Conference on Grey Literature and Repositories

Národní technická knihovna
2016

conference on grey literature and repositories

proceedings 2016

CONFERENCE ON GREY LITERATURE AND REPOSITORIES

Proceedings

National Library of Technology, 2016

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

English conference website

(**http://nrgl.techlib.cz/conference/9th-conference-on-grey-literature-and-repositories**)

Czech conference website

(**http://nusl.techlib.cz/konference/9-rocnik-konference**)

Publisher: National Library of Technology, Technická 6/2710, Prague, Czech Republic

Editor: Mgr. Hana Vyčítalová

ISSN: 2336-5021

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

Programme Committee:

PhDr. Eva Bratková, Ph.D., Charles University

Ing. Jozef Dzivák, Slovak Chemistry Library

Dr. Dominic Farace, GreyNet

Ing. Martin Lhoták, Academy of Sciences Library

Ing. Jan Mach, University of Economics, Prague

Doc. JUDr. Radim Polčák, Ph.D., Masaryk University

Dr. Dobrica Savić, Nuclear Information Section, IAEA

Organizing Committee

Mgr. Michaela Charvátová, National Library of Technology

Mgr. Lenka Patoková, National Library of Technology

PhDr. Petra Pejšová, National Library of Technology

Mgr. Hana Vyčítalová, National Library of Technology

List of Reviewers:

Stephania Biagoni, Italian National Research Council

Ing. Lukáš Budínský, Tomas Bata University in Zlín

Dr. Jan Dvořák, Charles University

Dr. Dominic Farace, Greynet

Mgr. Tomáš Foltýn, National Library of the Czech Republic

Mgr. Jan Hutař, Archives New Zealand

MgA. Michal Indrák, Ph.D., Moravian Library

Mgr. MgA. Jakub Míšek, Masaryk University

Doc. PhDr. Richard Papík, Ph.D., Charles University

Mgr. Zuzana Petrášková, National Library of the Czech Republic

PhDr. Jindra Planková, Ph.D., Silesian University in Opava

Doc. JUDr. Radim Polčák, Ph.D., Masaryk University

Mgr. Pavla Rygelová, VŠB – Technical University of Ostrava

Małgorzata Rychlik, Adam Mickiewicz University in Poznań

Joachim Schöpfel, University of Lille 3, France

Mgr. Václav Stupka, Masaryk University

Marcus Vaska, University of Calgary

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

Table of Contents

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

IMPACT OF CURRENT INFORMATION TECHNOLOGY TRENDS ON THE FUTURE OF GREY LITERATURE

Dobrica Savić

d.savic@iaea.org

International Atomic Energy Agency (IAEA), United Nations

Abstract

This paper deals with emerging information technology (IT) and other trends and their impact on grey literature. It is based on analysis of the most prevalent trends in general information management and new IT solutions, which will define and impact the digital future of related information management activities, as well as that of grey literature. The analysis was done based on seven reports issued in 2016 by five world leading consulting and service companies that have a special interest in researching the impacts of IT on our business environments, work procedures and behaviours.

These emerging IT developments make big changes in related activities not only possible, but also increasingly necessary. IT has become a driving force for change, innovation and new opportunities by offering powerful tools for furthering the digitization of our work processes, and the services we offer, as well as the creation of information products such as grey and other literature. This emerging new digital environment has the potential to affect all aspects of the way we do business, and the way we relate to our customers and the world around us. In order to meet expectations and benefit from this challenging opportunity, information managers need to be cognizant of new IT trends and the possibilities they offer in order to define the best strategies and action plans for their successful implementation in the future.

This paper elaborates the emerging IT and other related trends and their potential benefits, and offers some concluding guidelines relevant to the field of grey literature management, its use, and corresponding challenges.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

Keywords
Information Technology, Information Management, Trends, Grey Literature

---

Introduction
During the last few decades, we have witnessed a real revolution in computing and communications. There was a dramatic increase in the processing power of new information technologies (IT) accompanied by a decrease in the cost of communication. According to Moore's law, the processing power of microchips is doubling every 18 months. This resulted in a tremendous increase of the processing power of computers from 1956 to 2015 by 1 trillion times[1]. This and many other IT advances are a driving force behind changes which offer significant opportunities, but also pose some considerable challenges to the way we conduct our business and our lives.

