threats (Abazari, Madani and Gharaee, 2016). Vulnerabilities of computer and other networked devices (like smartphones and other smart devices) consist of "weaknesses in a system which can be exploited by the attackers that may lead to dangerous impact" (Jouini, Rabai and Aissa, 2014, 490). At this moment, there are many active security threats related to the use of the internet (Sherr, 2017; Burgess, 2017; Schroeder, 2017) that could exploit weaknesses in computer and other networked systems and cause substantial financial and other damages. To prevent realization of threats, an individual must be well informed about potential weaknesses in the operating system he or she uses on a computer or other device as well as about weaknesses in applications used in one's daily work. One such group of users of ICT are students which are highly active users of computers, smartphones, tablets and networking. As such they could easily become

victims of different security threats which could end in financial, intellectual, academic, reputation and other damages. To investigate the current level of awareness of students about well-known internet security threats and their exposure to these threats as well as their awareness about security threats countermeasures, a research study was initiated. This paper will present results from this research study.

Security threats

A (computer related) threat is "action or potential occurrence (whether or not malicious) to breach the security of the system by exploiting its known or unknown vulnerabilities. It may be caused by (1) gaining unauthorized access to stored information, (2) denial of service to the authorized users, or (3) introduction of false information to mislead the users or to cause incorrect system behavior (called spoofing)" (Threat). According to Technopedia, "threats are potentials for vulnerabilities to turn into attacks on computer systems, networks, and more". In addition to computer related threats, a category of their own are internet related threats. According to Symantec, 11 most common security threats include virus, spam, spoofing, phishing and farming, spyware, keylogging, adware, botnet, worm, Trojan horse, blended threat (combination of several threats at once), denial of service attack (DOS). In addition to viruses, Vernon included hacks into the list of threats, and hacks have become very frequent in recent periods of time. While some threats aim at a single system vulnerability, other involve multiple exploits (Technopedia) and target both businesses and individuals. The oppose threats, individuals must protect information with the prevention and detection of unauthorized actions by users of a computer (Microsoft). To reach the adequate level of security, one must also apply adequate countermeasures or protective measures. Online literature on internet threats and protective measures is abundant and will be limited to a small selection of references due to the space restrictions: Singh, Kumar, Singla and Ketti (2017) wrote about internet attacks and intrusion detection system; Byrne, Dvorak, Peters, Ray, Howe and Sanchez (2016) investigated user's perspective on risks relative to benefit associated with using the internet; Berriman (2017) wrote about youth using the internet and the governance of their use of the internet; Google introduced a program for teaching safe online exploration (2017); van den Berg and Keymolen wrote about the problem of internet regulation focusing on control vs trust issue; journal Education journal published an article on the internet safety measures (2017). The internet users can inform themselves about the latest internet threats by using many available information resources as the internet security is a topic of high interest to the widest possible circle of users of the internet. In addition to informing oneself about the internet threats and protective measures by using online courses and written, video and audio materials available on the internet, students at the Faculty of Humanities and Social Sciences at the University of Zagreb (who are in focus of this paper)

have a possibility of acquisition of knowledge about internet security. As part of the study program they can take part in courses related to use of ICT such as "Communication technology fundamentals", "Data protection", "Cryptology" and "Internet culture". In addition to the formal study program courses, students can choose short ICT training on site or online courses like "IT security" at the University computer centre in Zagreb, Croatia. The courses are intended primarily for students and teaching staff at the academic institutions in Croatia.

Research methodology

This research is a follow-up of previous researches on students' perceptions about the internet security threats (Vrana, 2012) and online social networks and security of their users (Vrana, 2013). To find out details about the current awareness and exposure of students to most common and active security threats and their awareness about security threats countermeasures, a research study was initiated. The purpose of this research study was to detect the level of awareness and exposure of students to security threats in order to offer them additional education about security threats countermeasures. The research study aims to answer the following research questions: RQ1: Are students exposed to internet security threats?; RQ2: Are students able to recognize the most common and most frequent internet security threats?; RQ3: Are students applying the basic security threats countermeasures? Online questionnaire with 16 closed type questions was chosen as the research method. While having some drawbacks as a research method, questionnaire is still valuable and applicable research method for researching a large number of potential respondents. The invitation for participation in the research study was sent by students' mailing list at the Faculty of Humanities and Social Sciences in Zagreb (University of Zagreb) in Croatia and it was also published on the main Web page of the same Faculty. The participation invitations were sent on May 9th 2017 with the closing date of May 17th 2017. Convenience sample was used a sampling method to attract as much students as possible for participation in the research study. The research study was closed on May 17th 2017 with 152 answer sets collected.

Research findings

Due to the space restriction, partial research study results will be presented in the following part of the paper.

The following part of results answers two research question; RQ1: Are students exposed to internet security threats? And RQ2: Are students able to recognize the most common and most frequent internet security threats?

Internet security threats recognized and encountered by respondents

Ability to recognize security threats represents an important step towards more secure use of the internet and avoiding well-known and well-documented threats. In this question, the respondents were given an opportunity to select (by

media) frequently announced security threats. The results indicate that the respondents are well acquainted with the most widely known threats except for more sophisticated and more complex threats like farming and spoofing. It is also surprising to see social engineering so low on the list of recognized threats since it quite a common threat. In the second part, the respondents were given the same list of security threats as in the previous question and a possibility to choose the threats they actually encountered. The results show that the majority of threats are spam and classic security threats like viruses, Trojan horses and worms. Adware is also highly ranked as it is present in many free (non-fee based) software applications. On the positive side, it is satisfactory to see phishing and identity theft ranked so low as consequences of their activities could be very severe.

Table 1. Internet security threats recognized by the respondents (multiple answers) (N=151) and internet security threats encountered by respondents (multiple answers) (N=149)

	Threats recognized by respondents		Threats encountered by respondents	
	N	%	N	%
Spam	145	96.0	121	81.2
Virus	143	94.7	105	70.5
Identity theft	135	89.4	101	67.8
Trojan horse	132	87.4	74	49.7
Adware	128	84.8	42	28.2
Spyware	123	81.5	27	18.1
Worm	111	73.5	21	14.1
Phishing	84	55.6	11	7.4
DDOS	43	28.5	8	5.4
Man in the middle	28	18.5	5	3.4
Social engineering	27	17.9	5	3.4
Farming	25	16.6	4	2.7
Spoofing	21	13.9	4	2.7
None of the above	1	0.7	1	0.7

Personal data used when signing / logging in into internet services

The choice of login data is usually predetermined by the internet service owner(s) and cannot be chosen / selected by internet users. Personal data protection is increasingly becoming topic of interest as many new online services require entering personal data. While some of the personal data are less secure either because of their shortness (PIN) or because they can sometimes be guessed based on available information about a particular internet user (login or e-mail address), other methods like user's picture are less frequently used (for instance, on smartphones in the process of face recognition) and can be falsified.

Table 2. Personal data used when signing / logging in into internet services (multiple answers) (N=151)

	N	%
E-mail address	139	92.1
Login name consisting of your first and last name	113	74.8
First name	77	51.0
Last name	72	47.7
PIN	56	37.1
Mobile phone number	33	21.9
Your picture	30	19.9
Personal identification number	11	7.3
Some other data	4	2.6

Unlock procedure used when accessing mobile phone

The aim of this question was to detect most commonly used mobile phone unlock procedure as a part of the phone access security. The unlock procedure is also a possible point of attack and must be taken into account when researching the problem of user's secure use of the internet. While PIN remains most popular phone unlock procedure, it is worth noting that not so insignificant number of users do not use any unlock procedure leaving their mobile phone openly accessible to anyone who can get into possession of the phone. The availability of some unlock procedures is directly related with the hardware installed in the mobile phone (for instance, fingerprint scanner), so, they are not available to all respondents.

Table 3. Unlock procedure used when accessing mobile phone (N=151)

	N	%
PIN	42	27.8
None of the above	38	25.2
Screen pattern	36	23.8
Fingerprint	21	13.9
User name or password	14	9.3
Retina scan	0	0.0

Frequency of following URLs sent in e-mail messages

The most recent phishing campaign that happened to Google in May 2017 (Levin, 2017) demonstrated clearly how important is for internet users to recognize valid from invalid URLs in their e-mail sent by hackers. The respondents who always follow URLs are also prone to phishing or virus infections more frequently than those respondents who do not follow URLs in their e-mails.

Table 4. Frequency of following URLs sent in e-mail messages (N=151)

	N	%
Seldom	70	46.4
Never	42	27.8
Occasionally	30	19.9
Often	8	5.3
Always	1	0.7

The next part of the research answers the following research question: RQ3: Are students applying the basic security threats countermeasures?

Informing oneself about internet security threats and informing oneself about internet security threats countermeasures

Informing oneself about the most recent and most dangerous internet security threats is a priority in establishing the behavior patent that helps in secure use of the internet. The aim of this questions was to discover sources of information the respondents use to inform themselves about the internet security threats. The most frequently chosen information sources are those at hand: friends and the university. It is interesting to see that some respondents use some other ways to inform themselves, however, some of them provided answers in which they state that they do not inform themselves at all. Similarly, to the first question, the question about and informing oneself about internet security threats countermeasures aimed at discovering sources of information for the respondents about internet security threats countermeasures. Except the most frequently chosen answer (friends), other answers differ from the previous question showing that the respondents seek information about countermeasures from professional sources which indicates that they are aware of existence of such sources and their potential quality when dealing with security threats.

Table 5. Informing oneself about internet security threats (N=150) and informing oneself about internet security threats countermeasures (N=149)

	Informing oneself about internet security threats		**Informing oneself about internet security threats countermeasures**	
	N	**%**	**N**	**%**
At the university	47	31.3	36	24.2
Internet security companies Web sites	30	20.0	45	30.2
Courses outside the university	4	2.7	2	1.3
Daily newspapers	16	10.7	11	7.4
Friends	81	54.0	80	53.7
General purpose news Web portals	38	25.3	28	18.8
Other ways of informing	63	42.0	47	31.5
Popular computer and internet related magazines	29	19.3	26	17.4
Radio	12	8.0	2	1.3
Relatives	30	20.0	30	20.1
Safe internet use specialized Web portals	45	30.0	62	41.6
TV	31	20.7	13	8.7
Weekly magazines	2	1.3	1	0.7

Solving internet security threats

In addition to being informed, the respondents have to also be able act for themselves in order to resolve the security threat issue. A significant number of them help themselves while other seek help from friends, experts and relatives. Dealing with the security threats by themselves indicate that these respondents are confident in their own knowledge and skills.

Table 6. Solving internet security threats (N=150)

	N	%
By myself	60	40.0
By the help of friends	33	22.0
By the help of experts	27	18.0
By the help of relatives	25	16.7
Other ways of solving threats	5	3.3

Antivirus software installed

Today, when there are free of charge antivirus software applications available globally, there is no excuse why one wouldn't have such a software installed on the device intended for access to the internet. The results in this question (N=151) indicate that 90,1% (N=136) have antivirus software installed on their device while 9,9% (N=15) of the respondents still do not have such a protection which could lead to security problems.

Frequency of operating system (OS) and application update on a device most frequently used for access to the internet; frequency of creating a backup copy of content on a device most frequently used for access to the internet

Software update is one of the most common and straightforward defense methods against security threats. Most recent OS-es and application offer automatic check-up for availability of updates and their installation thus helping users to avoid security holes in OS and applications they frequently use. With every new update, new safety features are added as it was evidently necessary in case of the most recent global ransomware attack (Hern, 2017.). The results (N=151) show that OS is updated occasionally and applications (N=151) often which is good as it raises the level of security. A more frequent application of updates would improve the security even more. Finally, creating a backup copy (N=151) of user data and applications is another critical activity in achieving the necessary level of security of computer and other networked systems. Unfortunately, the respondents are creating backup copies only occasionally and seldom which put them in danger of data and applications loss.

Table 7. Frequency of operating system (OS) and application update and frequency of creating a backup copy of user data (N=151)

	Operating system update		Applications update		Creating backup	
	N	%	N	%	N	%
Never	4	2.6	4	2.6	34	22.5
Seldom	22	14.6	15	9.9	43	28.5
Occasionally	53	35.1	38	25.2	46	30.5
Often	33	21.9	53	35.1	15	9.9
Always	39	25.8	41	27.2	13	8.6

Frequency of change of passwords in internet services one uses

Frequent password change is one of the best methods of protections against user account intrusion. The consequences of not doing so could lead to sever consequences as showed by the resent publication of a database with 560 million of user passwords on the internet (Broida, 2017). The answers to this questions indicated poor management of user accounts protected by passwords as almost one fifth of the respondents do not change passwords at all, while almost half of them do it less than once a year. Very few respondents change their passwords once in three months or even more frequently, which should be the standard procedure.

Table 8. Frequency of change of passwords in internet services one uses (N=151)

	N	%
Never	29	19.2
Less than once a year	64	42.4
Once a year	26	17.2
Once in 6 months	23	15.2
Once in 3 months	5	3.3
Once a month	3	2.0
Once a week	0	0.0
Daily	1	0.7

Estimation of students' knowledge about internet security

The final question aimed at receiving estimation of the respondents' knowledge about internet security in general. Almost half of the respondents showed that their knowledge is insufficient or sufficient which should be immediately improved given the situation with the severe security incidents occurring every week. Levels of knowledge stating good and very good could be also improved.

Table 10. Estimation of students' knowledge about internet security (N=152)

	N	%
Insufficient	37	24.3
Sufficient	38	25.0
Good	46	30.3
Very good	26	17.1
Excellent	5	3.3

Conclusion

Internet security is important at the university and outside of it. Students are very active users of the internet because university study programs require from them participation of ICT related activities. At the same time, they are also very exposed to every kind of internet security threats as almost any other group of frequent users of the internet services. Students inform themselves about the newest security threats as much as possible through various available channels of communication and by mediation of different people to remain up to date with the current internet security developments. The research study successfully provided answers to all three research questions: RQ1: the research study confirmed that students were exposed to internet security threats; RQ2: students were able to recognize the most common and most frequent internet security threats; RQ3: students were applying the basic security threats countermeasures. All three special hypotheses of the research study were confirmed: H1: students are well acquainted with the existence of most common and frequent security threats some of which they encounter in their daily academic activities; H2: students possess knowledge about the basic security threats countermeasures; H3: the level of knowledge of students about internet security is still low. To improve the situation, students should be offered additional courses which would enable to acquire additional theoretical and also hands-on knowledge about the internet related security.

References

Abazari, F.; Madani, A.; Gharaee, H. Optimal Response to Computer Network Threats. // 8th International Symposium on Telecommunications (IST'2016), IEEE, 2016, 729-734.

Berriman, L. Framing internet safety: the governance of youth online. // Information, Communication & Society. 20 (2017), 1829-1830.

Broida, R. 560 million more passwords were exposed -- was yours?. 16.5.2017. https://www.cnet.com/how-to/protect-yourself-from-the-latest-database-breach/ (18.5.2017.)

Burges, M. Another large cyberattack is underway and it could be worse than WannaCry. 18.5.2017. http://www.wired.co.uk/article/adylkuzz-cyberattack-malware (19.5.2017.)

Byrne, Z. S.; Dvorak, K. J.; Peters, J. M.; Ray, I.; Howe, A.; Sanchez, D. From the user's perspective: Perceptions of risk relative to benefit associated with using the Internet. // Computers in Human Behavior 59 (2016), 456-468.

Communication technology fundamentals. http://inf.ffzg.unizg.hr/index.php/en/38-instruction/instruction-undergraduate-study/583-information-technology-fundamentals (13.5.2017.)

Data protection. http://inf.ffzg.unizg.hr/index.php/en/38-instruction/instruction-undergraduate-study/561-data-protection (13.5.2017.)

Google program teaches safe online exploration. // American School Board Journal. 204 (2017), 20.

Hern, A. How to protect your computer against the ransomware attack. (16.5.2017.) https://www.theguardian.com/technology/2017/may/15/windows-xp-patch-wannacry-ransomware-wecry-wanacrypt0r (18.5.3027.)

Internet culture. http://inf.ffzg.unizg.hr/index.php/en/39-instruction/instruction-graduate-study/576-internet-culture (13.5.2017.)

Jouini, Mouna, Rabai, Latifa Ben Arfa, Aissa, Anis Ben. Classification of security threats in information systems. // Procedia Computer Science. 32 (2014), 489–496.

Levin, S. Google Docs users hit with sophisticated phishing attack in their inboxes. 3.5.2017. https://www.theguardian.com/technology/2017/may/03/google-docs-phishing-attack-malware (18.5.2017.)

New drive on internet safety. // Education Journal. (2017), 7.

Schroeder, S. There's another hacking attack right now, and it's making more money than WannaCry. 18.5.2017. http://mashable.com/2017/05/18/adylkuzz-wannacry-attack/#riLJeC185Eqm (19.5.2017.)

Security Threats. https://msdn.microsoft.com/en-us/library/cc723507.aspx (18.5.2017.)

Singh, R.; Kumar, H.; Singla, R. K.; Ketti, R. R. Internet attacks and intrusion detection system. // Online Information Review 41 (2017), 171-184.

Sherr, Ian. WannaCry ransomware: Everything you need to know. 18.5.2017. https://www.cnet.com/news/wannacry-wannacrypt-uiwix-ransomware-everything-you-need-to-know/ (19.5.2017.)

Sveučilišni računski centar. http://www.srce.unizg.hr/osnovni-tecajevi/popis-tecajeva (13.5.2017.)

The 11 most common computer security threats. http://www.symantec-norton.com/11-most-common-computer-security-threats_k13.aspx (18.5.2017.)

Threat. http://www.businessdictionary.com/definition/threat.html (13.5.2017.)

van den Berg, B.; Keymolen, E. Regulating security on the Internet: control versus trust. // International Review of Law, Computers & Technology. 31 (2017), 188-205.

Vrana, R. Making the Internet a safer place: students' perceptions about Internet security threats // Proceedings of the 23rd Central European Conference on Information and Intelligent Systems, University of Zagreb Faculty of Organization and Informatics, 2012, 91-98.

Vrana, R. Online social networks and security of their users: an exploratory study of students at the Faculty of humanities and social sciences Zagreb // Central European Conference on Information and Intelligent Systems, University of Zagreb Faculty of Organization and Informatics, 2013, 214-221.

Professional paper

Cyber-attacks as a Threat to Critical Infrastructure

Roman Domović
Zagreb University of Applied Sciences
Vrbik 8, Zagreb, Croatia
roman.domovic@tvz.hr

Summary

In today's world, hybrid warfare is present like never before. The top spot holds information operations, perception management and various types of cyber-attacks. Cyber-attacks that can be carried out via critical infrastructure can disable the normal functioning and development of a society or state for a long time. Analyzing such threats and designing solutions to reduce or eliminate these threats is a challenge for current and upcoming generations of computer security and legal experts. In this research paper, some examples of cyber-attacks on critical infrastructure from this decade are analyzed to see which attack vectors are the biggest threats and what can be done to avoid or minimize its impact.

Key words: hybrid warfare, cyber-attacks, critical infrastructure, defence strategies

Introduction

In the modern world, various states, centers of power, various activist groups and individuals are trying to spread their influence. The spread of influence can be done in two ways: through hard power and through soft power. Hard power refers to military power threats and the realization of these threats. Soft power relies on the ability to shape priorities of others by influence. Instead of pushing someone to do something, the same goal is achieved through co-operation. [1] But there is a gray zone, something that is neither purely conventional warfare, nor peaceful diplomatic and economic action. It is so-called hybrid warfare, where superiority is achieved by handling information and information and communications technology (ICT) infrastructure. Apart from the problems that may arise in private and business networks, special problems can arise out of threats on critical infrastructure. Given the role critical infrastructure has, it can have serious consequences for the stability and integrity of states. There are many different attack vectors like email attachments, insecure network connection, physical access to an insufficiently protected device, web pages, operating system ex-

[1] Joseph Nye: Soft Power : The Means of Success in World Politics, pp. 5, 2004.

ploits, social engineering and human error. The aim of this research paper is to analyze several examples of cyber-attacks on a critical infrastructure, to see which attack vector represents the greatest danger and what can be done to avoid or minimize the impact of certain cyber-attack.

Critical infrastructure

Although there is no universally agreed definition, critical infrastructure is generally understood as "those facilities and services that are vital to the basic operations of a given society, or those without which the functioning of a given society would be greatly impaired".[2] According to the directive of the Council of the European Union, "critical infrastructure means an asset, system or part thereof located in Member States which is essential for the maintenance of vital societal functions, health, safety, security, economic or social well-being of people, and the disruption or destruction of which would have a significant impact in a Member State as a result of the failure to maintain those functions."[3]
The sectors covered by these definitions differ from country to country, but generally include transportation systems (air, rail, road, sea); energy production and shipping; government facilities and services, including, in particular, defense, law enforcement and emergency services; information and communication technology; food and water; public health and health care; financial institutions.[4] Today, all these resources are managed by means of information-communication technology, which opens up a special attack vector. What makes it an advantage for easier management and control is at the same time a weakness subject to attacks. Why is information-communication technology at the same time a weakness?