IT developments and other changes also affect the way we currently create, disseminate and use grey literature, and it will continue to impact it in the future. This paper is based on analysis of the most prevalent trends in general information management and new IT solutions, which will define and impact the digital future of related information management activities, as well as grey literature. The analysis was done based on seven reports[2] issued in 2016 by five world leading consulting and service companies that have a special interest in researching the impacts of IT on our business environments, work procedures and behaviours.

These emerging IT developments make big changes in related activities not only possible, but also increasingly necessary. IT has become a driving force for change, innovation and new opportunities by offering powerful tools for furthering the digitization of our work processes, and the services we offer, as well as the creation of information products, such as grey and other literature. The emerging new digital environment has the potential to affect all aspects of the way we do business, and the way we relate to our customers and the world around us. In order to meet expectations and benefit from this challenging opportunity, information managers need to be cognizant of new IT trends and the possibilities they offer, in order to define the best strategies and action plans for their successful implementation in the future.

This paper elaborates the emerging IT and other related trends and their potential benefits, and offers some concluding guidelines relevant to the field of grey literature management, its use, and the corresponding challenges. In particular, it looks at the impact on grey literature through the technology deployed, products and services offered, and customers and staff working directly with grey literature.

The impact of current information technology trends on the future of grey literature is presented through IT progress and the present state of information management, followed by the status and challenges of grey literature today. Based on the current IT trends, a set of information management relevant trends was elaborated, while the impact on grey literature is examined

[1] **http://pages.experts-exchange.com/processing-power-compared/**

[2] A complete list of reports is available as part of the References at the end of this paper.

through a prism of the actual technology used, products and services offered, and through the changes impacting the work environment.

The progress of information technology

Ever since the creation of Z1, the first programmable computer, in 1936 by Konrad Zuse (Computer Hope, 2016), the introduction of the ENIAC[3] in 1946, and the first IBM personal computer in 1981, the progress of information technology has been characterized by tremendous developments, boundary-pushing innovations, and constant change; all happening at a very fast pace.

"Moore's law" characterizes some parts of this fast progress, where the number of transistors in a dense integrated circuit doubles approximately every 18 months and the processing power of computers from 1956 to 2015 increased 1 trillion-fold. Beside personal computers, mobile telephones also make up a portion of IT history and progress. In 1994, the first mobile phone to feature software applications (IBM Simon) was introduced; followed in 2007 by the introduction of iPhone (first commercial smartphone to use finger input), and the Samsung Galaxy S in 2010. The speed of cell phone introduction is especially significant. It took 13 years, from 1975 to 2008, to sell one billion PCs, while in 2013 alone, the sale of cell phones reached 1 billion! The introduction of the Internet and particularly of the World Wide Web in 1991 gave another impetus to the growth of IT. Connecting web technology and smart phones represents a quantum leap for the information management. Today, 89% of China's 668 million Internet users access the web from their mobile devices. The situation is similar with other developing nations. In January 2014, mobile phone Internet usage overtook PC Internet usage, opening yet another great window of challenge and opportunity for information suppliers, users and managers.

Parallel to all of these areas of information progress, another development is taking shape and becoming a practical reality. Artificial intelligence is now a new frontier that has made huge progress from the first Alan Turing tests in 1950, to Google's AlphaGo in January 2016, which crossed a major artificial intelligence threshold by besting human grandmaster Lee Sedol at the famously complex game of Go. Present today in the form of small intelligent snippets and larger smart agents, artificial intelligence represents a tremendous area of potential use and implementation in information management and, especially, in the area of grey literature identification, processing and dissemination.

Present state of information management

The present state of information management, and, in particular, certain sectors, such as libraries and information centres, is somewhat discouraging. The disappearance of many libraries around the world and their diminishing budgets and power is evident everywhere. Public and specialized libraries, and in particular libraries belonging to small and medium corporations, have been hit particularly hard. As an illustration, in the United Kingdom alone,

[3] Electronic Numerical Integrator And Computer (ENIAC)

almost 8,000 jobs, a quarter of all library staff, have disappeared in the last 6 years. During the same period, 343 libraries were closed, leading to fears about the future of the profession[4].

As staff cuts and professional work decreases, the number of volunteers working in libraries has increased, leading to the belief that library services can be run by volunteers. Although volunteers are welcome, they cannot deliver adequate professional and ethical services offered by information professionals. Skill gaps are evident and they have a negative impact on library patrons and information users.