About cyber-attacks

According to Bruce Schneier, "there are a bunch of reasons for this, but primarily it's:

1. the complexity of modern networked computer systems and
2. the attacker's ability to choose the time and method of the attack versus the defender's necessity to secure against every type of attack".[5]

[2] NATO Parliamentary Assembly. Document 162 CDS 07 E rev 1 – The protection of critical infrastructures, 2007, http://www.nato-pa.int/Default.asp?SHORTCUT=1165. (Access date: 02. 04.2017).

[3] The Council of the European Union: Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection. (Text with EEA relevance). http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008L0114 (Access date: 02.04.2017).

[4] NATO Parliamentary Assembly. Document 162 CDS 07 E rev 1 – The protection of critical infrastructures, 2007, http://www.nato-pa.int/Default.asp?SHORTCUT=1165. (Access date: 02. 04.2017).

[5] Bruce Schneier. Attack vs. Defense in Nation-State Cyber Operations. https://www.schneier.com/blog/archives/2017/04/attack_vs_defen.html (Access date: 09.04.2017).

Why can we consider that Schneier is right? In the protection of information-communication systems it is essential to keep its functionality, easy manageability, scalability and supervision of the system. In the data security it is essential to stick to the classical principle of cryptography, i.e. to preserve the confidentiality, authenticity and integrity of the data. If we look at the examples, we can see that this is a very demanding task. For example, GSM technology and its upgrades are the world's largest security system. There are more than four billion active security features in it.[6] The implementation of GSM technology for practical purposes leads to a number of large and complex information systems. Since the information and communication system consists of not only software and hardware, but also the people who manage it and protocols and procedures under which it is being operated, it is obvious that there are many points where an error or omission may occur. Attackers who actively monitor what is happening can take advantage of it. Smaller systems, even the smallest, are subject to the same.

The attackers in front of them have a concrete system with concrete technological solutions. They have advantage over defence which has to anticipate attacks, which is practically impossible because the attacks are enhanced by the development of technology and knowledge. And it requires time that defence doesn't have at its disposal. At the time of launching a system in operation, attacks on it or its parts may be completely unknown and impossible to predict. In addition, sharing knowledge in computer security is often hampered by business secrets and the secrecy of scientific discovery until its publication in a journal or at a conference. This benefits the group of attackers who unite resources in an attempt to attack a newly discovered weak point in a system.

In general, cyber-attacks can be divided into four categories according to the type of the attack:[7]

a. hacktivism – political propaganda and protest, fun or self-proving,
b. cyberespionage – strategy aimed at obtaining critical governmental or corporate information by breaking into computer networks and systems,
c. cybercrime – motivated by economic gains through illegal penetration of computer networks and relatively non-violent in nature,
d. cyberwarfare – actions by a nationstate to penetrate another nation's computers or networks for the purpose of causing damage or disruption.

Whoever runs the attacks, needs resources to do it. Resources cost. But equally, resources also cost the defense. Rebecca Slayton in her detailed analysis of the balance between cyber offense and cyber defense balance says that improvement of various defensive practices "will not produce invulnerable organiza-

[6] Dan Forsberg, Gunther Horn, Wolf-Dietrich Moeller, Valtteri Niemi. LTE Security, pp.28, 2010.

[7] Tobby Simon. Critical Infrastructure and the Internet of Things, 2017. https://www.cigionline.org/sites/default/files/documents/GCIG%20no.46_0.pdf (Access date: 09.04.2017).

tions, but they can increase the costs to attackers and decrease the costs of defenders". And also that "innovation in software development processes and technologies can make attack much more difficult". She concludes that "offensive advantages are not inevitable in cyberspace, and they cannot be eliminated by a technological fix. Instead, gaining defensive advantage will require persistent investments in technological management, innovation, and skill".[8] All this applies to cyber-attacks on critical infrastructure. By analyzing the cyber-attacks on critical infrastructure, it is possible to draw conclusions about which attack vectors are currently the most common and which critical points of the system within critical infrastructure are the most vulnerable.

Threat analysis

For the purpose of this paper five major cases of cyber-attacks on critical infrastructure from this decade have been analyzed.

1. In 2010. Stuxnet worm was detected and it was a first worm known to attack Supervisory Control And Data Acquisition systems (SCADA). It destroyed a number of Iranian nuclear centrifuges. Symantec Security Response team did a thorough examination of Stuxnet and concluded that it was created with the aim "to reprogram industrial control systems (ICS) by modifying code on programmable logic controllers (PLCs) to make them work in a manner the attacker intended and to hide those changes from the operator of the equipment". To increase chances of success, Stuxnet authors implemented various components such as zero-day exploits, a Windows rootkit, the first ever PLC rootkit, antivirus evasion techniques, complex process injection and hooking code, network infection routines, peer-to-peer updates, and a command and control interface.[9] Stuxnet was created to attack specifically Siemens S7-300 system running centrifuges in Iran's nuclear-enrichment program. It installs malware on the PLC that monitors the Profibus of the system and under certain conditions it periodically modifies that frequency, which results in that the connected motors change their rotational speed.[10] Eventually, it leads to destruction of centrifuges. Infection starts by plugging in a USB flash drive or from the internal network if an infected machine exists. In this case, the attack vector is a human error, an error made by the operator that works in the nuclear facility complex, who inserts the infected USB into the computer connected to the facility network.

[8] Rebecca Slayton. What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment. International Security, 2017(41), No. 3, pp. 72-109.

[9] Nicolas Falliere, Liam O Murchu, Eric Chien. W32.Stuxnet Dossier, 2011. /w32_stuxnet_ dossier.pdf (Access date: 16.04.2017).

[10] Stamatis Karnouskos. Stuxnet Worm Impact on Industrial Cyber-Physical System Security, 2011, http://papers.duckdns.org/files/2011_IECON_stuxnet.pdf (Access date: 16.04.2017).

2. In 2011 there was an attempt made to breach the Information Technology (IT) systems of Lockheed Martin, the American global aerospace, defense, security and advanced technologies company. But story begins earlier when the so-called Advanced Persistent Threat (APT) attack was carried out on United States security company RSA, consisting of three phases. The first phase of such attack is studying the targets and conducting social engineering over the target by which it is fooled, causing it to install malware. The second phase is a breakthrough in the network and the search for the appropriate parts of the information system, such as a user with as many administrator access rights to servers as possible. The third phase is the ultimate activity such as data gathering, data modification, data deletion, etc. In this attack, the attacker sent two different phishing emails within two days to two small groups of employees, who at the first glance were not worth the effort, just doors to higher levels. The assumption is that the attacker had previously gathered information about these employees. The email subject was "2011 Recruitment Plan". One employee opened an Excel file that was in the email attachment, entitled "2011 Recruitment plan.xls". The Excel document contained a zero-day exploit that installs backdoor through Adobe Flash vulnerability (CVE-2011-0609). After that, Poison Ivy malware was installed on the computer, which enabled the attacker to supervise the employee's computer and break in further into the company network. The attacker found certain RAR files on one server and sent them via an FTP server to an external server.[11] There were indications that a database was stolen which links serial token numbers called RSA SecureID and "seed" that each token fills so it becomes unique. There are also indications that this data was used to attack Lockheed Martin. The attack was possible because Lockheed Martin employees, along with thousands of employees from other companies, use RSA SecureID tokens to log onto computers and other sensitive parts of information systems. In this case, the attack vector is also human error, an error made by the operator that works in the company, who opened infected file on a computer connected to the internal company network.
3. In 2014 a hacker group known as Dragonfly or Energetic Bear attacked companies from energy sector in Europe and United States. In the attack they used malware "Havex" to run into the control system of the attacked companies.[12] When Havex infiltrated these systems, he sent sensitive

[11] RSA FraudAction Research Labs. Anatomy of an Attack, April 1, 2011, URL: http://blogs.rsa.com/anatomy-of-an-attack/ (Access date: 16.04.2017).

[12] Trend Micro. Report on Cybersecurity and Critical Infrastructure in the Americas, 2015, https://www.trendmicro.de/cloud-content/us/pdfs/security-intelligence/reports/critical-infrastructures-west-hemisphere.pdf (Access date: 29.04.2017).

information back to hackers. Havex is known to be distributed to targeted users through three arrival vectors: spam emails, exploit kits and trojanized installers planted on compromised vendor sites. Security experts from F-Secure company discovered that the main components of Havex malware are a general purpose Remote Access Trojan (RAT) and a server written in PHP. They also discovered how Havex operates. "Once the Havex malware has been delivered to the targeted users and installed on a machine, it scans the system and connected resources accessible over a network for information of interest. This information includes the presence of any Industrial Control Systems (ICS) or Supervisory Control And Data Acquisition (SCADA) systems present in the network. The collected data is then forwarded to compromised websites, which surreptitiously serve as remote Command and Control (C&C) servers."[13] In this case, the attack vector is also human error, an error made by certain users who have released malware into the networks.

4. In 2014. an attack was launched on a steel factory in Germany. According to a report by Germany's Federal Office for Information Security (Bundesamt für Sicherheit in der Informationstechnik – BSI), the attackers first infected the steel factory office network by spear-phishing emails and smart social engineering. From there, they progressed through a network and other systems including systems that control the plant's equipment, to cause frequent falls of individual control components and various systems. Consequently, the operators were unable to adequately regulate and immediately turn off a blast furnace. BSI stated that final result was "massive damage to the plant".[14, 15] In this case, the attack vector is also human error, an error made by certain employees who have activated backdoor and released malware into the network.
5. On December 23, 2015 an event that according to gathered evidence points to a cyber-attack at three regional electric power distribution companies, called Oblenergos, has caused a power outage in Ukraine and made impact on approximately 225,000 customers. According to ICS-CERT report, "the cyber-attack was reportedly synchronized and coordinated, probably following extensive reconnaissance of the victim networks. According to company personnel, the cyber-attacks at each company occurred within 30 minutes of each other and impacted multiple

[13] F-Secure. Backdoor:W32/Havex : Threat description, 2014, URL: https://www.f-secure.com/v-descs/backdoor_w32_havex.shtml (Access date: 29.04.2017).

[14] Trend Micro. Report on Cybersecurity and Critical Infrastructure in the Americas, 2015, (Access date 29.04.2017.). URL: https://www.trendmicro.de/cloud-content/us/pdfs/security-intelligence/reports/critical-infrastructures-west-hemisphere.pdf (Access date: 29.04.2017).

[15] F-Secure. Backdoor:W32/Havex : Threat description, 2014, URL: https://www.f-secure.com/v-descs/backdoor_w32_havex.shtml (Access date: 29.04.2017).

central and regional facilities. During the cyber-attacks, malicious remote operation of the breakers was conducted by multiple external humans using either existing remote administration tools at the operating system level or remote industrial control system (ICS) client software via virtual private network (VPN) connections. The companies believe that the actors acquired legitimate credentials prior to the cyber-attack to facilitate remote access." Attackers executed the KillDisk malware and wiped some systems, probably in the way that "KillDisk malware erases selected files on target systems and corrupts the master boot record, rendering systems inoperable". More damage has been done by KillDisk's overwriting of Windows-based human-machine interfaces (HMIs) embedded in remote terminal units, corrupting firmware of Serial-to-Ethernet devices and making them inoperable and scheduling disconnects for server Uninterruptable Power Supplies (UPS) via the UPS remote management interface. Companies also reported that they had been infected with BlackEnergy malware which was delivered via spear-phishing emails with malicious Microsoft Office attachments. It is suspected that it may have been used as an initial attack vector to acquire legitimate credentials.[16] ICS-CERT report does not confirm that BlackEnergy played a role in this cyber-attack, but it looks so and other sources support it.[17, 18] In this case, the attack vector is also human error, an error made by certain employees who have activated backdoor and released malware into the network.

Discussion

Attacks have occurred and the damage is done. Based on the performance of the attacks and the spotted defects in defense it can be analyzed which scenarios should occur so that the attacks are unsuccessful and without or with less damage. Unfortunately, all essential attacks detail and countermeasures needed for thorough in-depth analysis are not available. Therefore, after the analysis carried out on the basis of available information, a synthesis of the necessary defense countermeasures can be made.

Particularly interesting case is a Stuxnet breach into the uranium enrichment plant in a desert outside Natanz in central Iran. This facility is buried more than

[16] ICS-CERT. Alert (IR-ALERT-H-16-056-01) : Cyber-Attack Against Ukrainian Critical Infrastructure, 2016. URL: https://ics-cert.us-cert.gov/alerts/IR-ALERT-H-16-056-01 (Access date: 30.04.2017).

[17] Trend Micro. Frequently Asked Questions: BlackEnergy, 2016. URL: https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/faq-blackenergy (Access date: 30.04.2017).

[18] F-Secure. Backdoor:W32/BlackEnergy: Threat description. URL: https://www.f-secure.com/v-descs/backdoor_w32_blackenergy.shtml (Access date: 30.04.2017).

15 meters beneath the desert surface. Its wall and roof are reinforced with concrete and covered with layers of earth and it is heavily guarded.[19] This prevents unwanted surveillance over the facility, physical entry into the facility and partially protects it from missiles and different kinds of armed attacks. Regarding to cyber-attacks, it prevents side channel attacks on the internal electronic devices and communication. Internal network, computers and various electronic devices such as International Atomic Energy Agency (IAEA) digital surveillance cameras installed inside the facility are all air-gapped. It means that they are isolated from the internet or any other external network, which prevents direct intrusion by remote attackers. Policies and procedures are arranged so that the air-gap could not be bypassed. It looks like all the necessary protection measures have been taken so that any type of cyber-attack taken from the outside is not possible. But still, the successful attack has been done. How? Since it is impossible to access internal devices from the outside because of the air-gap, the air-gap needs to somehow be bypassed. Now, components inside the plant must be updated from time to time. Whether it is updating of operating systems, software, hardware or firmware, whether it is adding new components to a facility that should be connected to others, these operations are usually conducted by outside contractors. Technicians who are employees of companies who as outside contractors collaborate with a nuclear facility, have the ability to enter the plant and perform upgrades of the system. For the upgrade, it is necessary to add new parts of the upgrade to the existing components. To do this, it is necessary to connect existing components to a laptop, tablet or USB flash drive of an external technician, or to insert a CD / DVD into it. And there is the air-gap bypass.

Because of the need for upgrading, an absolute air-gap is not possible. Stuxnet was sent to spread across the world to increase the possibility of breach, with the main target – USB flash drives of four carefully selected companies that were outside contractors of Natanz nuclear facility, dealing with "manufacturing products, assembling components or installing industrial control systems".[20] These companies were a gateway and its infected technicians were carriers which passed Stuxnet inside Natanz facility and bypassed the air-gap. So, a scenario in which no air-gap bypass comes up includes a thorough check of every device entering the plant, conducted on separate pieces of equipment that are also air-gapped. It is time consuming and requires additional resources but reduces the possibility of breach. For such cases, an optimal solution should be found, but it cannot be presented in this paper because any such solution depends on the specific situation in a specific environment.

[19] Kim Zetter: Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon (2015).

[20] Ibid.

In four other cases, there were procedures under which employees should not open suspicious files or click on suspicious links. Despite this, the breaches occurred by phishing emails and compromised web sites which download payloads to an access computer. A scenario in which such breaches cannot occur involves administrator procedure of blocking every incoming file on the email server, sending message to a recipient that there is a file for him and that it should be checked on the air-gapped machine before being delivered to a specific computer. But this procedure can create a massive queue on email checking which can affect working efficiency and cause harm to operational capabilities of certain critical infrastructure. There can be other solutions and again, an optimal solution depends on the specific situation in a specific environment.

From all analyzed cases it is possible to extract elements of cyber-attacks that represent the greatest threat. The greatest threat that cannot be entirely resolved will remain the existence of zero-day exploits and carefully programmed malware which can remain undetected despite sophisticated digital-forensic equipment, all-rounded procedures, staff knowledge and experience. Secondary to that is the impact of social engineering. If the precautionary measures are intensified, it may be significantly reduced.[21] It would be irresponsible to say that it can be completely eliminated.

Also, it is possible to extract elements of cyber-attack countermeasures that are usually implemented, but must be improved. These are:

- incomplete procedures > that should be all-rounded to be able to prevent the air-gap bypassing,
- insufficient education of employees who are subject to social engineering > there must be constant raising of awareness of the methods of social engineering, data protection, information-system security and above all awareness of the need to protect critical infrastructure as a whole,
- insufficient coordination with outside contractors > there must be service level agreement (SLA) which determines the course of action in accordance with the defensive strategies of a particular critical infrastructure.

Critical infrastructure must be protected physically and procedurally from all known types of attack. In addition, as much as possible, new types of attacks must be foreseen and accordingly there has to be a modular defense strategy. Implementation of defensive strategies depends on certain type of critical infrastructure, its business operation and cost/benefit analysis of cyber security investments. Due to the importance of critical infrastructure for the functioning of a society, defence strategies must be fully met.

[21] For an example see Bullee, Montoya, Pieters, Junger and Hartel: The persuasion and security awareness experiment: reducing the success of social engineering attacks. Journal of Experimental Criminology, March 2015, Volume 11, Issue 1, pp. 97-115.

Conclusion

Hybrid warfare is a present danger. Cyber-attacks are being carried out frequently and the consequences are very expensive. Especially when it comes to attacks on critical infrastructure. From analyzed examples it can be concluded that the attacks are executed according to the APT attack pattern. In all cases, such attacks began by gathering information about certain employees who are then manipulated and deceived by social engineering which forces them to unconsciously install malware into the computer. That action allows attackers further penetration into the network and causing damage.

Part of the package of solutions that provide resilient critical infrastructure, in addition to high-quality security specialists and technology solutions, must include a scenario in which vital part of information systems must be procedures that prevent air-gap bypassing as well as continuing education of personnel in terms of computer security and how to not become a victim of social engineering who makes fatal errors.

For a thorough study on this topic, more examples should be analyzed with in-depth look at targeting infrastructure and types of cyber-attack methods.

References

Bullée, Jan-Willem H.; Montoya, Lorena; Pieters, Wolter; Junger, Marianne; Hartel, Pieter H. The persuasion and security awareness experiment: reducing the success of social engineering attacks. // Journal of Experimental Criminology. March 2015, Volume 11, Issue 1, pp. 97-115.

Falliere, Nicolas; O Murchu, Liam, Chien, Eric. W32.Stuxnet Dossier: Version 1.4. Symantec Security Response, 2011.

https://www.symantec.com/content/en/us/enterprise/media/security_response/whitepapers/w32_stuxnet_dossier.pdf (16.04.2017).

Forsberg, Dan.; Horn, Gunther.; Moeller, Wolf-Dietrich.; Niemi, Valtteri. LTE Security. Wiley; 2 edition, 2012.

F-Secure. Backdoor:W32/BlackEnergy: Threat description. https://www.f-secure.com/v-descs/backdoor_w32_blackenergy.shtml (30.04.2017).

F-Secure. Backdoor:W32/Havex: Threat description, 2014. https://www.f-secure.com/v-descs/backdoor_w32_havex.shtml (29.04.2017).

ICS-CERT. Alert (IR-ALERT-H-16-056-01): Cyber-Attack Against Ukrainian Critical Infrastructure, 2016. URL: https://ics-cert.us-cert.gov/alerts/IR-ALERT-H-16-056-01 (30.04.2017).

Karnouskos, Stamatis. Stuxnet Worm Impact on Industrial Cyber-Physical System Security. // IECON 2011 – 37th Annual Conference on IEEE Industrial Electronics Society / Institute of Electrical and Electronics Engineers (IEEE), 2012, pp. 359-364 http://papers.duckdns.org/files/2011_IECON_stuxnet.pdf. (16.04.2017).

NATO Parliamentary Assembly. Document 162 CDS 07 E rev 1 – The protection of critical infrastructures, 2007, http://www.nato-pa.int/Default.asp?SHORTCUT=1165 (02.04.2017).

Nye, Joseph. Soft Power: The Means of Success in World Politics. 2004, New York: Public Affairs, 2004.

RSA FraudAction Research Labs. Anatomy of an Attack, April 1, 2011, http://blogs.rsa.com/anatomy-of-an-attack/ (16.04.2017).

Schneier, Bruce. Attack vs. Defense in Nation-State Cyber Operations. https://www.schneier.com/blog/archives/2017/04/attack_vs_defen.html (09.04.2017).

Simon, Tobby. Critical Infrastructure and the Internet of Things. Centre for International Governance Innovation, Chatham House, 2017, https://www.cigionline.org/sites/default/files/documents/GCIG%20no.46_0.pdf (09.04.2017).

Slayton, Rebecca. What Is the Cyber Offense-Defense Balance? Conceptions, Causes, and Assessment. International Security. Winter 2016/17, Vol. 41, No. 3, pp. 72-109.

The Council of the European Union. Council Directive 2008/114/EC of 8 December 2008 on the identification and designation of European critical infrastructures and the assessment of the need to improve their protection. (Text with EEA relevance). http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32008L0114. (02.04.2017).

Trend Micro. Frequently Asked Questions: BlackEnergy, 2016, https://www.trendmicro.com/vinfo/us/security/news/cyber-attacks/faq-blackenergy (30.04.2017).

Trend Micro. Report on Cybersecurity and Critical Infrastructure in the Americas, 2015, https://www.trendmicro.de/cloud-content/us/pdfs/security-intelligence/reports/critical-infrastructures-west-hemisphere.pdf (29.04.2017).

Zetter, Kim. Countdown to Zero Day: Stuxnet and the Launch of the World's First Digital Weapon. Broadway Books; Reprint edition (September 1, 2015).

E-HEALTH APPLICATIONS AND SOLUTION

Original scientific paper

Digital Technology as a Tool in Self-management of Painful Low Back Syndrome

Mirjana Berković-Šubić
Health Centre in Zagreb County
Gajeva 37, Samobor, Croatia
mirjana.berkovic@gmail.com

Gilbert Hofmann
Health Resort Veli Lošinj
Podjavori 27, Veli Lošinj, Croatia
gilbert@net.hr

Biserka Vuzem
Special Hospital for Medical Rehabilitation Krapinske Toplice
Gajeva 2, Krapinske Toplice, Croatia
biba.vuzem@gmail.com

Summary

The development of electronic media has enabled the health service users an open access to different medical information. The patients are not satisfied only with the medical report and diagnosis made by a professional, they want to know more. Therefore digital technology is used to gain more information about the painful syndromes in the lower back. This process can be called e-low back pain. The pain that occurs in the lower back is called lumbago and it can have several causes. The aim of this paper is to analyze the lumbago patients' active participation in order to achieve new knowledge and to expand the already existing ones using different internet sources. The hypothesis stated in the research: H1: Participants show interest and positive attitude in using information technology in order to assist in the back pain treatment. This research carried out the result analysis of the E-back pain questionnaire about the usage of information technology and the digital communication capabilities in order to achieve new knowledge and to expand the already existing one related to the back pain. 120 participants responded to the research, dominantly middle-aged (average age 49.25 years). In conclusion, participants express their positive opinion about using modern information technologies in acquiring and expanding their knowledge about the back pain.

Key words: lumbago, back pain, electronic media, e-back pain, open access, the Internet

Introduction

The development of information and communication technology has caused radical changes in new knowledge acquisition which has become easily accessible to interested individuals.

When technology like computers, internet, and multimedia devices are included in the learning process, we can call it e-learning. Easily accessible contents on Internet enable the progress in upgrading and expanding the existing cognition. In fast-growing economies profit plays the most important role and the employees are forced to work even if they have musculoskeletal problems and they cannot exercise their right to get a granted medical leave because of pain. In the last decades the number of back pain patients has increased. (Deyo et al, 2014). Chronic back pain treatment is one of the most common reasons for visiting a physician or a physiotherapist (Lin et al, 2011),

In modern society many new professions have emerged and the employees' standing or sitting position is required throughout the whole working hours which often results with the low back pain (LBP). The consequence of a long-term sitting or standing position is the improper load of musculoskeletal system. Therefore some muscles are overloaded while the others weaken due to inactivity. If the improper body positions last longer and are repeated on a long term basis, the result is a chronic, painful condition. The back pain lasting longer than twelve weeks is called chronic lumbago (American Pain Society, 2007). Lifelong back pain prevalence is up to 80% (Smith et al, 2014).

In order to change their painful condition, the patients visit a physician to set a correct diagnosis and to determine the cause of pain. Patients often use e-media to expand their knowledge and also to find information about some therapeutic exercises they could practice in a home environment to treat the back pain and by doing that, they take part in forming the concept of e-back pain. In the past the information about health care was quite unavailable to patients and they could be given some only in direct contact with a particular medical practitioner. Nowadays the availability of all kinds of medical educational content has enormously increased due to the broad media support.

In the process of physiotherapy, patients are active participants, so the physiotherapist gives them advice how to overcome the daily activities, perform exercises, and the adaptation at work. Due to inability to get a granted leave, a person with back pain practices learned exercises at home on a daily basis to improve their painful condition. There are many people who do not have medical education but want to know more about their illness and its outcomes. The most widely used media is the Internet that offers countless opportunities. The Internet is a global data network that is publicly available and offers various services (Croatian declaration on open access, 2012).

Internet technology offers a wide range of solutions to acquire and improve the already existing knowledge. For example, the social networks like Facebook, as the most accessible one, publishes a large amount of visual and textual material

from different Internet portals, and the readers can choose the content according to their interests or medical difficulties.
Since the back pain is a serious medical condition, its treatment should be approached according to the recommendations of the world and national guidelines for the low back pain treatment (Grazio et al, 2012)
The aim of this research is to determine how much the respondents use the electronic media in order to expand their knowledge about the back pain and to assist in the chronic back pain treatment.
Hypothesis H1: The participants show interest and positive attitude in using the electronic media in order to assist in the back pain treatment.

Materials and methods

Participants for this study (N=120) were randomly selected from a group of low back pain treated patients at Special Hospital for Medical Rehabilitation Krapinske Toplice.
The approval for conducting this research was obtained by the Special Hospital for Medical Rehabilitation Krapinske Toplice Ethical Committee and with the consent of the hospital director.
The fundamental ethical integrity of the respondents was also respected and no data abuse used for other purposes.
The research was carried out in the period from January to April 2017, by anonymous, a specially constructed questionnaire with 10 questions for the research purposes created by the author.
By completing the survey, each respondent was familiar with the reason and research protocol confirmed their participation with signature. The questionnaire contained besides demographic questions also questions related to the use of electronic media and Internet capabilities in purpose of assisting in the back pain treatment. The questions were formulated with the ability to respond to Likart's scale (5 answers). The results are statistically processed and expressed as average values, the differences were compared by variance analysis, and statistical significance was determined as $p<0.05$.
The comparison of average attitude values in relation to age, gender and degree of education was made.

Results

This research describes the participants' interest and attitudes regarding the use of modern technologies to assist in the chronic back pain treatment. Among 120 participants (N=120) there were 55% women and 45% men, average age 49.25.
The participants aged from 25 to 77 were divided into four age groups:

- group aged less than 40
- group aged from 41-50
- group aged from 51-60
- group aged more than 60.

The majority of the participants were in the age group of 51 to 60 years of age, 33.3% (N 40), and the least of them 16.7% (N20) were in the age group over 60 years (Table 1).

Table 1. Presentation of the respondents by gender and age

Respondents	Gender		Age (Grades in years)			
	M	F	<40	41-50	51-60	>60
Number	54	66	29	31	40	20
Percentage (%)	45%	55%	24.2%	25.8%	33.3%	16.7%
Total (N)	120		120			

Presentation of the demographic characteristics of the respondents

According to the education level, the majority of the respondents belong to a group of unskilled or qualified workers, 50%, 31.66% of secondary education respondents and 16.33% of the respondents have college or university degree.

Table 2. Presentation of the respondents according to their qualifications in relation to the age group

Age	Qualifications			Total	%
	unskilled/ vocational education	high school	college/ university degree		
<40	17	8	4	29	24.0%
41-50	13	12	6	31	26.0%
51-60	18	14	8	40	33.3%
>60	12	4	4	20	16.7%
Total	60	38	22	120	100.0%

Distribution of frequency response by individual statement

In statement T1: The use of modern information technologies (Internet, social networks) helps me in gaining and expanding knowledge about back pain; 93 respondents expressed positive attitude, 20 neutral and 7 of them negative attitude.

In statement T2: Available tips in the form of a movie or a back pain image are useful; 102 of the respondents expressed positive attitude, 15 neutral and 3 negative attitude.

In statement T3: My physiotherapist's information on the possibilities of using information technology to meet my illness was useful; 95 respondents expressed positive attitude, 17 neutral and 8 expressed negative attitude.

In statement T4: The most information about the back pain I got by using the Internet; 28 respondents expressed positive attitude, 18 neutral and 74 negative attitude.

In statement T5: I use tips from scientific biomedical databases to treat my back pain; 43 respondents expressed positive attitude, 31 neutral and 46 negative attitude.

In statement T6: The Internet useful to me as a reminder to carry out previously learned back pain treatment exercises; 82 respondents expressed positive attitude, 17 neutral and 21 negative attitude.
In statement T7: I do not dare to use Internet and social networking self-help back pain tips; 55 respondents expressed positive attitude, 25 neutral and 40 negative attitude.
In statement T8: The lack of skill for modern information technologies (Internet, social networks) is a problem for me in acquiring new knowledge about my health; 31 respondents expressed positive attitude, 17 neutral and 72 negative attitude.
In statement T9: No knowledge of English is a problem for me to acquire new information about back pain; 51 respondents expressed positive attitude, 10 neutral and 59 negative attitude.
In statement T 10: I cannot do the back pain treatment without a specialist; 108 respondents expressed positive attitude, 3 neutral and 9 negative attitude.
All the asserted claims were handled individually and express the respondents' attitudes about the modern technology for the purpose of finding useful tips for the back pain treatment and the obtained results are expressed as average results (Table 3 and Table 4).

Table 3. Presentation of the results of the individual statements and the respondents' attitudes about the usefulness of information on the Internet about the back pain in relation to the age

	Age (grades in years)				**Anova**	
	<40	**41-50**	**51-60**	**>60**		
Statements	M	M	M	M	F	Sig.
S1	3.76	4.13	4.08	3.35	5.712	0.001
S2	3.90	4.26	4.03	3.60	4.639	0.004
S3	3.86	3.97	4.13	3.60	1.711	1.69
S4	2.55	2.77	2.58	1.90	3.512	0.018
S5	2.52	3.29	3.10	2.25	6.138	0.001
S6	3.55	3.87	3.60	3.05	3.156	0.027
S7	3.28	2.90	3.13	3.65	2.319	0.079
S8	2.28	2.77	2.30	2.95	2.860	0.040
S9	2.24	2.90	2.85	3.70	7.014	0.000
S10	2.24	4.13	4.13	4.50	1.009	0.391
N	29	31	40	20	120	

Overall looking at the results obtained, it is evident that the positive attitude of the respondents over the use of modern technologies prevails for the purpose of the assisting in the back pain treatment.
Men express slightly more positive attitudes than women, but there was no statistically significant difference in the views of these two groups. By comparing different age groups in attitudes towards the use of modern technologies for the purpose of the assisting in the low back pain treatment, it is evident that positive

attitudes are expressed by respondents aged 41-50 and 51-60, and with the age increase (more than 60 years) this positive trend decreases. Compared to the age, a statistically significant difference in positivity was confirmed in relation to the statements. S1= p< 0.001, S2= p< 0.004, S4= p< 0.018, S5= p< 0.001, S6= p< 0.027 i S8= p< 0.040. A slightly lower attitude was found in the youngest age group (<from 40 years), however, in this group most of the respondents are with lower qualifications who are exposed to heavier physical activity (and the problems with back pain), as well as fewer opportunities to use information technology at the workplace, therefore a lower inclination to the same. The correlation between occupation and professional background and attitudes about the usefulness of Internet information about the low back pain shows a statistically significant difference in the claims S8= p< 0.032 i S9 = p< 0.017, where the possible reasons mentioned is the ignorance of the use of the information technology and the foreign language (English).

Table 4: Presentation of the results of the individual statements and the respondents' attitudes about the usefulness of information on the Internet about the back pain in relation to the level of education

	Level of education			**Anova**	
	unskilled vocational	**High school**	**Coll. Univ**		
Statements	M	M	M	F	Sig.
S1	3.80	3.92	4.09	1.143	0.322
S2	3.88	4.08	4.09	1.385	0.254
S3	3.92	3.92	4.00	0.077	0.926
S4	2.43	2.61	2.55	0.362	0.697
S5	2.75	3.05	2.86	0.918	0.402
S6	3.53	3.58	3.64	0.096	0.909
S7	3.22	3.32	2.91	1.121	0.330
S8	2.73	2.47	2.05	3.544	0.032
S9	3.10	2.82	2.27	4.238	0.017
S10	4.28	4.26	3.95	1.288	0.280
N	60	38	22	120	

Based on the obtained results, we confirm H1 hypothesis that the respondents show interest and positive attitude towards the use of information technology to assist in the treatment of the low back pain.

Discussion

The aim of this paper was to determine how many patients use modern information and communication technology to help with acute pain. The use of digital technologies and the Internet is of great help to medical service users in their efforts to study their painful conditions and apply what they've learned in everyday life using digital information. Global Health Network Supercourse includes cooperation from 81 countries with more than 750 experts from global

health, epidemiology and the Internet and gives the reader a lot of medical information (Global Health Network Supercourse, 2016).

The results obtained in the research show an overall positive attitude of respondents to using modern digital technologies in order to help treat back pain, the sex of the respondents brings no significant difference, and the positive attitude decreases with increase in age and lower qualifications.

By applying information and communication technologies, the communication between all participants in the health care process is improved: the doctor, the physiotherapist and the patient. There are very useful freely available websites that are mostly regulated by physiotherapists from private healthcare institutions and physiotherapists' associations. Such online rehabilitation programs give the patient the assurance that they are therapeutically valid because they are recommended by a health professional from that clinical area.

Due to the long-term inaccessibility of physiotherapists in public health institutions, such form of remote consultation and patient education over the Internet improves the health care of patients in need.

In this paper, a positive attitude on the use of modern technologies (Internet, social networks) with the goal of gaining and extending knowledge on back pain is in 77.5% of respondents. The availability of advice in the form of a video or pictures of the back pain is considered useful by 85% of the respondents. Physiotherapists' guidance on the use of digital technologies for learning about the illness is considered to be useful by 79% of respondents. To the statement "I received the most information about back pain by using the Internet" 23.3% of respondents agree, which is not unusual when we look at the structure of the respondents (according to the degree of education most of the respondents belong to the group of unqualified workers (50%) and 31.66% of respondents have a high school diploma). A positive attitude on the use of advice from biomedical databases for treating back pain is present in 35.8% of respondents.

A highly positive attitude towards using the Internet as a reminder to conduct already learned exercises for treating back pain is present in 68.3% of respondents. With the purpose of self-treating back pain 45.8% of respondents are afraid of using advice from the Internet or social media, and 33.3% have a negative attitude about it. 25.8% respondents believe the problem lies in not knowing informational and communicational technologies in order to gain new knowledge about their condition. No knowledge of English is a problem in acquiring new knowledge about back pain in 49.1% of respondents because they have a lower level of education and older generations don't use English in everyday communication. Considering our respondents/patients suffer from back pain, which is why they are in a rehabilitation facility in the first place, their concern for gaining new information about the condition that is present in their everyday life is perfectly justified. One also can't neglect the fact that respondents with low and medium levels of education (50%) aren't familiarized with digital technologies because their work status is tied to heavy physical work.

Younger generations use the Internet but have given average scores because they aren't in major contact with the condition and have more faith in professionals and therapeutic procedures with professionals where they can conduct their rehabilitations in person. Older generations give lower scores because of their lower computer literacy as they have a hard time searching the Internet for information. Middle-aged respondents between the ages of 41 and 60 (59.1%) can use computers, search the internet and use the information and knowledge gained so they have given higher scores as they already know medical information gained from the Internet is useful and reliable.

Conclusion

The rapid development of information and communication technologies has brought great benefits and advances in the ability to acquire new knowledge for healthcare users. Increased education and use of the Internet for self-treatment allows people with back pain to have an easier way of learning about their condition and using the pain relief programs offered. In this paper, the acquisition of new knowledge is rated as a high positive rating, which confirms that respondents are aware of the need for the use of modern technologies and forms of digital communication for the purpose of acquiring new knowledge and their application. The positive attitude of respondents to the use of modern technology to find useful counselling in treating back pain decreases as the age increases. The relationship of interest, expertise and attitudes about the usefulness of Internet information about back pain depends on knowledge of digital technology and foreign language (English).

References

Deyo, Richard A; Dworkin, Samuel F; Amtmann, Dagmar; Andersson, Gunnar; Borenstein, David; Carragee, Eugen; et al.. Report of the NIH Task Force on research standards for chronic low back pain. //Journal Pain. 2014;15(6):569-85

Lin, Chung-Wei Christine; Haas, Marion; Maher, Chris G; Machado, Luciana AC; van Tulder, Maurits W. Cost-effectiveness of guideline-endorsed treatments for low back pain: a systematic review. //Eur Spine J. 2011;20(7):1024-1038

American Pain Society 2007. Available at: http://annals.org/article.aspx?articleid=736814, (Accessed on 2.1. 2016.)

Smith, Benjamin E; Littlewood, Chris and May, Stephen. An update of stabilization exercises for low back pain: a systematic review with meta-analysis.// BMC Musculoskelet Disord. 2014; 15:416

Croatian declaration on open access. Available at: http://www.uaos.unios.hr/artos/Hrvatska_deklaracija_o_otvorenom_pristupu.pdf (Accessed 6. 10. 2015.)

Grazio, Simeon; Ćurković, Božidar; Vlak, Tonko; i sur. Dijagnostika i konzervativno liječenje križobolje: pregled i smjernice Hrvatskoga vertebrološkog društva. //Acta Med Croat 2012; 66: 259-94

Global Health Network Supercourse. http://www.pitt.edu/~super1/ (Accessed on 2.1. 2016.)

COMMUNITY INFORMATICS AND SERVICE-LEARNING

Service-Learning and Digital Technologies

Sara Semenski
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
sara.semenski@gmail.com

Aidan Harte
National University of Ireland Galway
University Rd, Galway, Ireland
aodhan-mac-airt@hotmail.com

Nives Mikelić Preradović
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
nmikelic@fzg.hr

Summary

This paper describes the project developed during the Europe engage project student tour, where students from 11 European universities have collaborated in a multicultural service-learning experience. The main goal of the project was to produce a video which would enhance the work and mission of community partner – Cell EXPLORERS. The video represents the vision and mission of the Cell EXPLORERS workshop programme and it will be useful for each of the Cell EXPLORER's university partners to increase the student volunteer base and participation in STEM subjects in schools, while simultaneously creating highly trained and competent educators. The video is targeted at a broad audience and it incorporates drawings made by workshop participants, along with volunteer interviews, documenting the fact that volunteers and participants of the Cell EXPLORERS workshops were involved in the creative process of the video making. Since the aim of the workshop and the video was to dispel the prevailing stereotypes of scientists and to increase participation in STEM subjects, children's voices were also incorporated to match the stills of the children's drawings.

Key words: service-learning, community-based learning, videos, digital technology, information and communication technology, Europe Engage, community informatics

Introduction

Service-Learning (SL) or community-based learning is widely defined as a form of experiential education that integrates meaningful community service into the curriculum. SL contains two main elements: engagement within the community (service) and reflection on that engagement (learning) [6]. These two elements should be balanced in a way that students "participate in an organized service activity that meets identified community needs", and "reflect on the service activity in such a way as to gain further understanding of course content, a broader appreciation of the discipline, and an enhanced sense of civic responsibility" [1].

SL represents a means of university's community outreach, tying the goals of higher education to the community needs through active participation of university students in structured activities that address community needs [2]. Utilizing their knowledge and skills for the improvement of local communities, students develop many transversal skills, including critical thinking and interpersonal skills [5]. Student reflection in SL occurs before, during and after the student service, so that students can recognize the importance and impact of the service on the local community and on their own learning.

Although being well established in the institutions of higher education in North America, Western Australia, and New Zealand, in 2015, before the *Europe Engage* project, little was known about SL within the European universities, apart from the isolated institutional experiences in some countries and a few national networks (Campus Engage in Ireland: http://www.campusengage.ie, Service-Learning University NetworkApS(U) in Spain: http://sites.google.com/site/redapsuniversitario and the University Network on Social Responsibility and Higher Education in Germany: http://www.netzwerk-bdv.de/content/home/index.html).

The overall aim of the project *Europe Engage – Developing a Culture of Civic Engagement through Service-Learning within Higher Education in Europe* [Reference 2014-1-ES01-KA203-004798] is to promote service-learning (SL) as a pedagogical approach that embeds and develops civic engagement within European higher education, students, staff and the wider community [4].

The project partners represent a breath of 12 universities in Europe committed to civic engagement and service-learning: Autonomous University of Madrid (Spain); National University of Ireland, Galway; University of Zagreb (Croatia); University of Brighton (United Kingdom); University of Duisburg-Essen (Germany); Erasmus University of Rotterdam (Netherlands); Instituto Superior de Psicología Aplicada (Portugal), University of Bologna (Italy); Vytautas Magnus University (Lithuania); Ghent University(Belgium); University of Applied Science-Krems (Austria) and University of Helsinki (Finland).

During the three years, the project has benchmarked existing SL practices across the disciplines in all 12 EU countries and mapped the repository of knowledge (a database of SL trainers, training materials and bibliography)

within each of the partners' countries. In the final year of the project, students from partner universities have collaborated in a multicultural service-learning experience organized at the National University of Ireland. The student project described in this paper is the result of that collaboration.