One of the factors having a negative impact on information management, and libraries in particular, is the notion that everything is already on the web. All we need to do is search Google and all of our information needs are met. There is very little understanding that in order to appear as a book for sale on Amazon, or as a Google search result, a piece of information first needs to be placed and maintained on a website. The reliability, authenticity and correctness of information resources is not given adequate consideration, which leads to errors, lost time, and erroneous outcomes. This competitive positioning of Amazon and Google vs. libraries is of no benefit to the actual users who need information on daily basis to do their work.

Another element that defines the state of information management today is the remarkable increase in price for information content by almost all suppliers, making access even more difficult. A study done by Times Higher Education[5] found that the amount paid to Oxford University Press rose by 49.2 per cent between 2010 and 2014. The amount paid to Springer rose by 36.3 per cent and the amount to Wiley by 33.5 per cent. The smallest rise – 17.4 per cent – was in subscriptions to Elsevier journals. Overall subscription cost increased by 23.9 per cent. This price hike is also evident in the increasing cost of new library management systems and their related applications.

The protection of intellectual property rights present information management with challenges at all stages of work, starting from the creation of digital documents, and the selection and acquisition of external materials, to access rights and long-term preservation. Electronic publications offer an opportunity to open the access to information, and the flood of information through the Internet illustrates this. However, at the same time, it also requires information managers to increase access control and more closely monitor the use of information. All this is taking place at the same time as users increase their demands for faster delivery of information in a variety of formats and with certain value added to the raw information and documentation.

Grey literature challenges

Grey literature is part of the wider information management arena so its activities and related challenges are in many ways shared. Still, there are some specific challenges related to grey

[4] BBC News, 29 March 2016. **http://www.bbc.com/news/uk-england-35707956**

[5] Times Higher Education, 30 October 2014 **https:/goo.gl/yLlmDL**

literature and they include the actual concept of grey literature, grey literature processing, and its sustainability and usability.

Examples of grey literature include conference papers, reports, patents, dissertations, fact sheets, lectures, newsletters, course materials, memoranda, interviews, policy statements, posters, government documents, legislations, press releases, personal communication, photographs, bibliographies, speeches, physiological specimens, and others. Conceptually, grey literature should have a clear distinction from other forms of literature[6]. Grey literature should not be thought of as strictly 'literature', but rather as grey 'resources' as it can encompass many different formats depending on the discipline (Bichteler, 1991; Tyndall, 2008). It is accepted that it covers document types which are not commercially controlled by publishers (Schopfel, 2010), while the other view regards grey literature as the diverse and heterogeneous body of material that is made public, but is not subject to traditional academic peer-review processes (Adams et al., 2016). However, this distinction from other forms is not easy to achieve, especially since grey literature today covers many diverse types of documents including electronic forms, such as emails, blogs, webinars, comments, Tweets or Facebook postings.

There are many challenges related especially to the new forms of grey literature, such as blogs, various forum posts, comments, Tweets, etc. Collecting, processing, managing and preserving that type of grey literature is especially challenging. For example, blog archive incorporated as part of the blog page is important, not only from a historical and preservation point of view, but it is also important to the blog's success mainly for giving it special depth and credibility, characteristics that are becoming more and more valued by users.

Another set of challenges relates to processing and its reliability, since some key metadata elements are often missing, and to the lack of bibliographic controls and systematic collection. The core reasons for difficulties in identifying and acquiring this type of literature are due to its "poor bibliographic information and control, non-professional layout and format, and low print runs" (Augur, 1989). The implementation of bibliographic control through ISBNs, ISSNs or report number systems is not well organized and, therefore of not much help for grey literature repositories.

Sustainability is also a major challenge that grey literature faces. It is rare to find projects or plans for long-term preservation of grey literature, which, when combined with changes of its hosts, and lack of permanent location identifiers, directly affects its usability and sustainability. Substantial requirements for continuous financing in order to provide longer sustainability and usability of this valuable resource also represent a serious obstacle, since available funds are in most cases very scarce. Copyright ‘overprotection’ and other intellectual property related issues hinder the success of the open access movement and their drive to have as much literature as possible available to the users. This is particularly the case in the area of science and technology documents that need to be shared among researches and developers if they are to make the desired impact on speedy implementation of technical solutions.

[6] The 'Luxembourg definition', developed and approved during the Fourth International Conference on Grey Literature in 1999, defines grey literature as "that which is produced on all levels of government, academics, business and industry, in print and electronic formats, but which is not controlled by commercial publishers".