Digital Technologies and Service-Learning

Due to its strong emphasis on the community engagement, group work and the on-site engagement, SL was for a long time assumed quite incompatible with ICT, the latter implying individual work, (mostly) online communities and digital literacy. However, in the past decade, new terms such as *Technology-based Service-Learning*, *e-Service-Learning*, *Digital Service-Learning* and *Service-e-learning* have emerged. The most prominent among them, *Service-e-learning*, is defined as "an integrative pedagogy that engages learners through technology in civic inquiry, service, reflection, and action" [10]. The main aim of this pedagogy is to link digital technology to a meaningful community service, to utilize ICT in order to further improve the quality of the civic engagement and to fill the digital gaps within the local community.

Also, the analysis of SL course syllabi that are available online shows a growing trend of video journaling. Reflective journaling is commonly used in SL courses as a means of critical reflection that provides structure to often unpredictable and unstructured experiences [9]. But, all until recently, students in SL courses were most often requested to write journals, while the use of video blogging and video journals was scarce [8]. The increasing availability of cameras, video editing software, and hosting space has allowed students to submit their critical reflection in SL courses through weblogs and online journals. Using video journals provides service-learning with an additional layer of complexity, visibility, and learning for all participants [7]. Video journaling is especially useful for international SL courses, where students spend the semester in a society that is culturally different than their own and video journaling enables them to express not only their personal experiences (thoughts, reactions, and emotions), but also references from the course content and their new understanding of that content based on their international SL experience.

Service-e-learning project: video planning and realization

As a part of the *Europe Engage* project, students from partner universities collaborated in a multicultural service-learning experience designed by the Community Knowledge Initiative (CKI) staff at NUI, Galway. Students chose to collaborate with one of five wide-ranging community partners; Cell EXPLORERS, Galway Community Circus, Saol Cafe, ReelLifeScience, and Smart Consent. Students who chose the same community partners were then grouped together in groups of twos and threes for their respective projects.

The main goal of the project was to produce a digital media resource which would enhance the work and mission of community partners through a collabo-

rative process with students. It was collectively decided upon that a brief, two minute video would best serve each project and, following a training workshop with TechSpace, students were sufficiently equipped to make short creative video using *Adobe Premiere Elements 12* software. Planning implementation, idea development, media creation, skill application, and outcome reflection were the main stages in the production process. The first stage was crucial in the process, however, as it depended upon clear communication between community partners and students to simultaneously understand what the community partners wanted to express, and what the students expected to achieve. After a brief PowerPoint presentation by each of the community partners, students and community partners met in smaller groups to discuss which direction their projects would take, conscious of both process risks and end product.

Cell EXPLORERS, the community partner in the service e-learning project described in this chapter, was established in 2012 (http://www.cellexplorers.com/how-we-began) to inform, inspire, and involve people in the excitement of science. With the concept of Community-Based Learning (i.e. Service-Learning) as a guiding framework, the key was to understand the benefits for all who participated in the Cell EXPLORERS programmes; including universities, communities, academic staff, volunteers, and children, and how best to compress that concept into a two minute video. The brief included some key aspects of the Cell EXPLORERS programme which academic staff wanted to highlight in particular, and this resulted in a theme for the video concept. Inspired by Cole [3], the founding director, Dr. Muriel Grenon, created an interactive workshop programme targeted at both primary and post-primary students, with the objective of dispelling prevailing stereotypes of scientists, complementing the didactic methods of teaching, while increasing participation in STEM subjects. Involvement by both undergraduate and postgraduate students in the facilitating of workshops provides the opportunity to nurture their teaching skills, develop self-confidence, and increase competency at their work.

On that basis, it was decided that, although not always a good idea as a marketing strategy, the video would attempt to appeal to a broad target audience and contain some comedic value. With the restrictive timeframe, and students' limited skillset, a video, which represents the fact that volunteers and participants of the Cell EXPLORERS workshops were involved in the creative process of the video making, was created. Both the volunteers and the participants of the Cell EXPLORERS workshops were included in each stage of video's development. The video incorporates drawings by workshop participants, along with volunteer interviews and children's voices that match the stills of their drawings. After many long hours of filming, recording, editing, discussion, and a plethora of email correspondences between students and Cell EXPLORER staff, this service-e-learning project was completed. It is publicly available on the following link: https://www.youtube.com/watch?v=YM6f3G3Jh0s.

Service-e-learning project results and benefits

The Cell EXPLORERS video represents the vision and mission of their workshop programme, and will hopefully form part of their broader public awareness campaign. Ultimately, the digital images, videos, and audio files will be utilized by each of the Cell EXPLORER university partners in a public awareness campaign to increase its student volunteer base and participation in STEM subjects in schools, while simultaneously creating highly trained and competent educators in the process.

As mentioned in the video, the training of volunteers and demonstrators is taken very seriously at Cell EXPLORERS, and volunteers must achieve a high standard of training before facilitating workshops. Hopefully, this will help to increase trust from primary and post-primary science teachers, and raise participation by regional schools.

Discussion

The process of video editing in this service-e-learning project presented its own challenges, as students' combined backgrounds were in sociology, political science, librarianship, Croatian language and literature and advocacy. This meant that, at the beginning of the project, students were complete novices at video editing. Students have quickly adapted, however, and through the exchange of opinions, sharing to new ideas, and the overcoming of technical problems, worked together in a team of students (both primary & 3rd level), and academic staff to deliver a video which fitted the brief accordingly.

The inclusion of staff and volunteers in the production process presented its own set of challenges, and students had to decide upon the level of autonomy over each stage of the project. This has increased students' ability to negotiate through the process which neither community partner nor Europe Engage students had much prior experience in. The Cell EXPLORERS team contributed to the creative project from the start, which was beneficial for the entire production process, and they provided great support at every stage. This included their involvement in interviews, the procuring of source materials, along with invaluable production tips.

Digital technologies also played an integral role in the entire project, and the final product met its brief entirely, increasing students' knowledge of the challenges facing the scientific community today, and will hopefully lead to increased participation in STEM subjects in schools.

Finally, student participation, both in the production process of the videos, truly felt like students’ voices were finally being accepted as valid contributions to the whole service-learning debate.

Conclusion

The Europe Engage Tour represented an extraordinary opportunity for students, especially for those who had been involved in a variety of SL projects prior to the arrival in Galway. The diversity of political, cultural, economic, and social backgrounds of each of the participants contributed to a broad range of topics being discussed throughout the two weeks of the tour. These included conversations concerning democratic values in the various EU member states, the rise of right-wing politics in the EU, Brexit, and the global refugee crisis, amongst others. Concepts of identity were also explored, and how each democratic state within the EU is increasingly exposed to capitalist, free-market policies which have contributed to the much global instability experienced in recent years.

Involvement with the community partners truly highlighted the need for universities and students to take the lead in promoting service-learning as a valid form of pedagogical accreditation. The community partners were from a wide range of backgrounds but they each shared one common bond, and it was that they each based their practices on solid, academic research, while placing the student and, by extension, the university, in the heart of the community in a meaningful context once again.

Acknowledgements

This paper was written under the framework of the project *Europe Engage – Developing a Culture of Civic Engagement through Service-Learning within Higher Education in Europe* [Reference 2014-1-ES01-KA203-004798]. We wish to acknowledge all 12 partner countries that contributed to the work in this project, especially the partners from the National University of Ireland, Galway, who organized the international service-learning multicultural experience within community for all visiting students. We also wish to thank all participants in the video; including volunteers Damilola Arosomade, Sarah Carroll, & Ben Nolan, and school children Antoine & Batiste Grenon, and Lily Sage. Finally, we wish to thank Dr. Muriel Grenon and Dr. Claudia Fracchiolla, for without their guidance and support, we might not have produced such a video.

References

[1] Bringle, R. G., & Hatcher, J.A. (1995). A service-learning curriculum for faculty. *Michigan Journal of Community Service-learning, 2*, 112-122.

[2] Bringle, R. G., & Hatcher, J.A. (1996). Implementing Service Learning in Higher Education. *Journal of Higher Education, 67(2),* 221-239.

[3] Cole, M. (2006). *The 5th Dimension: An After-school Program Built on Diversity*. Russell Sage Foundation.

[4] Europe Engage project website: https://europeengage.org/ [accessed 2017-09-25].

[5] Eyler, J., Giles, D. E. Jr., Stenson, C. M., & Gray, C. J. (2001). At A Glance: What We Know about The Effects of Service-Learning on College Students, Faculty, Institutions and Communities, 1993-2000: Third Edition. *Higher Education*. 139. http://digitalcommons.unomaha.edu/slcehighered/139

[6] Mikelić Preradović, N. (2015). Service-Learning. In Peters, M. (Ed.), *Encyclopedia of Educational Philosophy and Theory* (1-6). Springer Singapore: Springer.

[7] Mikelić Preradović, N., & Jandrić, P. (2016). Using video journals in academic service-learning. *Polytechnic and design, 4(4)*. doi:10.19279/TVZ.PD.2016-4-4-06

[8] Robinson, L., & Kelley, B. (2007). Developing reflective thought in preservice educators: Utilizing role-plays and digital video. *Journal of Special Education Technology, 22(2),* 31-43.

[9] Stevens, D. D., & Cooper, J. E. (2009). *Journal keeping: How to use reflective writing for effective learning, teaching, professional insight, and positive change*. Sterling, VA: Stylus Pub.

[10] Waldner, L. S., McGorry, S. Y., & Widener, C. (2012). E-service-learning: The evolution of service-learning to engage a growing online student population. *Journal of Higher Education Outreach and Engagement, 16(2),* 123-151.

The Influence of Ad Blockers on the Online Advertising Industry

Tomislav Ivanjko
Department of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb
Ivana Lučića 3, Zagreb, Croatia
tivanjko@ffzg.hr

Tanja Bezjak
AdCumulus
bezjaktanja@gmail.com

Summary

Online advertising developed in 1994 as a means of financing website content production, but a constantly decreasing number of ad impressions and a significant rise in ad blocker users have incurred the loss of 30% of total industry revenue in 2015. However, studies reveal that two-thirds of current ad blocker users would to be willing to turn their blockers off should overall user experience be improved. As marketers continue debating whether the ad block revolution will finish online advertising, this paper presents an alternate viewpoint. It suggests that the rise of the ad block has been beneficial to the industry, because exposing the weaknesses of the current advertising model and the reasons for blocker popularization prompts the damaged model to change. It gathers best practices in advertising and recommendations for creating ads which do not need to be blocked.

Key words: online advertising, ad blocking, user experience, banner, ad

Introduction

Value exchange exists in every community, primal and modern ones alike. Whether it is trading berries for a piece of meat or currency for products and services, it is implied that both sides should obtain value. How did we then begin to feel entitled to go online and get content without any duty to compensate? There is a general expectation that online content should be "free"; at the same time, there is no general understanding that *online advertising* is what makes it possible (IAB, 2016). Since 1994, websites have sold their media space to finance content production. The online advertising chain is profitable for all its participants; *publishers* maintain "free" access for users, *advertisers* promote products and services to obtain new customers, *ad networks* gather publishers and advertisers, and website *visitors* use ad impressions as the cur-

rency for accessing content (Davar, 2013). Evans (2009) suggests that advertising online is potentially the most advanced form of marketing because the use of technology facilitates the meeting of supply and demand. Marketers target relevant marketing messages to objective audiences and direct them towards precise offers, and buyers do not waste time looking at ads that do not interest them.

However, many online users have started using ad blocking software, which prevents the ad from loading by cutting off communication between the ad server delivering ads and the user's computer (Bubna, 2013). It uses predefined *blacklists* to determine the ad types, publishers, and websites that should be blocked. The website's appearance does not change in the process and there are no blanks left in the content (IAB UK, 2016).

This paper examines the current online advertising model, which provoked the development of the ad blocking software. It explores the main areas of user frustration with online advertising, examines industry standards and trends, and proposes a redefinition of the online advertising model by suggesting the characteristics of ads which would not need to be blocked.

Background research on online advertising

Such is the importance ad blocking has gained in the last couple of years that it was identified as a global Internet trend in the 2016 Kleiner Perkins Internet Trends Report (PageFair, 2017). A literature review reveals a number of articles contemplating the reasons which have caused the discontent of online users, resulting in 236 million desktop and 380 million mobile active ad blocker users globally, as reported by the Global Adblock Report (PageFair, 2017). To understand what motivates the constantly increasing use of ad blockers and demonstrate that things have not always been this way, the paper will look at the online advertising industry's development.

The online ad which started it all at the end of 20th century was a banner by telecom company AT&T. The banner was a part of an integrated 2-year TV broadcast campaign (McCambley, 2013), published on 27th October 1994 on *HotWired*. Turning 23 this year, AT&T's ad was the standard banner size of 468x60 pixels (Barker, 2013). The ad copy read "Have you ever clicked here? You will", and an ad click redirected visitors to a landing page containing relevant, entertaining and informative content, and asking users for feedback and improvement suggestions. What started as a means of financing content production (Greenfield, 2014) has since transformed into a 60 billion dollars heavy industry in 2015 alone (PwC, 2015). In the novelty of the World Wide Web and its 14 million most progressive users, 10 thousand of which were *HotWired* readers (Greenfield, 2014), it is no wonder the first banner ad was an absolute success. Compared to today's average click-through rate of just 0.08%, AT&T's ad got a whopping 44% of website visitors' clicks (Barker, 2013). The ad's creative team claims that they aimed to reward the visitor for the click (Greenfiled,

2014), offer useful information, entertain, impress and provide an enjoyable user experience (McCambley, 2013).
And while the value of ads still lies in the sum of clicks and impressions, the number of views and interactions ads get today has decreased significantly since the banner's golden times. There were 3.7 billion Internet users in March 2017 (Internet World Stats, 2017); just as the number of Internet users grew, the number of websites in the online universe increased, too; more and more advertisers were keen to reach out to potential customers through online advertising. The novelty wore off and a coping mechanism for distracting advertising called *banner blindness*, which entails ignoring the margins of pages where ads are located, kicked in (Shopify, 2016). And whereas ignoring ads is a choice of an individual, using ad blockers implies employing technology which enables the destruction of a business model (Vallade, 2009).

Ad blocker development

When ignoring ads was no longer enough, ad blocking software emerged as the weapon for ad elimination, costing the online advertising industry 30% of its annual revenue in 2015 (PageFair & Adobe, 2015). The founder of the ad blocking software was IT student Henrik Aasted Sørensen from Copenhagen, who built the browser extension AdBlock in 2002. Today, there are more than 100 million users of the most popular ad blocking solution, Eyeo's Adblock Plus (Williams, 2016), and the number is constantly increasing. Reportedly, there were 77 million active desktop ad blocker users in Europe in 2016 (PageFair & Adobe, 2015) and another 50 million users in the USA (PageFair, 2017).
Although ad blocking stigma tends to victimize only the end users, publishers and advertisers actually bear more significant damage (Davar, 2013). Some portals like Forbes (Morrissey, 2015) have fought back, introducing *paywalls* to request financial compensation from ad blocker users who refuse to turn it off in order to access content (IAB, 2016a); other websites have implemented software which blocks the ad blocker's functioning (IAB UK, 2016), switching the blocked ad with an alternative one. When IAB realized outsmarting one another was not sustainable in the long term, they did some fact-checking to tackle the problem systematically and discover who blocks ads, why and how to win them back.
The goal of Interactive Advertising Bureau's (2016b) study "Who blocks ads, why and how to win them back" was to determine the demography of ad blocker users, detect the biggest problems with online advertising which prompt ad blocker use, and propose solutions for convincing users to abandon ad blocking software. Out of 1292 respondents, 330 were existing blocker users, 478 had never used it nor intended to start, 260 were former users and 224 expressed affinity to starting to use an ad blocker. The study found men between the ages of 18 do 34 to be both the most common users and the group most likely to stop using it; PageFairs's 2017 Global Adblock Report confirmed it.

IAB also dedicated a small portion of their research to mobile users, but as 94% of global mobile usage of ad blockers occurs in the Asia-Pacific region (PageFair, 2017), and the paper focused mainly on the European and Northern American markets, it was not taken into consideration.
The most important conclusion IAB's study found was *hope*. Namely, Internet users do not hate all kinds of online advertising (IAB, 2016b). This is backed up by research conducted by Adblock Plus (2016), which revealed that only 25% of current ad blocker users want *all* ads eliminated, while the majority is willing to support content production if they are granted satisfactory user experience. Furthermore, US Desktop Adblock Survey Data (PageFair, 2016) confirmed that users do not reject digital advertising in general, but rather disagree with some aspects of its execution.

Main areas of frustration and recommendations

PageFair's 2017 Global Adblock Report (PageFair, 2017) revealed more than 70% of online users cannot isolate only one reason for using ad blocking software. Since a number of sources referenced in this article reveal similar problems with online advertising, they were grouped into seven main categories: privacy anxiety, malwertising, user experience, ad format, creativity, relevance and optimization.
Privacy anxiety is the most pronounced objection users hold against online advertising, and a great deal of its notoriousness can be attributed to personal data malpractices (PageFair, 2017). As the industry thrives on user data becoming ever more sophisticated, users and their actions are being tracked and stalked (Evans, 2009) using various technologies which create fear and discomfort (Searls, 2015a). Bits and pieces of data are collected, processed, and mushed together into *Big Data* with the goal to respond to users' needs accurately. As noble as that cause may sound, data misuse is common, and websites often collect valuable personal data of their visitors in order to sell to third parties (Searls, 2015a), organizations and companies which need user data. Evans (2009) believes educating users about data collecting activity could help minimize their anxiety, because information shortage intensifies it. Besides having data privacy concerns, online users are worried about installing viruses and other malware on their devices by clicking on ads (Bubna, 2013). Users feel safer browsing online when they are protected from ads which might transmit harmful software without approval, and they do not tolerate ads which contain fake exit or download buttons, and pressure users to click on them before proceeding to the wanted page (Davar, 2013).
Consider a typical Internet usage scenario. You turn on the computer either to socialize, browse, look something up or enjoy different content (Shopify, 2016). As you are focusing on your activity, all of a sudden an ad pops up, covering the content and destroying your concentration and flow. Distracted and annoyed, you immediately close the ad window and perceive advertising as a neg-

ative and intrusive occurrence (Brajnik and Gabrielli, 2010). Furthermore, if the same ad keeps popping up over and over again, it is really not that difficult to understand the motivation behind installing ad blockers. The latter can be prevented with *frequency capping* (PageFair, 2017), which is setting the maximal number of ad occurrences.

Redundancy intensifies irritation, especially with the most hated ad formats. These are *interstitial* ads which take over the whole page, *overlay* ads which cover the page content, *pop-ups* and *pop-unders* which appear out of nowhere, *pre-roll* videos which cause video abandonment (Vallade, 2009), ads heavy on visual effects which distract the user from viewing content, *scroll-down* ads which follow you as you scroll down the page and *rich-media* ads which consume huge portions of your data plan and slow down the page load time, blinking and playing audio and video automatically (IAB, 2016b) (IAB UK, 2016). Therefore, IAB (2016) suggests ads should be limited in size and dimensions, have a fixed position on the page, and only load once the web visitor scrolls to the part of the page where the ad is located.

Getting the user to view the ad is the easy part; getting them to click on it is significantly more demanding. If you would not click an ad, why would your audience? It seems advertisers have forgotten an important lesson AT&T's creative team tried to teach; offer useful information, entertain, impress and give online users an enjoyable experience (McCambley, 2013). A lot more thought should be given to engaging the users, and following *the three Is* recipe might be useful; make your ad *inviting*, *interesting* and *innovative* and top it off with valuable content (Shopify, 2016). Ads which have irrelevant content (Brajnik and Gabrielli, 2010) and lack connection to users' interests are advertising money poured down the drain. Some online users reacted positively to the possibility to define the fields of interest for which they wish to receive ads. That way, the user gets precisely targeted ads, advertisers promote directly to their target customers and publishers get ad interactions. Furthermore, users should be given the possibility to provide feedback for the ads they see, rating them positively or negatively, and refining their preferences even further. Finally, all ads should be continually tested, audience response tracked and measured, and analytic tools used to monitor key performance indicators and optimize if necessary (Shopify, 2016).

Standards which propose solutions to the blocked Web

Several notable proposals and initiatives by respectable industry institutions to get online advertising malpractices under control stand out, with the focus on introducing methodical changes and standardizing the online advertising model. Such is the *LEAN* ads standard, which was first presented to IAB's "Who blocks ads" study respondents as one solution against bad advertising and ad blocking. It was applauded and praised by the majority of respondents, who said they would turn off their ad blockers if online ads fulfilled the four main pre-

requisites implied in *LEAN*. According to the standard, ads should be *light* in order to load fast and not consume too much data, *encrypted* to secure user data and protect users from privacy breaches, *Ad Choice* supported to let visitors control information regarding their interests, and *non-invasive* in their appearance, not do disrupt user experience or be distractive (IAB, 2016b).
And while *LEAN* caters to end users, IAB's *DEAL* standard appeals to the other important players in the game; the publishers. It promotes initiating dialogue with the user who visits a website with an active ad blocker. In such event, the standard suggests undertaking the following steps. When ad blocker usage is detected, the visitor should be educated about the "value exchange", namely informed that advertising finances the website's content production. Then, the visitor should kindly be asked to disable the software, temporarily or permanently, in order to access the content they wish to view. Finally, the publisher should either limit access to content or let the visitor view it, depending on the visitor's decision to disable the software. However, PageFair's 2017 report has deemed this approach ineffective, stating that 74% of users who encounter an *adblock wall* simply abandon the website which requires the blocking software to be disabled, with the exception of content which cannot be found anywhere else (PageFair, 2017).
Acceptable Ads initiative is the most controversial solution proposal, as it was introduced in 2011 by the most popular ad blocking software company itself (Solon, 2016). Adblock Plus started the initiative as an attempt to bring the online advertising industry back into balance. Aware that not all ads are bad, they gave online users the possibility to only block ads which threatened pleasant user experience, and otherwise support advertisers whose ads were deemed acceptable by a proposed set of guidelines, pertaining mainly to ad content, positioning, and dimensions (Adblock Plus, 2016). The fact that an ad blocking company itself attempted to find a way to unblock the Web speaks about the importance of the role which online advertising plays in the creation and consumption of the content available online.

Trends and the future of online advertising in the context of ad blocking

As notorious as online advertising has become nowadays and despite the fact that, statistically, a mere 8% of total Internet users account for 85% of all banner clicks (Morrissey, 2013), recent data revealed that the static banner ad is still the preferred ad format among users (PageFair, 2017). Its click-through rate recorded the lowest point in 2008, after which it stabilized around 0.04% (Morrissey, 2013).
However, ad serving technology has advanced, offering innovative possibilities for segmenting audiences and targeting users precisely (Oberoi, 2013). As the technology evolves further, much attention is given to revenue maximization

and not enough to content optimization. *Programmatic advertising*, which is the automated buying and selling of advertising space, is quickly gaining popularity and threatening to become the weapon to direct existing problematic ads to larger audiences (McCambley, 2013). On the other hand, MacDonald (2015) insists that the best way to get consumers to voluntarily engage with advertisements is not to make ads at all.

Instead, she suggests attracting visitors' attention with useful content, which is exactly what advertisers have been doing on Facebook. Although its primary purpose was socializing, the fact that 20% of world population created profiles has turned Facebook into an advertising super machine with approval to access end users' personal data, interests, and affinities (Oberoi, 2013). Its powerful tools are impossible-to-block *in-feed* ads and a very comprehensive analytics and efficiency reporting system, allowing constant campaign optimization for maximum performance. After all, social platforms are where we spend our free time nowadays, and relevant ads for products and services we have expressed our preference for might be the right direction to go.

Due to the fact that many users ignore traditional ad formats but focus on content, Ming & Yazdanifard (2014) praise native ads, which mimic the appearance and style of the page they are located within. However, their intention is not to deceive online users; although they look like editorial pieces, they are clearly indicated as ads and may only contain one-third of advertising content. Among preferred ad formats, native ads have been graded *neutral* by online users (PageFair, 2017).

The new advertising model

Ad blocking poses a serious threat to destroying the proper functioning of the advertising-financed value exchange powering "free Internet", i.e. users' privilege to access online content for free. This paper gathers the main areas of user frustration, the standards proposed by industry influencers, and the marketing strategies trending on the advertising market, summarizes them into a proposal for the redefinition of the online advertising model and gives recommendations for creating ads which do not need to be blocked.

Table 1: Key aspects of the redefined model of online advertising

Current model	Redefined model
Privacy anxiety	Data safety
Revenue maximization	Content optimization
Unpredictable ad formats	User-approved non-invasive ads
Interruption and redundancy	Enhanced flow and frequency capping
Banner blindness	Native advertising

As most respondents of IAB's (2016) and Page Fair's (2017) research expressed, data safety is the primary concern regarding online advertising. It should be secured with encrypted ads, ensuring no malware gets in the way and

no private information leaks. This has already been attempted with filters such as *Acceptable Ads*, but marketers should also consider users' preferences and enable their feedback in the future. When users have more control, there is less room for worrying, doubting and needing defence. Furthermore, instead of refining the ad serving technology to achieve revenue maximization, marketers should focus more on what they are serving. The answer to optimizing the quality of ads might lie in the three approaches which have been trending on the online advertising scene recently: social networks as the platform, *content marketing* as the strategy (MacDonald, 2015), and *native advertising* as the format (Ming and Yazdanifard, 2014). The consequences of *banner blindness* should not be fought with the use of irritable and unpredictable ad formats. Marketers should abandon pop-ups and replace them with non-invasive, user-approved static banners or native ad formats, with limited frequency capping to prevent overwhelming visitors. Finally, users' clicks should be attracted with quality, useful and relevant content (MacDonald, 2015) and original design.

Conclusion

Marketers have overwhelmed online users, who have in turn responded by eliminating ads by using ad blocking technology. Since most of the content production available online is financed with advertising revenue, the right to "free Internet", which most World Wide Web users take for granted, is at stake. With the goal of introducing methodical changes to repair the damaged online advertising model, IAB carried out a study which revealed two-thirds of online users are willing to abandon the blocker if user experience improves.

Based on IAB's study, a number of industry-relevant sources and advertising best practices, this article provides recommendations for the redefinition of the online advertising model. If the focus is placed on data safety, uninterrupted user experience and quality creative content in the future, there is a promise of an improved sustainable solution which will allow the continuation of the value exchange powering the Web as we know it.

References

Adblock Plus. Allowing Acceptable Ads in Adblock Plus. https://adblockplus.org/acceptable-ads (2017/04/10)

Barker, Dan. The First Ever Banner Ad (& How it Performs Today). 2013/10/30. http://barker.co.uk/banner (2017/04/13)

Brajnik, Giorgio; Gabrielli, Silvia. A Review of Online Advertising Effects on the User Experience. // *Journal of Human-Computer Interaction*. 26 (2010), 971-997.

Bubna, Josiah. The Ethics of Adblock. 2013/05/06. http://bubnaphotography.com/josiah/writing/TheEthicsofAdblock.pdf (2017/04/16)

Evans, David S. The Online Advertising Industry: Economics, Evolution, and Privacy. // *Journal of Economic Perspectives*. 23 (2009), 3; 37-60.

Greenfield, Rebecca. The Trailblaizing, Candy-Colored History of The Online Banner Ad. 2014/10/27. https://www.fastcompany.com/3037484/most-creative-people/the-trailblazing-candy-colored-history-of-the-online-banner-ad (2017/04/10)

IAB. Publishers Making DEALs to Battle Ad Blocking in the U.S. and in Europe. 2016/04/25. https://www.iab.com/news/publishers-making-deals-to-battle-ad-blocking/ (2017/03/23)
IAB. Ad Blocking: Who Blocks Ads, Why and How to Win Them Back. June 2016. http://www.iab.com/wp-content/uploads/2016/07/IAB-Ad-Blocking-2016-Who-Blocks-Ads-Why-and-How-to-Win-Them-Back.pdf (2017/04/16)
Internet Advertising Bureau UK. AdBlocking FAQs 2016. March 2016. https://www.iabuk.net/sites/default/files/Ad%20blocking%20FAQ%20March%202016.pdf (2017/04/29)
Internet World Stats. Internet Usage Statistics: World Internet Users and 2017 Population Stats. 2017/03/31. http://www.internetworldstats.com/stats.htm (2017/04/16)
MacDonald, Muriel. Better than Banner Ads: Smart Ways to Spend Your Ad Dollars in 2015. 2015/01/12. http://www.tintup.com/blog/better-than-banner-ads-smart-ways-spend-ad-dollars-2015-muriel-macdonald/ (2017/04/29)
McCambley, Joe. The First Ever Banner Ad: Why Did It Work So Well? // *The Guardian, [online]*. 2013/12/12. https://www.theguardian.com/media-network/media-network-blog/2013/dec/12/first-ever-banner-ad-advertising (2017/04/10)
Ming, Wong Qi; Yazdanifard, Rashad. Native Advertising and its Effects on Online Advertising. // *Global Journal of Human-Social Science (E)*. 14 (2014), 8; 11-14.
Morrissey, Brian. 15 Alarming Stats About Banner Ads. 2013/03/21. https://digiday.com/publishers/15-alarming-stats-about-banner-ads/ (2017/04/16)
Morrissey, Brian. Forbes Starts Blocking Ad-Block Users. 2015/12/21. http://digiday.com/publishers/forbes-ad-blocking/ (2017/04/29)
Oberoi, Ankit. The History of Online Advertising. 2013/07/03. http://adpushup.com/blog/the-history-of-online-advertising/ (2017/04/16)
PageFair and Adobe. The Cost of Ad Blocking: Ad Blocking Report. 2015. https://downloads.pagefair.com/wp-content/uploads/2016/05/2015_report-the_cost_of_ad_blocking.pdf (2017/04/10)
PageFair. Adblocking Goes Mobile: 2016 Mobile Adblocking Report. November 2016. https://pagefair.com/wp-content/uploads/2016/05/Adblocking-Goes-Mobile.pdf (2017/03/23)
PageFair. The State of the Blocked Web: 2017 Global Adblock Report. 2017/02/01. https://pagefair.com/downloads/2017/01/PageFair-2017-Adblock-Report.pdf (2017/04/16)
PwC. IAB Internet Advertising Revenue Report: 2015 Full Year Results. April 2016. https://www.iab.com/wp-content/uploads/2016/04/IAB-Internet-Advertising-Revenue-Report-FY-2015.pdf (2017/03/23)
Searls, Doc. How Will the Big Data Craze Play Out? // *Linux Journal*, *[online]*. 2015/11/04. http://www.linuxjournal.com/content/how-will-big-data-craze-play-out (2017/04/29)
Sharetrough. Native Ads vs Display Ads. 2013. http://www.sharethrough.com/resources/native-ads-vs-display-ads/ (2017/03/10)
Shopify. 50 Ways to Make Your First Sale: Buy Banner Ads (Chapter 27). 2016. https://www.shopify.com/guides/make-your-first-ecommerce-sale/banner-ads (2017/04/13)
Solon, Olivia. Adblock Plus Launching Platform to Sell 'Acceptable' Ads. // *The Guardian, [online]*. 2016/09/13. https://www.theguardian.com/business/2016/sep/13/adblock-plus-launching-platform-to-sell-acceptable-ads (2017/04/16)
Target Marketing. Consumers Prefer Targeted, Relevant Online Advertising in Exchange for Trade-Offs. 2010/07/08. http://www.targetmarketingmag.com/article/consumers-prefer-targeted-relevant-online-advertising-in-exchange-trade-offs/ (2017/03/10)
Vallade, Jilian. Adblock Plus and the Legal Implications of Online Commercial-Skipping. // *Rutgers Law Review*. 61 (2009), 3; 823-853.
Williams, Ben. Adblock Plus and (a Little) More: 100 Million Users, 100 Million Thank-yous. 2016/05/09. https://adblockplus.org/blog/100-million-users-100-million-thank-yous (2017/04/13)

INTEGRATION OF ICT IN EDUCATION

Original scientific paper

ICT in Higher Education: Teachers' Experiences, Implementation and Adaptations

Andrea Miljko
Faculty of Humanities and Social Sciences, University of Mostar
Matice hrvatske b.b., Mostar, Bosnia and Herzegovina
andreaivankovic@gmail.com

Mateo Jurčić
Faculty of Humanities and Social Sciences, University of Mostar
Matice hrvatske b.b., Mostar, Bosnia and Herzegovina
mateo.jurcic@gmail.com

Tonćo Marušić
Faculty of Science and Education, University of Mostar
Matice hrvatske b.b., Mostar, Bosnia and Herzegovina
tonco.marusic@gmail.com

Summary

Every aspect of human life is evolving so fast and with great quality. Nowadays, education has become more complex due to the enormous social change and a new view in the field of pedagogy. We live in the age of information and communication technology (ICT), and we cannot allow to ignore it in the education business management. The use of ICTs has great potential in teachers' preparation to deal with various challenges and responsibilities they have to fulfill in their educational environment. Thanks to fast-growing development of ICT in past decades, the topic of ICT integration in education is being intensively discussed at different levels. Considering this, teachers' ICT competences should play an important role, but the approach to ICT competences assessment – in general as in cases of teachers' profession – rarely exists. The presented article focuses on teachers' experiences, implementation and adaptations of ICT in higher educations. The research was conducted on the Faculty of Humanities and Social Sciences at the University of Mostar where it was determined the frequency of ICT use and the subjective evaluation of teachers' ICT understanding and their level of ICT literacy. Descriptive parameters of test variables as well as frequencies and percentages are demonstrated in the statistical data analysis.

Key words: ICT, teacher, higher education, ICT literacy

Introduction

In the time of information and communication technologies the computer takes an important educational and scientific role. The influence of computers can be positive and negative, depending on the needs and ways of using the computer. In both cases, those impacts are rarely simple and direct and usually affected by many social and other factors.

In the past, education was highly elitist, but the era of information and communication technologies and the flexibility of using these technologies have enabled the participation of many students in the educational process. The technology education standards are extremely important for improving the quality of the educational process.

One of the basic requirements for education in this era of information explosion is to prepare learners for participation in a networked information society. This basic requirements can be available only when teachers are very aware of ICT [2].

Theoretical background

Teachers who wish to update and upgrade their teaching and learning designs using new learning technologies have some difficult issues to confront. Whether they work in schools, colleges, or universities, the incorporation of new technologies into their teaching requires them to acquire a very different approach to teaching and learning [9].

We live in the age of information and communication technology, and we cannot allow to ignore it in management of education business. The use of ICTs has great potential in teachers' preparation to deal with various challenges and responsibilities they have to fulfill in their educational environment. Therefore, many recent research studies on this subject show that many institutions do not manage to integrate technologies into the existing context. The teaching staff, although they have enough skills and they are innovative and they easily overcome obstacles, have not integrated technology as the means of learning and teaching [4]. Furthermore, [4] highlights the continuous problems in getting the teaching staff to acquire ICT and the need for further researches about how ICT can improve education.

Successful implementation of technology in education requires teacher's support and a positive attitude to a great extent. If the teachers feel that the ICT integration in teaching learning is not fulfilling their and their student's demands, then they will be somehow reluctant in using technology in teaching [7].

The main obstacle for using ICT among teaching staff, stated in [3], can be the following:

- The lack of self-confidence;
- The lack of competency;
- Resistance to change and negative attitudes.

Nowadays teachers' ICT competence is considered to be a part of their professional competence, which is not a strictly defined area (i.e. containing e.g. only technical knowledge and skills related to the use of ICT in education), but an area which is coherent and consequent with other areas of teacher's professional competence (subject, pedagogical, didactic and psycho-didactic, diagnostic and intervention and others). The core of ICT lies in the interconnection of ICT with teacher's educational activity [8].

There are many complex factors that determine how university teachers employ ICT to change their teaching practices and/or the learning practices of their students. Evidence from studies into how ICT can enhance or transform educational processes states only one influence upon teachers [6]. Some others, often more pervasive, include:

- Individual differences in teachers' attitudes to the adoption of innovations;
- Individual differences in teachers' conceptions of and approaches to teaching;
- The established departmental / faculty / institutional ethos and ways of working; and
- Competing demands of discipline-based research and administration [6].

ICT presents an entirely new learning environment for students, requiring a different skill set to be successful. ICT is changing teaching and learning processes by adding elements of vitality to learning environments including virtual environments for the purpose. ICT is a potentially powerful tool for offering educational opportunities. It is difficult and maybe even impossible to imagine future learning environments that are not supported by ICT [10] in one way or another. The implementation of ICT by the teaching staff can have multiple benefits which can increase if the students are allowed to use ICT in the process of learning. Students can become more ICT literate if the teaching staff use ICT in their process of teaching. According to the analysis of literature, ICT competencies needed from teaching staff who work in a technologically enriched environment are:

- Recognition every single problem faced by students during the process of learning;
- Careful consideration of the choice regarding the use of media;
- Verification of given information;
- Development of effective search techniques and the ability to effectively do research on the computer;
- The ability to use standard software confidently and competently;
- The ability to make wise and critical decisions based on found information.

The majority of the teaching staff believes that the usage of technologies is important for the education. However, they lack the confidence and understanding

during the process of integration. Further on, they should have skills and competencies needed for designing, delivering and the assessment because „successful integration of technology requires not only the knowledge of the technology and its potential use but also the skill to plan and execute a good lesson (of which the technology is only a part). When technology usage is aligned with the instructional goal, where technology is integral to teaching, successful integration might be succeeded" [4].

The use of new technologies in education implies new teacher roles, new pedagogies and new approaches to teacher´s education. The successful integration of ICT into the classroom will depend on the ability of teachers to structure the learning environment in new ways, to merge new technology with a new pedagogy, to develop socially active classrooms, encouraging co-operative interaction, collaborative learning and group work. This requires a different set of classroom management skills [13].

Recent research on teachers' use of ICT in education shows a concern for the difficulties teachers face when trying to use ICT in their daily educational practices. One way to frame these difficulties concerns the many and varying ways in which ICT can be used [12].

The ICT knowledge and skills needed by teachers are never permanent and learning must be continual. Although new recruits to the profession will have live with a variety of ICT applications, even they need to engage in constant professional learning both about the technology and its pedagogical applications [1].

Successful integration of ICT in teaching and learning process is highly dependent on the preparation and attitudes of teachers. Undeniably that there are ICT tools available and easily accessible by teachers but even with the existence of these tools, teachers still failed to fully utilize ICT into their lessons. And, if they did utilize these tools, most employed the use of ICT in teaching language skills and even fewer studies paid attention to teachers' attitudes in using ICT in literature lessons. Teachers are one of the factors that determine the development and innovation in public education because they are the people who use ICT investment for the development of education. [11]

UNESCO's Framework emphasizes that it is not enough for teachers to have ICT competencies and be able to teach them to their students. Teachers need to be able to help the students become collaborative, problem-solving, creative learners through using ICT so they will be effective citizens and members of the workforce. [13] In order to have these competencies, the teaching stuff need to acuire a high level of ICT literacy.

ICT literacy has become an important prerequisite in the socialization and professional caree. Therefore, education as an important social factor plays a key role in ICT literacy. Basic ICT knowledge and skills contantly need to be updated in order to follow the fast development of ICT.

The purpose of this paper is to investigate teachers' experiences and views on ICT in higher education and the ICT literacy which is presented in the next chapters.

Research questions

In order to achieve the aims of the study, the following research questions were formulated:

- Do teachers use educational software for teaching purposes?
- Do teachers believe that the quality of scientific research work is increased by using computer technology?
- Do teachers believe that ICT literacy plays a key role in the educational process?
- Do teachers believe that they are ICT literate and in what measure?

Research method

The research was conducted among 65 teachers of the Faculty of Humanities and Social Sciences, University of Mostar. The research was carried out among 44 women (64.69%) and 21 men (32.31%). Among the tested sample, 22 (34.92%) belong to the scientific field of social sciences and 41 (65,08%) to humanities.[1] As for the scientific-teaching titles, 53.85% (N=35) are research assistants or senior research assistants, 32.31% (N=21) are assistant professors, 10.77% (N=7) are associate professors and 3.08% (N=2) are full professors.

The research was conducted through a questionnaire consisted of two parts: the first part of analysis refered to the frequency of computer use as well as the subjective evaluation of computer understanding by the teaching stuff. The second part was a test with 15 questions written in order to establish the level of ICT literacy.

Teachers come across different tools, softwares, terms and symbols while using ICT every day for different purposes. Therefore, the questions on the test were written to determine whether the subjects knew the things they use (e.g. do they know how the computer works, what is an operating system; what is a web browser).

The questionnaire was based on empirical insight of the teaching process conducted on the Faculty of Humanities and Social Studies of the University of Mostar.

Descriptive parameters, as well as frequencies and percentages, of researched variables were shown in the statistical data analysis.

The results of the ICT literacy test were given by a very simple linear compination of correct answers.

[1] Two subjects did not indicate their science field

Spearman's rank correlation coefficient was used to test the correlation of ICT literacy and some habits of computer use. The differences in the level of ICT literacy between the teaching stuff and students, teachers of different fields and teaching and research positions were tested by the Chi-square test and/or a non-parametric test for testing differences between two/three independent groups. The results were interpreted on the significance level of 5%.
The computer program STATISTICA (data analysis software system), version 7 was used for the statistical analysis.

Findings and discussion

A. Computer usage frequency and the understanding of computers efficacy

Table 1 shows the computer usage frequency as well as the self-evaluation of their understanding of computers efficacy.