Current information technology trends

Many companies and consulting agencies deal with technological forecasts, particularly those relating to information technology. For this review, seven reports created by five leading consulting IT companies were selected and are presented here.

Gartner	Forbes	Forrester
1. The device mesh 2. Ambient user experience 3. 3D printing materials 4. Information of everything 5. Advanced machine learning 6. Autonomous agents and things 7. Adaptive security architecture 8. Advanced system architecture 9. Mesh App and service architecture 10. Internet of things architecture and platforms ***Gartner's top 10 strategic technology trends for 2016***	1. Connecting customers 2. Embracing millennials 3. Remote employee development and training 4. Strength based leadership 5. Add extra value to commodity products you sell 6. Corporate culture of customer service 7. Deliver results, not just solutions 8. Engage customers through fun and games 9. Integrate impartial content to support customer decisions 10. Develop “selling/solving” skills for non-salespeople ***Top 10 Business Trends That Will Drive Success In 2016***	1. Smart connected world 2. Systems of insight 3. APIs as strategy 4. Digital CX limitations 5. Security and risk rethink 6. Hyper-connected hyper-adopters 7. Business tech acceleration 8. Infrastructure snowballs 9. Software as part of the brand 10. Workforce technology ***The Top Technology Trends To Watch: 2016 To 2018*** 1. From customer-aware to customer-led 2. From data-rich to insight-driven 3. From perfect to fast 4. From silos to connected ***The Operating Model for Customer Obsession***
Deloitte	**Accenture**	
1. Right-speed IT 2. Augmented & virtual reality 3. Internet of Things: From sensing to doing 4. Reimagining core systems 5. Autonomic platforms 6. Blockchain: Democratized trust 7. Industrialized analytics 8. Social impact of exponential techn. ***Tech Trends 2016: Innovating in the digital era*** 1. Organizational design 2. Leadership 3. Culture 4. Engagement 5. Learning 6. Design thinking 7. Changing skills of HR organization 8. People analytics 9. Digital HR 10. Workforce management ***Global Human Capital Trends 2016***	1. Intelligent automation 2. Liquid workforce 3. Platform economy 4. Predictable disruption 5. Digital trust ***Technology Vision 2016 - People First: The primacy of people in a digital age***	

Table 1: Current information technology trends

Although somewhat different in their approach and the way they look at the future, identified current IT trends offer a very good starting point for establishing at least major IT influences or trends that will define the way we do business in the near future. After looking at the reports and their main elements summarised above, four different groups of trends dealing with different business areas were identified. They are:

Technology	**Customers**
▪ Secure architecture ▪ Autonomous agents ▪ Machine learning (algorithms) ▪ Internet of things (from sensing to doing) ▪ Application Program Interface (API) ▪ 3D printing	▪ Customer culture ▪ Connected world ▪ User experience ▪ Engage customers ▪ From data-rich to insight-driven
Products/services	**Employees**
▪ Added value ▪ Deliver results, not just solutions ▪ Social impact ▪ Predictable disruption ▪ Digital trust ▪ Analytics	▪ New generation ▪ Liquid workforce ▪ Remote work ▪ Learning & training ▪ New skills (leadership, sales) ▪ From silos to connected

Table 2: Categories of information technology trends

An attempt was made here to list the business areas and related trends in some order of priority and importance. However, it should be kept in mind that because identified trends come from different reports and different observations, they do not necessarily represent the best order of priorities, but rather an attempt to bring some order and importance to multiple factors. For example, it is very probable that from a technological point of view, secure IT architecture, and security in general, will represent the most important factor in the future. From customer related trends, many companies and organizations around the world are trying to build and implement "customer culture" – a culture where all the products, services and workflows are devoted to, geared towards and determined by customers.

In the area of products and services, we are presently experiencing growing demand to increase the value added to everything delivered, closely followed by both, the comprehensive delivery of results, not just outside solutions, and the impact of social media and social responsibility. The number of determining factors regarding employees is significant and their importance, viewed individually, is hard to rank by relevance. The introduction of a new generation of workers, with new demands and specific behaviours, is a constant, but 'liquid workforce'[7] and remote work are becoming more and more frequent. Learning new skills and continued training are of great significance, especially in dynamic areas such as information technology.

[7] According to Accenture, the 'liquid workforce' requires constant re-training in order to stay relevant in the midst of the digital revolution. A Liquid Workforce is one that is able to rapidly adapt and change based on the environment that they are in.