Table 1: Computer usage frequency and the understanding of computers efficacy

No.	Question	Given answers	Results	
1.	Do you use computer technologies in your class?	Yes	f	62
			%	96.87
		No	f	2
			%	3.13
2.	If you do, circle how often you use it:	Rarely	f	3
			%	4.76
		Sometimes	f	4
			%	6.35
		Often	f	10
			%	15.87
		Very often	f	46
			%	73.01
3.	How often do you use the computer?	More than 3 hours a day	f	43
			%	66.15
		2-3 hours a day	f	14
			%	21.54
		1-2 hours a day	f	7
			%	10.77
		0-1 hours a day	f	1
			%	1.54
4.	What do you use your computer for in class?	Search the Internet	f	40
			%	61.54
		Writing and creating different documents	f	57
			%	87.69
		Communication with students	f	49
			%	75.38
		Make a presentation	f	6
			%	9.23
		Paper presentations	f	1
			%	1.54

No.	Question	Given answers	Results	
		Work in programs specified for certain college subjects	f	1
			%	1.54
		Preparing the lecture	f	2
			%	3.08
		Improving the class through quality control	f	1
			%	1.54
		Correcting papers	f	1
			%	1.54
		Translation	f	1
			%	1.54
		Reading books	f	1
			%	1.54
		Social media	f	1
			%	1.54
5.	Do you use any educational software for your class needs?	Yes	f	12
			%	19.35
		No	f	50
			%	80.65
6.	If you do, write which ones:[2]	Merlin	f	3
			%	27.27
		Moodle	f	2
			%	18.18
		Power Point	f	4
			%	36.36
		Ephorus	f	1
			%	9.09
		Pdf	f	1
			%	9,09
		VLC player	f	1
			%	9,09
		Hrčak	f	1
			%	9,09
		Busyteacher	f	1
			%	9,09
		Englishtips	f	1
			%	9,09
		thefreedictionary.com	f	1
			%	9,09
7.	Assess how well you understand computers and computer technologies:	1 (not enough)	f	1
			%	1.54
		2 (enough)	f	11
			%	16.92
		3 (well)	f	30
			%	46.15
		4 (very well)	f	17
			%	26.15

[2] Question no. 6 was an open-ended question. Their answers are represented in the table.

No.	Question	Given answers	Results	
		5 (excellent)	f	6
			%	9.23
8.	Does the use of computers imply the use of Internet?	Yes	f	5
			%	7.81
		No	f	59
			%	92.19
9.	Do you think that the knowing computers and computer technologies is highly essential in our everyday life?	Absolutely not	f	1
			%	1.56
		Partly	f	2
			%	3.13
		Moderately	f	8
			%	12.50
		Very	f	53
			%	82.81
10.	Do you think that the quality of scientific work increases with the use of computer technologies?	Absolutely not	f	3
			%	4.69
		Partly	f	1
			%	1.56
		Moderately	f	30
			%	46.87
		Very	f	30
			%	46.87
11.	Do you think that ICT literacy plays a key role in the educational process?	Absolutely not	f	1
			%	1.54
		Partly	f	1
			%	1.54
		Moderately	f	32
			%	49.23
		Very	f	31
			%	47.69
12.	Do you think that the use of computer technologies in class helps students master new units more easily?	Absolutely not	f	2
			%	3.08
		Partly	f	4
			%	6.15
		Moderately	f	47
			%	72.31
		Very	f	12
			%	18.46
13.	Do you think you are ICT literate?	Yes	f	55
			%	85.94
		No	f	9
			%	14.06

In the question “Do you use computer technologies in your class?” 96.87% of the subjects stated that they do use the computer, and only 3.13% that they do not use computer technologies. 73.01% of the teaching staff uses computers in class very often, 15.87% often, 6.35% sometimes and 4.76% rarely.

The research showed that 66.15% of the subjects used the computer more than three hours a day, 21.54% two to three hours a day, 10.77% one to two hours a day and 1.54% one hour or even less a day.
In the question "What do you use your computer for in class?" the majority of them 87.69% stated for writing and creating different documents, and the smallest percentage for paper presentations, correcting papers, work in programs specified for certain college subjects, preparing the lecture, Improving the class through quality control, translation, reading books or for social networks.[3] Further research has shown that only 19.35% of the teaching staff uses some of the educational software for class needs, while 80.65% do not use any software (Figure 1). Among those who use educational software, 27.27% named Merlin software, 36.36% Power Point, 18.18% Moodle software and 9.09% the use of Ephorus, Pdf, VLC player, Hrčak, Busyteacher, Englishtips or the freedictionary.com which they think are educational software.

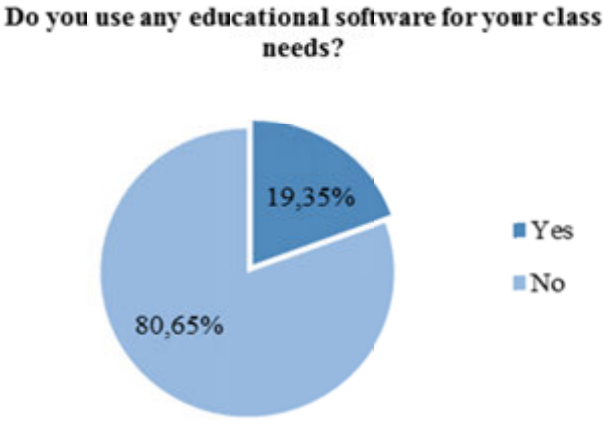


Figure 1. Question "Do you use any educational software for your class needs?"

As for self-evaluation of their understanding of computers and computers technology efficacy, the highest percentage believe that they understand well (46.15%) or very well (26.15%) computers and computer technologies software (Figure 2). Also, 92.19% of them do not think that the use of computers necessarily implies the use of Internet.
In the question "Assess how well you understand computers and computer technologies" the scientific-teaching staff both from the humanistic and social studies equally believe they understand computers and computer technologies The majority of subjects, 82.81%, believes that knowing computers and computer technologies is highly essential in our everyday life. Also, the highest

[3] The subjects were given the opportunity to give multiple answers.

number highlights that the quality of scientific work has significantly increased with the use of computer technologies.
The question "Do you think that ICT literacy plays a key role in the educational process?" shows that only 1.54% answered absolutely not or partly, 49.23% moderately, and 47.69% (Figure 3).

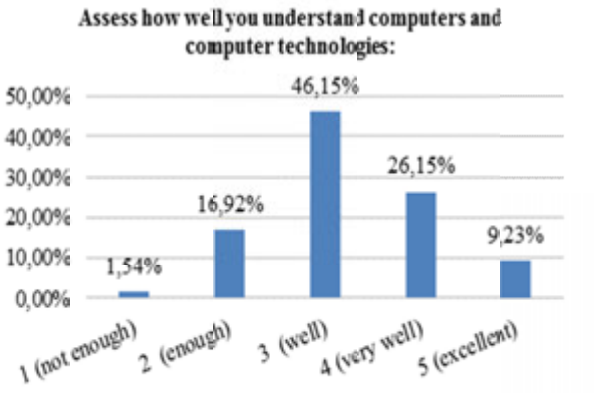


Figure 2. Question "Assess how well you understand computers and computer technologies

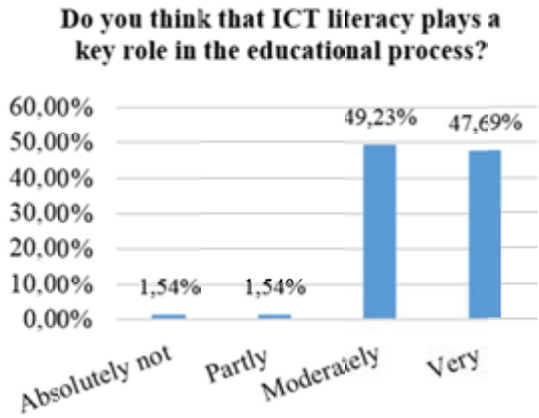


Figure 3. The question "Do you think that ICT literacy plays a key role in the educational process?"

The majority of the teaching staff thinks that the use of computer technologies in class moderately helps students master new units more easily.
The last question referred to the self-evaluation of ICT literacy in which 85.94% answered that they thought they are ICT literate and 14.06% stated contrary.

B. Results from the Knowledge test of teaching stuff´s ICT literacy

The level of teaching staff´s ICT literacy was tested by objective measurement. The questionnaire was taken from the paper "ICT Literacy among the Students of the Faculty of Philosophy, University of Mostar" in which all the same quastions were given to students. In the quoted paper students achieved average results. [5]

This research wanted to show the level of teaching staff´s ICT literacy.

According to data analysis the percentage of correct answers on the Test of ICT literacy varies from 37.10% to 98.41% (Table 2, Figure 4). The lowest percentage, only 37.10% of all the subjects, knew which operating system was produced in 2011 by Microsoft, while 98.41% knew which function on the computer reruns the system and the constituent part of the symbol @.

Table 2. Results from the Knowledge test of teaching stuff´s ICT literacy

No.	Question	Answers		
		f / %	Correct	Incorrect
1.	On which digits is the computer world built?	f	51	22
		%	82.26	17.74
2.	What of the listed represents an operating system?	f	56	7
		%	88.89	11.11
3.	What is Facebook?	f	61	2
		%	96.83	3.17
4.	What is http?	f	30	33
		%	47.62	52.38
5.	Whose founder was Steve Jobs?	f	58	5
		%	92.06	7.94
6.	Which number in the picture represents the space bar button?	f	57	4
		%	93.44	6.56
7.	Which shortcut means Cut?	f	42	18
		%	70.00	30.00
8.	What is the operating system produced by Microsoft in the 2011. year?	f	23	39
		%	37.10	62.90
9.	Which one of listed is an IP address?	f	53	8
		%	86.89	13.11
10.	Which function on the computer reruns the computer?	f	62	1
		%	98.41	1.59
11.	@ sign is an integral part of:	f	62	1
		%	98.41	1.59
12.	Which one of the listed represents a wireless network?	f	56	7
		%	88.89	11.11
13.	What does the shortcut PC mean?	f	59	4
		%	93.65	6.35
14.	What program is used for website browsing?	f	32	31
		%	50.79	49.21
15.	What file extension makes only an image file?	f	61	2
		%	96.83	3.17

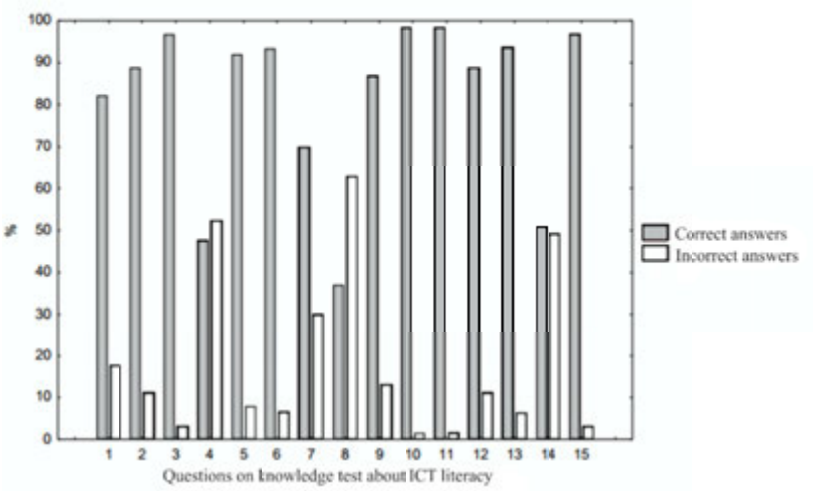


Figure 4. The presentation of results from the knowledge test of teaching stuff´s ICT literacy

The distribution of results on the Test of ICT literacy is negatively assimetrical which means that most teaching staff have accomplished extraordinary results. (Figure 5). According to the median value, there are 13.33 out of 15 correct answers. The median value of correct answers is 88.89%, and of incorrect 11.11%. The distribution of test results raises a question whether the test results are the real reflection of ICT literacy. This distribution shows that the test is insufficiently accurate: or the teaching staff is extraordinary literate or the test is too easy.

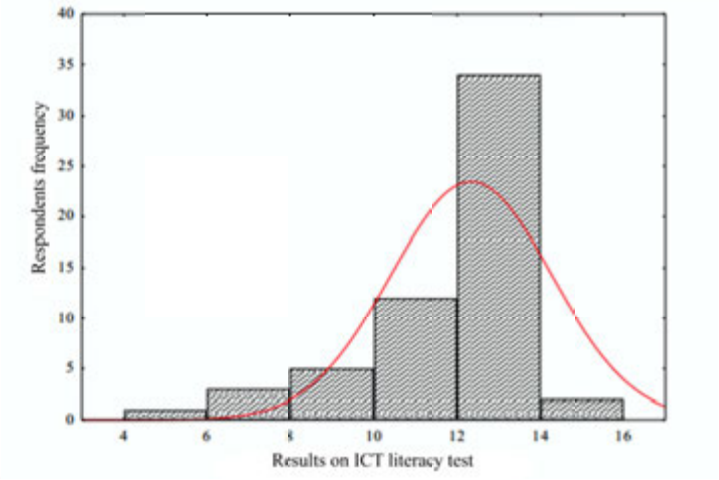


Figure 5. The distribution of results of the ICT literacy knowledge test of teaching stuff from the Faculty of Humanities and Social Studies of the University of Mostar

The data analysis showed the correlation between ICT literacy and habits of computer use, as well as the self-evaluation of computer knowledge and ICT literacy (Table 3). A significant positive statistical link was established between the subjective evaluaton of computers and the results of the ICT literacy test. The given correlation demonstrates that the subjects with better insight in computers have accomplished better results on the test. Apart the positive correlation, a negative correlation between the subjective evalution of ICT literacy and the objective level of ICT literacy has been established.
There is a tendency among some teaching stuff with higher results on the ICT literacy test to poorly assess their own knowledge and and vice versa, subjects with lower results, evaluate themselves more ICT literate. These results can be indirectly explained in the way that people with lower self-evaluation are more pron to perfectionism and therefore are more self-critical and doubtful.

Table 3. The correlation between ICT literacy and the habits od computer use and the self-evaluation of computer and ICT literacy knowledge.

Variables	ICT literacy
Daily use of computers	0.144
Knowledge self-evaluation of computers and computer technology	0.367*
Self-evaluation of ICT literacy	-0.336*

*p<0,05

Conclusion

Thanks to fast-growing development of ICT in past decades, the topic of ICT integration in education is being intensively discussed at different levels. Considering this, teachers' ICT competences should play an important role, but the approach to ICT competences assessment – in general as in cases of teachers' profession – rarely exists.
Technologies have a great potential for improving different aspects in everyday life, including education. Nowadays, technology plays a key role in the effective achievement of class aims and tasks in the educational system. Computers and computer technologies are considered to be efficient teaching tools. In an academic world students expect a positive learning atmosphere based on ICT. The teaching staff should provide ICT literacy skills in an environment like that. It could be said that ICT is an essential condition for higher education.
Our research has shown that almost all the teachers use computer technologies for their class needs. And that they mostly use the computer more than three hours a day.
The subjects mostly use computers for writing and creating different documents. However, this research has shown that the teachers do not use any educational software and that they do not even know what educational software is, because in their answers they use different programs and web pages that are not part of any educational software.
According to the self-evaluation of their understanding of computers and computers technology efficacy the highest percentage thinks that they understand

computers and computer technologies. Subjects believe that computers and computer technologies are highly essential in our everyday life, and that their use significantly increases the quality of scientific work. Furthermore, subjects also believe they are ICT literate.

The ICT literacy test has shown great results about the teaching stuff´s knowleddge, i.e. it has shown that they are ICT literate.

The ICT literacy focuses on process knowledge, critical thinking and problem solving which can be found among information in a digital environment. Altough the skills which are a part of the ICT literacy are acquired through education, they can be extended beyond the academic life to real life which enables us to function in an informational and technological enriched society. The given components are highly required in institutions based on the development of knowledge, skills, understanding attittudes and trends which ensure students a successful life and work in the 21st century.

References

[1] Albion, Peter; Knezek, Don; Adubra, Edem. (2011). *TWG3: Teacher Professional Development*. EDUsummIT 2011. Paris, France: UNESCO.

[2] Angadi, Gavisiddappa R. Teachers' Attitude towards Information and Communication Technology (ICT). // *International Journal of Education and Psychological Research (IJEPR)*. 3 (2014), 1

[3] Bingimlas, Bingimlas. Barriers to the Successful Integration of ICT in Teaching and Learning: A Review of Literature. // *Eurosia Journal of Matematics, Science and Technology Education*. 5 (2009); 235-245

[4] Gülbahar, Yasemin. ICT Usage In Higher Education: A Case Study on Preservice Teachers and Instructors. // *The Turkish Online Journal of Educational Technology – TOJET*. 7 (2008), 1

[5] Ivanković, Andrea; Špiranec, Sonja; Miljko, Dražan. ICT Literacy among Students of the Faculty of Philosophy, University of Mostar. // *Procedia - Social and Behavioral Sciences*. 93 (2013); 684-688

[6] Kirkwood, Adrian. ICT in higher education: policy perspectives // *ICT Leadership in Higher Education*. India: Hyderabad. (2013); 36-43

[7] Kler, Shikha. ICT Integration in Teaching and Learning: Empowerment of Education with Technology.// *Issues and Ideas in Education*. 2 (2014), 2; 255-271

[8] Kubrický, Jan; Částková, Pavlina. Teachers ICT Competence and Their Structure as A Means of Developing Inquiry-Based Education. // *Procedia - Social and Behavioral Sciences*. 186 (2015); 882-885

[9] Laurillard, Diana. Supporting Teacher Development of Competencies in the Use of Learning Technologies. // *ICT in teacher education: policy, open educational resources and partnership, IITE /* UNESCO. Moscow; UNESCO, 2010; 63-74

[10] Noor-Ul-Amin, Syed. An Effective use of ICT for Education and Learning by Drawing on Worldwide Knowledge, Research and Experience: ICT as a change Agent for Education. // *Scholarly Journal of Education*. 2 (2013), 4; 38-48

[11] Shah Md., Parilah; ESL Teachers' Attitudes towards Using ICT in Literature Lessons.// *International Journal of English Language Education.* 3 (2015), 1; 201-2018

[12] Talebiana, Sogol; Mohammadia, Hamid Movahed; Rezvanfar, Ahmad. Information and communication technology (ICT) in higher education: advantages, disadvantages, conveniences and limitations of applying e-learning to agricultural students in Iran. // *Procedia - Social and Behavioral Sciences*. 152 (2014); 300-305

[13] UNESCO. ICT Competency Framework for Teachers. Paris: UNESCO, 2011.

Original scientific paper

Creation and Use of Game-Based Learning Material

Josip Mihaljević
School for nurses Vrapče
Bolnička cesta 32, Zagreb, Croatia
jomix@gmail.com

Summary

Rising popularity of games and increased use of technology in schools have made game-based learning more popular in education. In this paper the advantages of game-based learning are analyzed and some research results on the effect of game-based learning are given. A pilot research on the use of educational games for motivating students was conducted on high school students and results are given. The paper also presents the author's websites which contain many custom educational games such as Tic-tac-toe, Snake, Tetris, Memory, Hangman, etc. These games were created free of charge and without any copyright restrictions using certain web services and available game codes from different websites. A number of crated games are played like traditional games they are based on but have educational content integrated into game mechanics, such as a question for every turn in the game. A number of games were created as a fun way to help students learn languages (Croatian, German, Latin) as well as the basic facts of Computer Science (recognizing parts of a computer). There are also archival games which were created in a similar way as museum games to promote archives and archivist's activities where a player can match pictures of archives to get basic information about them or solve puzzles containing the pictures of archival holdings.

Key words: education, free software, game creation, game-based learning, gamification, multimedia, motivation of students, websites

Introduction

Games are often considered as a media that combines other media such as text, sound, and picture. Their strength lies in their interactivity as players can interact with the previously mentioned media. Thus, it may seem that games are perfect for studying but the problem is that games aim at having fun so they are mostly used in entertainment industry rather than education. However, that does not mean that there is no place for games in education. Traditional teaching methods are slowly being modernized and adapted to specific needs of the information driven society primarily through various attempts to integrate computers and multimedia in the everyday classroom (Mitrović & Seljan 2008:

248). Due to their interactivity, games can, even more than movies, be used as a way of motivating students for certain subjects. Educational games can improve the outcome of learning if designed and used correctly in class (Gros 2007: 31-36). As console and PC games are too expensive to produce and distribute, the focus should be on smaller mobile or web games. Web games have evolved and now with the new HTML5 language for their creation, they can be played on devices with different screen sizes and different web browsers without installing additional software (Gasston 2013: 21). They are also much easier to make with various types of free technologies and codes available on the Internet and they can be distributed through hyperlinks.