Impact of IT trends on grey literature

It is of great importance to explore what impact, if any, these four current IT trends might have on grey literature in the future. Since four major areas or groups of trends were identified, the same four areas need to be examined from the grey literature perspective. In other words, a parallel needs to be drawn between the impact of general technology on information management and its direct effect on grey literature. The following table offers this parallel.

Technology	**Customers**
▪ More difficult access to GL due to security constraints ▪ Higher level of IT expertise required to access and process GL ▪ More dynamic docs – less GL ▪ New tech-driven forms ▪ Increased amount of big data	▪ High expectations (e.g. comprehensiveness, relevance, aggregation, added value) ▪ Interconnectivity ▪ Top of the line finding tools ▪ Web 2.0 features (e.g. social networking, collaboration, user generated content) ▪ Tools to analyse and exploit big data ▪ Mobile expectations of the new generation ▪ Lack of understanding of GL value and importance
Products/services	**Employees**
▪ Availability of HR and financial resources ▪ Competition with 'big players' ▪ Lack of interest to make GL available ▪ Difficulty with going beyond local repositories ▪ Intellectual property protection ▪ Disappearing e-archives, older materials	▪ Lack of proper training and education ▪ Limited career development ▪ Frequent change of jobs and interests (i.e. lack of continuity and long-term planning) ▪ Changing technical requirements ▪ Business focus ▪ Culture of preservation missing ▪ Multitasking and rapid delivery expectations

Table 3: Grey literature related categories of information technology trends

The above table shows that all four areas of grey literature, namely technology, customers, products/services, and employees will be impacted in the future by the identified current information management trends. The impact is already being strongly felt across the information management sector, and it will continue to be of major concern, bringing many challenges and requiring well-planned, well-established, and well-financed actions. Some of these forthcoming changes could be regarded as significant, potentially having a serious negative impact on grey literature in general. For example, IT security constraints and restrictions, dynamic documents, available resources, intellectual property protection, required relevance of search results, inadequate training and limited career development options are difficult to overcome. Quite often, technical or product constraints get most of the attention and resources. However, most projects or new initiatives fail, not because of these factors, but rather because of ab untrained or unmotivated workforce, where the emphasis of actions should always have been. It would require many planned efforts to overcome all the anticipated challenges, and to make grey literature a valuable, viable and usable information future resource.

Conclusion

During the last few decades, IT developments have had an immense impact on the way we manage information in general, as well as on the way we create, disseminate and use grey literature. Based on the review of current IT trends and new solutions already in place, it can be concluded that this interdependency between IT developments and grey literature will continue in the future. Information technology will continue to define and affect the digital future of related information management activities, as well as grey literature.

In order to increase the future use of grey literature, it seems necessary (i) to open the relevant repositories and make them freely accessible to the public; (ii) to implement top performance technical solutions, such as modern databases, fast search engines, and efficient processing tools; and (iii) to provide immediate and free access to the full-text of documents, preferably in different record formats.

In order to increase access to grey literature and meet user needs, it would be beneficial (i) to simplify the basic search interface, while providing the option to use an improved advanced search; (ii) to incorporate rich features at every possible stage of grey literature processing, retrieval and use, but make them as discrete as possible; and (iii) to offer big data analysis tools so that repositories become part of a dynamic solution, rather than reference placeholders.

Finally, in order to increase the visibility of grey literature, it is necessary (i) to incorporate grey literature repositories into search engines, such as Google, Google Scholar, Baidu, Yandex, and others; (ii) to invest in the promotion of grey literature, its value and usefulness; and (iii) to promote continuous training and education at all levels, from schools and academia, to business and government.

The identified trends and emerging new digital environments have the potential to affect all aspects of the way we do business, and the way we relate to our customers and the world around us. In order to meet expectations and benefit from this challenging opportunity, information and grey literature managers need to be cognizant of new IT trends and the possibilities they offer in order to define the best strategies and action plans for their successful implementation in the future.

References

Technology trends reports

ACCENTURE. *Technology Vision 2016 - People First: The primacy of people in a digital age*. 2016. Available from: **https://goo.gl/Uc4ltY**.

DELOITTE. *Tech Trends 2016: Innovating in the digital era*. 2016. Available from: **http://goo.gl/ZaboVL**.

DELOITTE. *Global Human Capital Trends 2016*. Available from: **https://goo.gl/JlbEOf**.

ALTMAN, Ian. Top 10 Business Trends That Will Drive Success In 2016. *Forbes*. 1 December 2015. Available from: **http://goo.gl/9X4jU5**.