Difference between Gamification and Game-Based Learning

The idea of gamification is to add game-elements to a non-game situation. Corporate reward programs are a good example of gamification. They reward users with badges and points for certain behavior they favor. These points can later be used to get items from the store for free or for a lower price. This is very similar to RPG (Role-playing game) where a player gets points during gameplay which he can spend on getting game items (Trybus 2009). In the classroom, gamification has been integrated in a more authentic manner as some classrooms have become a living, breathing game. The example of gamification in classroom is when the teacher displays students' points and rankings on the board. In this way, the students can see their progress and compare it with the progress of other students and be more motivated to improve themselves. Another example of gamification can be seen with certain plugins for Moodle which give each profile badges and achievements for passing exams or certain subjects (Hall 2015).

Game-based learning, as compared to gamification, actually uses video games as a medium for learning. There are games that are specifically designed for learning, like *My Coach* series of video games for Nintendo DS console which are used for learning foreign (e.g. French and Chinese) languages by solving different puzzles (Spencer 2007). Although games as Assassin's Creed and Total War Series were not designed for educational purposes they can also be used for learning history because they realistically depict the time period in which the story of the game takes place (Higgins 2017). Game-based learning is not so often used as gamification in a classroom situation and cannot replace traditional learning. However, games can be used as a good medium for motivating students for certain subjects and games with questions and assignments can be used for repeating what the students have learned. This is why there should be more games made for educational purposes especially now when technology has become more accessible, much easier to use and there are many open source free programs and software which can be used for efficient game creation (Ally & Samaka 2013: 1-2).

It has been proved that students generally enjoy working with computers, and this fact should be used to motivate them during learning process (Dovedan, Seljan & Vučković 2002: 73). Numerous studies support the view that games have a positive effect on learning. A growing number of researchers are committed to developing educational games for promoting student skills. One of the popular game learning platforms well known in Croatia is Kahoot, in which teachers can create quizzes for their students which they can access with their passwords. However, little is as yet known of how games can influence student acquisition of learning and innovation skills. Meihua Qian and Karen R. Clark (2016) conducted a research using the Academic Search Complete Database[1]. They analyzed 137 studies on game-based learning. Their research shows that the influence of games on the success of learning is mostly dependent on the way they combine educational content with certain game mechanics which are successful in the entertainment game industry as well. They think that specifically designed games for learning different subjects and aiming at different groups defined by age and gender mostly work better than typical commercial and educational games. The problem is how to develop specific games for specific groups of students because that requires programming and finding or creating appropriate graphical resources. Another problem that Clark and Qian stressed in their research is that there are some indications that there is as yet not enough evidence that game-based learning improves the learning process. While the majority of the results were statistically significant, the practical significance of many of the empirical findings remains unknown.

Other researchers think that games may stimulate information acquisition, enhance the ability to think quickly and analyze different situations as well as help with aspects of coordination and concentration on visual details (Anderson & Lawton 2009; Garris, Ahlers & Driskell 2002; Wilson et al. 2009). However, there are also disadvantages such as children being overstimulated by games (Barlett et al. 2009; Thomas, Gentile & Anderson 2008). Games may also be irrelevant to the content of the subject taught or linked to excessive or addictive play (McFarlane et al. 2002; Prensky 2001). Unfortunately, many researchers analyze only negative effects of violent videogames but really focus on the positive aspect that games can have on developing psychomotoric skills or how game mechanics can be used for learning purposes (Bilić, Gjukić & Kirinić 2010: 200-203).

[1] A comprehensive full-text database for multidisciplinary research designed for academic institutions. This database is a leading resource for scholarly research. It supports high-level research in the key areas of academic study by providing journals, periodicals, reports, books and more (EBSCO 2012).

Creation of Educational Websites

Three websites for different game contents where created: Language Games and Multimedia Display of Language, Computer Games and Other Multimedia Content, and Archival Games.

Language Games and Multimedia Display of Language

The undeniable advancements in technologies used for educational purposes have, in many ways, improved the traditional language teaching methods. Teachers can more easily find teaching materials and have access to various online multimedia sources which make it possible to improve the, sometimes dull, course materials (Mitrović & Seljan 2008: 251).

The website Language Games and Multimedia Display of Language[2] was created for the purpose of presenting and distributing language games for different languages. Language games are seen as a fun way of learning different languages and learning facts about language and literature. The idea to create this website came after the games for learning Glagolitic letters were made for The Old Church Slavonic Institute[3] and memory and spelling games were made for the Institute of Croatian Language and Linguistics[4]. The first web domain for the site was http://www.jezicneigre.tk which was hosted on free web hosting service 000webhost. After that, the site was transferred to a new web host and got a new domain http://www.jezicneigre.com. The old .tk domain is still active but it automatically redirects visitors to the .com site. The websites were created by using Wordpress CMS system. This site not only includes typical language games, e.g. quizzes that most language learning sites such as Duolinguo and Memrise use but present many other non-typical language games as well. The games on this site are based on popular games that are familiar to all generations such as Memory, crossword puzzles, Snake, Hangman, Word Search, Tic-Tac-Toe, and Tetris. There are also some games that are inspired by modern games such as Flappy Glagolitic which was based on the popular game Flappy Bird designed in 2013 for mobile devices (Williams 2014) and Glagoljatrix which was based on the FPS (first person shooter) game SUPERHOT made in 2016 for PC and new generation consoles such as Xbox ONE (Orland 2016).

[2] Language Games and Multimedia Display of Language. 2016. http://www.jezicneigre.com (2.4.2017)

[3] Glagolitic games. 2017. https://zci.stin.hr/hr/article/112/osmosmjerka_s_glagoljicom (2. 4.2017)

[4] Croatian language in games. 2017. http://hrvatski.hr/igre/ (2. 4.2017)

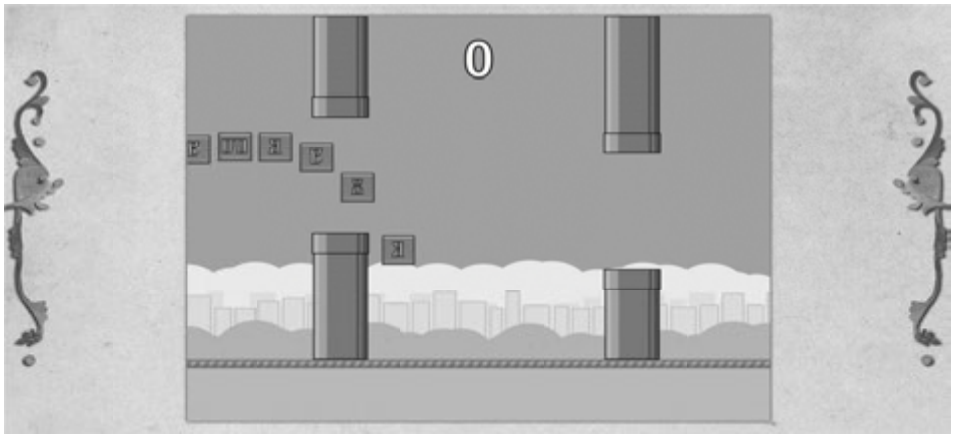


Figure 1. Flappy Glagolitic – a game inspired by Flappy Bird
which can be used for learning Glagolitic letters

The reason for adding *multimedia display of language* in the title of the website is not to exclude content such as interactive maps or timelines, e.g. the timeline presenting the history of Croatian dictionaries[5]. Some games for learning German were made in collaboration with a student of German.

The design, including the games, is web responsive, so games can be played on mobile devices. The games are divided into categories based on the language and each game has multiple tags (key words) which connect games having similar content and games of the same type. A Facebook page for the site was created and is used for promoting games as at present everything is spread through social media[6]. The site post for games has icons for sharing games through different social networks such as Facebook, Twitter, LinkedIn, and Pintrest. The site has no commercials, all of the games are free and it does not require registration. Many of the games were transferred to the new site Croatian in School[7] of the Institute of Croatian Language and Linguistics. On this website a special section is dedicated to language games and quizzes. Language games for learning German are also featured as a link among teaching materials on the website of the School for nurses Vrapče.

[5] Croatian dictionary history timeline. 2016. http://jezicneigre.com/hr/povijest-rjecnika/ (2. 4.2017)

[6] Language games and multimedia display of language – Facebook page. 2017. https://www.facebook.com/jezicne/ (2. 4.2017)

[7] Croatian in School. 2017. http://hrvatski.hr/ (2. 4.2017)

Computer Games and Other Multimedia Content

Similar site for learning Computer Science in school[8] was created. On this site students can play games for learning programming language generations, computer components, and other computer related topics.

This site was created for high school students who have Computer Science in their freshman and sophomore years. This site has a design similar to the site analyzed above. The games are divided into categories and have tags. Most of the games, e.g. quizzes, crossword puzzles, memory games and tic-tac-toe games are similar to the games on the site mentioned above but their graphical design if different and looks more computer-like. The content for students is being made based on the data that they are studying with their teacher. The most popular game on this site is Binary Tetris. It consists of two columns, one is for placing Tetris blocks and the other switches binary digits from 0 to 1 and vice versa depending on the position of the blocks. This game also uses a font that looks very 8-bit arcade like. Every time a player completes a row, the game pauses and a multiple choice question appears on the screen which the player has to answer to continue the game. If a player answers correctly, the game continues at normal speed and offers the player an extra point but if the answer is incorrect, the player loses a certain number of points and the game speeds up, which makes playing harder. The games on this site also have buttons which allow them to be shared on social networks but currently there is no official social website for these games.

Figure 2. Binary Tetris is a game of Tetris in which questions appear for every completed row

[8] Computer games and Other Multimedia Content. 2016. https://informatickeigre.com/ (2. 4.2017)

Archival Games

In addition to education, some web games can also be used for promoting cultural activities. That is why some museums and libraries create games for promoting their activities and materials. However, there are almost no games for promoting archives and archival activities. That is why archival games are presented on the website of the *Croatian Archival Society*[9]. The games were created during the development of this site. The purpose of the games is to promote archival practice and materials kept in archives as well as archival pedagogy. There is a memory game with Croatian archives, a jigsaw puzzle where players put together parts to get a picture of an artefact kept in the archive. When a player finishes the puzzle he is given some basic information about the artefact. There is also a typing game which is based on the *Archival Dictionary* where the player has to type quickly English words and phrases that are used in archival practice. For each word he types correctly he is given its Croatian translation. This game is designed for archivists who want to learn archival terminology and have fun while first two games are designed for anyone interested in archives and are located in the section for archival pedagogy. This content is also in an early stage of development. However, there are plans to expand archival games based on the world map of archives[10] which is also on this site and which has enough content to create quizzes, crossword puzzles, word search games, and other puzzle games.

Figure 3. A jigsaw puzzle with archive holdings

[9] Archival games. 23.10.2016. http://had-info.hr/arhivisticke-igre (2.4.2017)

[10] The world map of archives. 22.10.2016. http://had-info.hr/arhivi-svijet/ (2.4. 2017)

Technology in Creating Games

While creating games, the first step is to have a concept for a certain game type. When creating a game that is similar to an already well-known game, the game developer should first check if there is an existing template or code which can be reused free of cost. All of the games presented in this paper were made using free technology accessible through the Internet. All of the games were made using the HTML format that unlike Adobe Flash or Microsoft Silverlight works on all web browsers. Most of the games, e.g. Memory, crossword puzzles, Word Search games, Tic-tac-toe, that do not require keyboard controllers can be played on mobile devices. Some of the games were created by using the H5P web platform[11]. The site allows free creation of many different educational web games through the website GUI and there are a lot of customizing options which allow a more specific design. Many games, such as Croatian spelling quizzes, were made with a quiz game type which allows the creation of multiple type questions, e.g. multiple choice, which can have one or more correct answers, fill in the blanks, drag-n-drop, drag text and mark the words questions. Quiz creators can decide on the number of points for each question and the feedback messages for certain answers. There is also a hint button which can be enabled and options for customizing text messages in buttons and results so quizzes can be made for multiple languages. The solutions and points can be shown immediately after the answer is given or after the whole quiz is completed. There is also a memory game in which some text information is displayed when a pair is matched. This is used for explaining objects in images such as archives in the memory with Croatian archives which was made for the website of the Croatian Archival Society. There is also a flashcard quiz in which an image is displayed and the player must type what he sees on the picture, interactive video where a question is displayed at a certain point in the video, timeline tool for telling stories in a chronological order, image hotspot for marking parts of the image and adding dialog boxes to them which open when the user clicks on a marked part of the image.

The rest of the games were made through available codes and algorithms found on CodePen[12] website, where web designers and programmers can share and comment each other's work. The codes on that site are protected with MIT license which means that they are free to use and modify without any restrictions but have to include the copyright notice in all copies or other modified versions of the work. The code for certain games such as Tic-tac-toe was modified so that questions appear each time a player makes a move. If a player answers a question correctly, he is allowed to place his mark on the table, if his answer is wrong, he loses a turn. Games such as Flappy Glagolitic and Glagolitic Word

[11] H5P. https://h5p.org/ (2.4.2017)

[12] CodePen. http://codepen.io/ (2.4.2017)

Search were modified so that they show Glagolitic fonts and use Croatian words for their learning corpus. They were also graphically modified in the style of the Middle Ages, when Glagolitic letters were used. Memory with Glagolitic letters was created by using Construct 2[13] game engine which allows the creation of the HTML5 game, by using GUI interface, which allows placing graphic resources on canvas (game stage), then placing already finished blocks of code for movement and controls onto the resources. This allows a much easier and faster creation of web and mobile games for developers who don't have strong programming skills. Unfortunately, Construct 2 free version has certain restrictions because it allows publishing games that have only a certain number of events. Other memory games for learning German and Croatian were created using GDevelope[14], which is very similar to Construct 2 but is open-source, which means that it is completely free to use without any license restriction. Once the tool for creating games is selected and game mechanics programmed, there is a need to create the visual identity for the game by using available graphics and resources. All of the graphical resources used for creating games are license free. They were mostly found on Pixabay[15] site, which offers thousands of photos and cliparts free for use for any purpose. They especially have a wide range of clipart vector graphics which can be modified to look good on any screen resolutions. Drawings for the game *A Day in a Life* on the site of Language Games, where a player must drag sentences to a drawing that represents a typical working day was drawn by one of the high school students at a computer science class. The sound clips for Glagolitic memory and snake were created by the musician Ivan Mihaljević[16] and sound clips for other games were downloaded from Freesound.org site which offers creative-commons licensed sound for musicians and sound lovers. So everything was done free of charge. There are many programs and available codes which can be used free of charge for creating games and other educational content such as timelines with TimeLine JS[17], maps with StoryMap JS[18], and videos with Wirewax[19].

[13] Construct 2. 2011. https://www.scirra.com/construct2 (2.4.2017)

[14] GDevelop. 1.4.2015. http://compilgames.net/ (2.4.2017)

[15] Pixabay. 2010. https://pixabay.com/ (2.4.2017)

[16] Ivan Mihaljević official site. 2017. http://www.ivanmihaljevic.com/ (31.3.2017)

[17] TimeLine JS. 21.9.2016. https://timeline.knightlab.com/ (31.3.2017)

[18] StoryMap JS. 21.11.2016. https://storymap.knightlab.com/ (31.3.2017)

[19] WIREWAX. 31.5.2016. http://www.wirewax.com/ (31.3.2017)

Figure 4. An example of a quiz created in H5P platform

Pilot research

A short pilot research was conducted in the School for nurses in Vrapče in Zagreb based on the author's use of web games on the site Computer Games and Other Multimedia Content as a means to motivate students to learn and repeat essential facts they learned in Computer Science classes. The games were mostly used at the beginning or the end of a lesson as a way to repeat and practice the previous lessons. An online survey created with KwikiSurveys was given to each of the first and second-year students about the use of games for motivating them. 69 students filled the questionnaire and the results show that 77% like when the teacher uses games for motivation. 63% considered that the games helped them learn. However, 51% students admitted that they did not use the site for repetition before the exam. This means that most of them were not aware of the importance of games for the exam although in the games there were some questions that ended in the exam. Their favorite game (54%) was Tic-tac-toe with questions followed by Memory where they match computer parts (22%). As this survey did not have a large number of examinees, which is at the moment impossible because these games are only used in the first two years and only in one school, it points to the conclusion that most teenagers like when educational games are used in class but are not aware that by playing them they can learn for the test in an efficient way. The results would probably be even better with elementary school children since younger children enjoy games more and Computer Science is not an obligatory subject in Croatian elementary schools so mostly only children who like computers and games enrol in these classes.

Table 1. Answers to the first three questions

Questions	Yes	No
1. Do you like it when the teacher uses computer games for repeating lessons?	53 (77%)	16 (23%)
2. Did you use the games when you were repeating for the exam?	34 (49%)	35 (51%)
3. Do you think that the games helped you with the learning process?	39 (63%)	23 (37%)

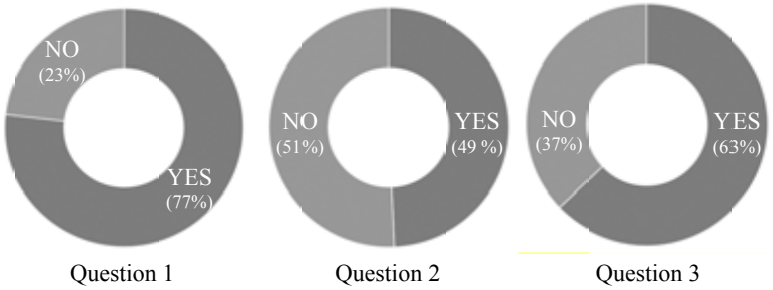


Chart 1. Answers to the first three questions

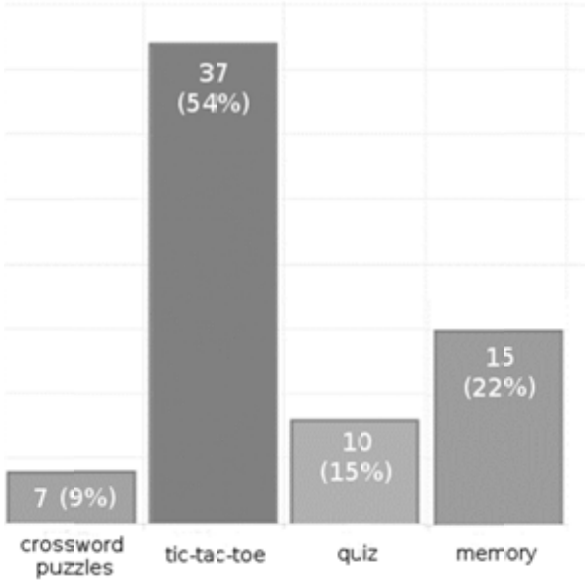


Chart 2. Answers to the question: Which of the game types do you like the most?

The last question in the questionnaire was which type of a game you would like to see next. This was an open question. Some students wrote names of popular commercial games such as Call of Duty, Grand Theft Auto and FIFA which cannot be used for educational purposes. Many students wrote Tetris, therefore the game Binary Tetris with questions from the field of computer science was created and it became one of the most popular games on this site. Other suggested games that can be made in future are Hangman and Pictionary.

Table 2. Relevant answers to the last question

Which game type would you like to see next?		
Tetris (7 answers)	Hangman (4 answers)	Pictionary (3 answers)

Conclusion

The idea of an interactive, highly engaging training and education is very old. An old Chinese proverb says: "Tell me, and I'll forget. Show me, and I may remember. Involve me, and I'll understand." (Popik 2012). While traditional educational methods such as reading, writing, and lecturing are still widely used, the gap between traditional methods and learning through modern interactive multimedia education technology is becoming narrower as technology is entering into more and more classrooms. With that in mind, the future of game-based learning seems to be promising. Even the results of the pilot research show that there is an interest for game based learning among high-school students. However, there is still a growing need to design and develop specific games for learning many different subjects and do research on the influence of gamification and game-based learning in specific classroom and learning situations. For that, a good collaboration between programmers, designers, and teachers has to be established or teachers have to learn how to design and create games for their subjects. Fortunately, technology is becoming more available and much easier to use so in a few years maybe many teachers will be able to create their own game-based learning resources. So in the future, the author hopes to expand his research to other high schools and/or elementary schools.

References

Ally, M.; Samaka, M. Open Education Resources and Mobile Technology to Narrow the Learning Divide. // The International Review of Research in Open and Distributed Learning. 14 (2013), 2; 1-2

Anderson, P. H.; Lawton, L. Business simulations and cognitive learning: Developments, desires, and future directions. // Simulation & Gaming. 40 (2009); 193-216

Archival games. 23.10.2016. http://had-info.hr/arhivisticke-igre (2.4.2017.)

Bilić, V.; Gjukić, D.; Kirinić, G. Mogući učinci igranja računalnih igrica i videoigara na djecu i adolescente. // Napredak : časopis za pedagogijsku teoriju i praksu. 151 (2010), 2; 195-213.

Binary Tetris. 13.1.2017. https://informatickeigre.com /1r/binarni-teris-internet/ (2.4. 2017.)

CodePen. http://codepen.io/ (2.4.2017)

Computer games and Other Multimedia Content. 2016. https://informatickeigre.com (2. 4.2017.)

Construct 2. 2011. https://www.scirra.com/construct2 (2.4.2017)

Croatian dictionary history timeline. 2016. http://jezicneigre.com/hr/povijest-rjecnika/ (2. 4.2017.)
Croatian in School. 2017. http://hrvatski.hr/ (2. 4.2017.)
Croatian language in games. 2017. http://hrvatski.hr/igre/ (2. 4.2017.)
Dovedan, Z.; Seljan, S.; Vučković, K. Multimedia in Foreign Language Learning. // Proceedings of the 25th International Convention MIPRO 2002: MEET + MHS. Rijeka: Liniavera, 2002. 72-75
EBESCO. A Comprehensive Full-Text Database for Multidisciplinary Research. 24.2.2012. https://www.ebscohost.com/academic/academic-search-complete (29.3.2017)
Garris, R.; Ahlers, R.; Driskell, J. E. Games, motivation, and learning: A research and practice model. // Simulation & Gaming. 33 (2002); 441-467
Gasston, P. Moderni web: responzivan web dizajn uz HTML5, CSS i JavaScript. Zagreb: Dobar plan, 2013
GDevelop. 1.4.2015. http://compilgames.net/ (2.4.2017.)
Glagolitic games. 2017. https://zci.stin.hr/hr/article/112/osmosmjerka_s_glagoljicom (2. 4.2017.)
Gros, Begoña. Digital Games in Education: The Design of Games-Based Learning Environments. // Journal of Research on Technology in Education. 40 (2007), 1; 31-36
H5P. https://h5p.org/ (2.4.2017.)
Hall, R. Moodle Tip: How to award badges in Moodle based on different levels of performance. 4.8.2015. http://blog.howtomoodle.com/blog/moodle-tip-how-to-award-badges-in-moodle-based-on-different-levels-of-performance (30.3.2017.)
Higgins, M. Five Video Games That Owe Their Success To The Wars Of The Past. 9.1.2017. https://www.warhistoryonline.com/instant-articles/five-video-games-owe-success-wars-past.html (30.3.2017.)
Ivan Mihaljević official site. 2017. http://www.ivanmihaljevic.com/ (31.3.2017.)
Language games and multimedia display of language – Facebook page. 2017. https://www.facebook.com/jezicne/ (2. 4.2017.)
McFarlane, A.; Sparrowhawk, A.; Heald, Y. Report on the educational use of games. 2002. http://www.teemorg.uk/resources/teem_gamesined_full.pdf (31.3.2017.)
Meihua, Q.; Clark, Karen R. Game-based Learning and 21st century skills: A review of recent research. // Computers in Human Behavior. 63 (2016), 50-58
Mitrović, P.; Seljan, S. Computer Learning of Small Math Using MATΣMATIX in English Class. // Proceedings of the 31st International Convention on Information and Communication Technology, Electronics and Microelectronics MIPRO. Opatija: MIPRO, 2008. 248-252
Orlando, K. Superhot review: Time is on my side. 25.2.2016. https://arstechnica.com/gaming/2016/02/superhot-review-time-is-on-my-side/ (24.3.2017.)
Pixabay. 2010. https://pixabay.com/ (2.4.2017.)
Popik, B. "Tell me and I forget; teach me and I may remember; involve me and I will learn".19.12.2012. http://www.barrypopik.com/index.php/new_york_city/entry/tell_me_and_i_forget_teach_me_and_i_may_remember_involve_me_and_i_will_lear/ (31.3.2017.)
Slavorum. Croats developed a new game! Meet "flappy" glagoljica!. 17.11.2016. http://www.slavorum.org/croats-developed-a-new-game-meet-flappy-glagoljica/ (2. 4.2017.)
Spencer. Ubisoft to teach DS owners French and Spanish?. 15.5.2007. http://www.siliconera.com/2007/05/15/ubisoft-to-teach-ds-owners-french-and-spanish/ (30.3.2017.)
StoryMap JS. 21.11.2016. https://storymap.knightlab.com/ (31.3.2017.)
The world map of archives. 22.10.2016. http://had-info.hr/arhivi-svijet/ (2.4. 2017.)
TimeLine JS. 21.9.2016. https://timeline.knightlab.com/ (31.3.2017.)
Trybus, J. Game-Based Learning: What it is, Why it Works, and Where it's Going. 14.1.2009. http://www.newmedia.org/game-based-learning--what-it-is-why-it-works-and-where-its-going.html (2.4.2017.)

Williams, R. What is Flappy Bird? The game taking the App Store by storm. 29.1.2014. http://www.telegraph.co.uk/technology/news/10604366/What-is-Flappy-Bird-The-game-taking-the-App-Store-by-storm.html (24.3.2017)

Wilson, K. A.; Bedwell, W. L.; Lazzara, E. H.; Salas, E.; Burke, C. S.; Estock, J. L.; Conkey, C. Relationships between game attributes and learning outcomes. // Simulation & Gaming. 40 (2009); 217-266.

WIREWAX. 31.5.2016. http://www.wirewax.com/ (31.3.2017.)

CryptoBase
A cryptography-based learning application

Vedran Juričić
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
vedran.juricic@gmail.com

Ian Christian Hanser
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
cryptobase11235@gmail.com

Dino Smrekar
Faculty of Humanities and Social Sciences
Ivana Lučića 3, Zagreb, Croatia
dino.smrekar@outlook.com

Summary

The authors are exploring the possibility of creating a cryptography-based educational application intended for the classroom and personal use. The basic idea is to help students or people interested in cryptography in their learning activities by providing them with an application capable of presenting basic information about codes and ciphers. Another goal is to give the users a chance to test out all of the ciphers contained within the application. One positive aspect of this kind of software is the fact that it can be adjusted to the user by using different languages and their alphabets in the learning process. Aside from personal use, this software can be incorporated in the classroom as an educational aid. It can also be used by teachers as an aid in grading student's tests.

Keywords: cryptography, ciphers, encryption, decryption, education, security, classroom aids

Introduction

In these modern times, we consider information as one of our most valuable resources. This idea evolved through history as technology developed. Since we view information as an extremely valuable resource, the need for masking and hiding information from unwanted interceptors arose quite fast. The solution to this problem was the development of cryptography.
Other key advances were made in terms of teaching and learning by developing different presentation techniques to accommodate different learning styles. We

have also seen a drastic change in student's learning habits with the rise of the World Wide Web.

The purpose of this project is to present CryptoBase, a cryptography-based learning application that can be used as a personal learning aid, an educational tool in classrooms or as help for teachers while evaluating student's tests. The application focuses on the user's interaction by allowing him to choose between three languages and presenting the user with a form which encrypts or decrypts an input message. The reason for creating such a project stems from the idea of incorporating different learning aids to enhance the student's learning experience[1]. The application is not intended only for students and focuses on anyone interested in learning about cryptography.

Cryptography and its contemporary use

Cryptography is a scientific discipline which focuses on studying and developing methods of sending messages in a concealed form, that is, in such a way that they can be read only by the person for whom the message is intended[2]. The two main methods of concealing messages in cryptography are codes and ciphers, which are defined as[3]:

- A cipher refers to an algorithm which replaces the order of letters or replaces each letter with a symbol or a different letter based on a key. The algorithm consists of encryption and decryption steps (steps required to make a message readable or unreadable.) The second important factor which affects the cryptographic algorithm is the key (a single word, phrase or string used for encrypting and decrypting messages.) This way, the encrypted messages may be read only by people who know the key.
- A code refers to a method of replacing words in a message with certain symbols, numbers or individual letters. A code always requires the use of a code book, as it serves as a reference in both the encryption and decryption process.

A code book is a lookup table consisting of words or phrases, and their corresponding code[4] (there also may be multiple codes intended for one word or phrase.)

Although cryptography was first used for military purposes, we can see its presence in every aspect of our lives, ranging from bank transactions, data encryp-

[1] Manichander, T. Emerging Trends in Digital Era Through Educational Technology, Second Edition. Solapur: Ashok Yakkaldevi. 2016. pp. 42-45.

[2] Stinson, Douglas Robert. Cryptography: Theory and Practice, Third edition. Ontario: Chapman and Hall/CRC, 2006. p. 1.

[3] Churchhouse, Robert. Codes and Ciphers: Julius Caesar, The Enigma, and the Internet. Cambridge: Cambridge University Press, 2002. p. 5.

[4] Anderson, Ross. Security Engineering: A Guide to Building Dependable Distributed Systems, First edition. John Wiley and sons 2001. p. 79.

tion and the majority of our day-to-day communication. Reliance on this certain technology also means that it is one of the key factors of online privacy[5].

CryptoBase project overview

The application currently contains 8 traditional ciphers which belong into the subcategories of monoalphabetic, polyalphabetic and transposition ciphers. All subcategories and its containing ciphers are organized in order of their complexity and each cipher introduces a new cryptographic method. This is because some of the advanced traditional ciphers incorporate simple cryptographic methods. It also gives us the ability to produce new ciphers by combining different ciphers. This application provides the users an interface which enables him to encrypt or decrypt custom message. The user can also change the key used in the cipher and the algorithm will also provide the result of encryption or decryption and a detailed overview how each letter has changed. This data may also vary from cipher to cipher because of different methods used in the processes of encryption and decryption.

- In the case of monoalphabetic ciphers, the displayed data includes both the plaintext and the ciphertext alphabets along with the encrypted or decrypted message.
- In the case of polyalphabetic ciphers, since the encryption process calculates a new letter based on the plaintext letter and a key letter, the displayed data includes each letter of the plaintext, along with the corresponding key letter and, of course, the resulting ciphertext letter.
- Finally, in the case of transposition ciphers the program calculates a matrix with the same number of columns as the key's length, and fills it up with the plaintext letters. This way the users have a clear view of how these types of ciphers operate because, aside from the result ciphertext, the program displays a step by step description of the encryption or decryption process.

Another key feature of this application is that it is currently translated into the English, Croatian and German language. The user is given a choice to switch between languages while the app is running and all of the data is displayed in the user's preferred language[6]. While a language is selected, the program also alerts the user if any of the input letters do not exist in the selected language's alphabet.

Application layout – navigation bar

The navigation bar always contains the back button and the language selection drop box. This is to enhance navigation between different windows of the appli-

[5] Cobb, Chey. Cryptography For Dummies, First edition. Indianapolis: Wiley Publishing, Inc. 2004. p. 23.

[6] Vigdorchik, Igor. WPF localization for dummies 13 December, 2011. https://www.codeproject.com/Articles/299436/WPF-Localization-for-Dummies (17 May 2017).

cation and to allow language changing. Another additional drop box appears when inside the cipher specific view. This drop box contains ciphers that belong in the same subdivision as the selected cipher.

Application layout – the main menu

The main menu consists of a tab bar which is divided into three subdivisions of traditional ciphers. Currently present categories include: monoalphabetic, polyalphabetic and transposition ciphers. When a category is selected, the tab bar container fills up with the ciphers of the selected category. Links to specific ciphers are displayed in the form of a button which contains an icon and the cipher's name[7].

Figure 1. The main menu

Application layout – specific cipher view

This view is divided into two equal halves. The left half represents the interactive part of the application and it consists of:

- the cipher's name
- the message input field – This field is intended for the user's input, it can be either a plaintext or a ciphertext
- The key input field – This field enables the user to enter a key which will be used while decrypting or encrypting a message, sometimes it will be disabled since some ciphers do not require a key
- the output field – The result of encryption or decryption (based on the user's choice) is written here
- The encryption data field – after the application has encrypted or decrypted a message it will also print out a detailed description of how each letter was calculated. The look of this box varies from cipher to cipher.

[7]Stovell, Paul. WPF Navigation 2 October 2009. www.paulstovell.com/blog/wpf-navigation (17 May 2017.)

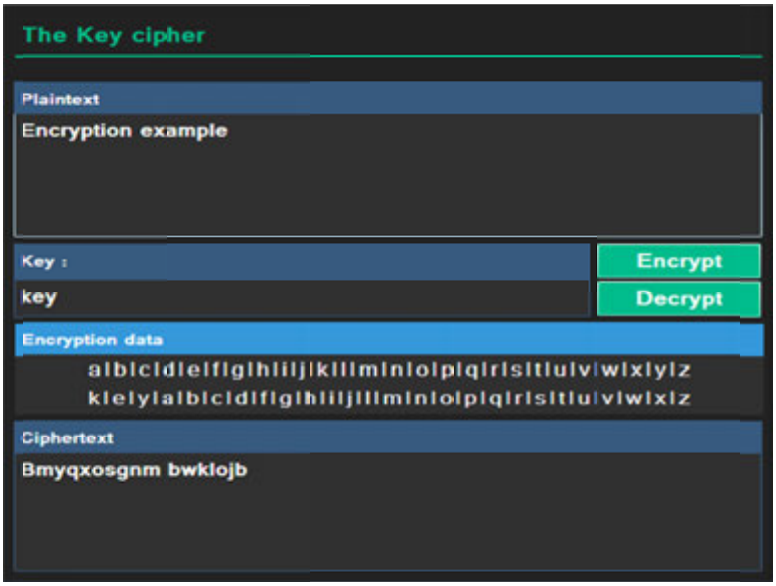


Figure 2. Specific cipher view – left half

In the case of monoalphabetic ciphers, the cipher always refers to one plaintext and one ciphertext alphabet. Because of that, the data is divided into two rows, with the first one containing the plaintext alphabet and the second one containing the ciphertext alphabet, while the number of columns varies based on the user's selected language.

In the case of polyalphabetic ciphers, every letter is calculated by assigning each letter of the alphabet a number and adding a key letter's value to a plaintext or ciphertext letter's value. That way we can print out the input, key and result letters, along with their numerical representations.

Finally, in the case of transposition ciphers, the resulting message is calculated by rearranging the letters of the input message. The program draws a different table based on the key and input message lengths along with indicators of how letters should be rearranged.

The Right side of this view is consisting of three different drop-down menus. The first menu is titled "about the cipher" and holds historical information about the cipher, such as: author details, historic use and development. The second menu is titled "encryption steps" and the third one "decryption steps" and each of these drop down menus holds a detailed description of the encryption and decryption processes presented in the user's preferred language. These drop boxes always contain one example of a message being encrypted or decrypted and all the data linked to the process.

Figure 3. Specific cipher view – right half

Conclusion

Although these present features serve a purpose and contribute to one's personal learning, there is still room for improvement. The first important step in advancing this application is to incorporate more languages, as to accommodate a wider audience. Also, to make the app run on different devices, it is possible to write a server API which can be accessed both by personal computers and mobile devices. In order to aid the users even more, it is necessary to build a question based system so the user can test his knowledge. In this mode, the program will provide the user with a random message from a database and will task him with encrypting or decrypting the given message.

References

Manichander, T. Emerging Trends in Digital Era Through Educational Technology. Second Edition. Solapur: Ashok Yakkaldevi. 2016.

Stinson, Douglas Robert. Cryptography: Theory and Practice, Third edition. Ontario: Chapman and Hall/CRC, 2006.

Anderson, Ross. Security Engineering: A Guide to Building Dependable Distributed Systems, First edition. John Wiley and sons 2001.

Cobb, Chey. Cryptography For Dummies, First edition. Indianapolis: Wiley Publishing, Inc., 2004.

Vigdorchik, Igor. WPF localization for dummies 13 December, 2011. https://www.codeproject.com/Articles/299436/WPF-Localization-for-Dummies (17 May 2017).

Churchhouse, Robert. Codes and Ciphers: Julius Caesar, The Enigma, and the Internet. Cambridge: Cambridge University Press, 2002.

Stovell, Paul. WPF Navigation 2 October 2009. http://www.paulstovell.com/blog/wpf-navigation (17 May 2017).

INFuture2017 reviewers

All papers were reviewed by at least two reviewers. INFuture relies on the double-blind peer review process in which the identity of both reviewers and authors as well as their institutions are respectfully concealed from both parties.

INFuture2017 gratefully acknowledges its reviewers:

Reviewer name	Institutional affiliation
Amelia Acker	Austin School of information, University of Texas, Texas, USA
Winton Afrić	University North, Koprivnica, Croatia
Željko Agić	Department of Computer Science, IT University of Copenhagen, Copenhagen, Denmark
Darko Babić	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Croatia
Mihaela Banek Zorica	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Maja Baretić	KBC Rebro, University of Zagreb, Zagreb, Croatia
Clive Billenness	University of Brighton, Brighton, UK
Roderic G. Broadhurst	Australian National University, Canberra, Australia
Maximilian Brosius	University of St. Gallen, Switzerland
Ethan Brown	Portland-based interactive marketing agency, Oregon, USA
Charles Buabeng-Andoh	Pentecost University College, Sowutuom, Ghana
Lluis-Esteve Casellas I Serra	Ajuntament de Girona, Catalonia, Spain
Janet Delve	University of Brighton, UK
Anca Dinicu	"Nicolae Bălcescu" Land Forces Academy, Sibiu, Romania
Ivan Dunđer	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Anna Dziemianko	Adam Mickiewicz University in Poznań, Poland
Kim Ebensgaard Jensen	University of Copenhagen, Denmark
Andrew J. Flanagin	Department of Communication, University of California, Santa Barbara, California, USA
Eliana E. Gallardo-Echenique	Faculty of Educational Sciences and Psychology, Universitat Rovira i Virgili Carretera de Valls, Tarragona, Catalonia, Spain
Chiara Garau	Department of Civil and Environmental Engineering and Architecture (DICAAR), University of Cagliari, Cagliari, Italy

Reviewer name	Institutional affiliation
Yasemin Gülbahar Güven	Ankara University, Turkey
John Horsfield	Alliance Consulting
Tomislav Ivanjko	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Rafal Jaworski	Faculty of Mathematics and Computer Science, Adam Mickiewicz University in Poznan, Poland
Debora Jeske	University Colleage Cork, Ireland
Vedran Juričić	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Ksenija Klasnić	Department of Sociology, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Lejla Kodrić Zaimović	Faculty of Philosophy, University of Sarajevo, Sarajevo, BiH
Mila Koeva	Faculty of Geo-information Science, University of Twente, Enschede, The Netherlands
Vlatka Lemić	Croatian State Archives, Zagreb, Croatia
Robert Lew	Adam Mickiewicz University in Poznań, Poland
Nikola Ljubešić	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Vjera Lopina	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Melina Lučić	Croatian State Archives, Zagreb, Croatia
Sonia Lupica Spagnolo	Architecture Built environment and Construction engineering (ABC) department, Politecnico di Milano, Italy
Martina Matovinović	KBC Rebro, University of Zagreb, Croatia
Shin'ichiro Matsuo	MIT Media Lab, Massachusetts, USA
William K. Michener	College of University Libraries & Learning Sciences, University of New Mexico, Albuquerque, New Mexico, USA
Nives Mikelić Preradović	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Pearse Murphy	Athlone Institute of Technology, Ireland
Dana Naous	University of Lausanne, Lausanne, Switzerland
Hilary Nesi	Coventry University, Coventry, UK
Marko Odak	University of Mostar, BiH
Santanu Pal	University of Saarlandes, Saarbrucken, Germany
Krešimir Pavlina	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia

Reviewer name	Institutional affiliation
Franjo Pehar	Department of Information Sciences, University of Zadar, Zadar, Croatia
Piotr Piasecki	iComply Investor Services Inc., Burnaby, British Columbia, Canada / Factom, Austin, Texas, USA
Martina Poljičak Sušec	Croatian Bureau of Statistics, Zagreb, Croatia
Anthony Pym	University in Tarragona, Catalonia, Spain
Danijel Radošević	Faculty of Organisation and Informatics, University of Zagreb, Varaždin, Croatia
Nuraidawany Rashid	Kementerian Pendidikan, Malaysia
Riccardo Rialti	University of Florence, Florence, Italy
Paul van Schaik	Teesside University, School of Social Sciences, Business and Law, Middlesbrough, UK
Sanja Seljan	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Roman Senkus	University of Toronto, Toronto, Canada
Vjeran Strahonja	Faculty of Organization and Informatics, University of Zagreb, Varaždin, Croatia
Olof Sundin	Department of Arts and Cultural Sciences, Lund University, Sweden
Nikša Sviličić	Proactiva Ltd., Croatia
Mária Šimková	L'udovit Štur Institute of Linguistics, Slovak Academy of Sciences, Bratislava, Slovakia
Sonja Špiranec	Department of Information and Communication Sciences, Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb, Croatia
Mireille Vale	School of Linguistics and Applied Language Studies, Victory University of Wellington, New Zealand
Daniel Vladušić	XLab, Ljubljana, Slovenia
Bencie Woll	University College London, UK

The INFuture2017 conference is supported by

The Miroslav Krleža Institute of Lexicography
http://www.lzmk.hr

SV Group
http://www.svgroup.hr/en

Ericsson Nikola Tesla
https://www.ericsson.hr/

Mikrocop
http://www.mikrocop.hr

Ideje.hr
http://ideje.hr/

Zagreb Tourist Board
http://www.zagreb-touristinfo.hr/