FORRESTER. *The Operating Model for Customer Obsession*. 3 November 2015. Available from: **https://goo.gl/hn1kLx**

FORRESTER. *The Top Technology Trends to Watch*: *2016 To 2018*. 8 September 2015. Available from: **https://goo.gl/XQbjvw**.

GARTNER. *Gartner's top 10 strategic technology trends for 2016*. 6 October 2015. Available from: **https://goo.gl/1Rq9fy**.

Other references

ADAMS at al., 2016. Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies. *International Journal of Management Reviews*. Available from: **http://onlinelibrary.wiley.com/doi/10.1111/ijmr.12102/full**.

AUGUR, Charles P., 1989. *Information Sources in Grey Literature*. London: Bowker-Saur, 1989.

BICHTELER J., 1991. Geologists and gray literature: Access, use and problems. *Science and Technology Libraries.* (11)*,* 39-50.

Computer Hope. When was the first computer invented? 5 September 2016. Available from: **http://www.computerhope.com/issues/ch000984.htm**.

SCHÖPFEL, Joachim, 2010. Towards a Prague Definition of Grey Literature*.* In *Twelfth International Conference on Grey Literature: Transparency in Grey Literature. Grey Tech Approaches to High Tech Issues*. Prague: GreyNet, 2010, p.11-26.

TYNDALL, J., 2008. How low can you go? Towards a hierarchy of grey literature. In: *Dreaming08: Australian Library and Information Association Biennial Conference*, Alice Springs, 2008. Available from: **http://hdl.handle.net/2328/3326.**

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

TIB AV-PORTAL: A RELIABLE INFRASTRUCTURE FOR SCIENTIFIC VIDEOS

Margret Plank

Margret.Plank@tib.eu

German National Library of Science and Technology, Germany

Abstract

With the AV Portal[1], the German National Library of Science and Technology (TIB)[2] in collaboration with the Hasso Plattner Institute (HPI)[3] has developed a user-oriented platform for scientific films. This portal offers free access to high-quality computer visualisations, simulations, experiments and interviews as well as recordings of lectures and conferences from the fields of science and technology. The automatic video analysis of the TIB AV Portal includes not only structural analysis (scene recognition), but also text, audio and image analysis. Automatic indexing by the AV Portal describes videos at the segment level, enabling pinpoint searches to be made within videos. Films are allocated a Digital Object Identifier (DOI), which means they can be referenced clearly. Individual film segments are allocated a Media Fragment Identifier (MFID), which enables the video to be referenced down to the second and cited. The creator of the audiovisual media segment can choose between an Open Access licence and a declaration of consent, enabling them to decide how they wish to permit TIB to utilise the material. TIB recommends the "CC-Namensnennung – Deutschland 3.0" licence, which ensures that the creator is acknowledged and permits the comprehensive use of audiovisual media in research and teaching.

[1] **https://av.tib.eu**

[2] **https://www.tib.eu/de/**

[3] **http://hpi.de/**

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

Keywords

Audiovisual Media, Audiovisual Portal, Multimedia Retrieval, Semantic Analysis

Introduction

Nowadays, publications produced by researchers often include a combination of an article, a dataset and a code of the scientific model as well as video and animation (Kraft et al, 2015). However, all these media types have different search, indexing and archiving requirements. In addition, the objects must also be connected to each other to enable cross-media research.

In order to tackle these challenges, the Competence Centre for Non-Textual Materials (KNM)[4] was founded at TIB in 2011. At KNM, an interdisciplinary team of experts in IT development, multimedia retrieval and ontologies, media archivists, information scientists and legal experts is engaged in fundamentally improving conditions of access and use for media types such as audiovisual media, 3D objects and research data.

KNM focuses on developing innovative solutions to problems in the areas of collecting, indexing, providing and (digitally) preserving non-textual materials. In the future, it should be possible for such material to be published, located, cited and made available on a permanent basis as easily as textual documents are today. To make this happen, KNM strives to develop infrastructures, tools and services to actively support users throughout the scientific process. In addition to finding solutions to specific users' needs and other object types, the team also keeps an eye on new domains of knowledge. To ensure that research approaches are transferred efficiently and effectively to everyday practice in digital libraries, the developments are consistently supported by user-centred software design, ensuring optimal usability of the portals and tools. Last but not least, the competence centre supports additional information facilities and providers as a knowledgeable point of contact in all matters concerning non-textual materials.

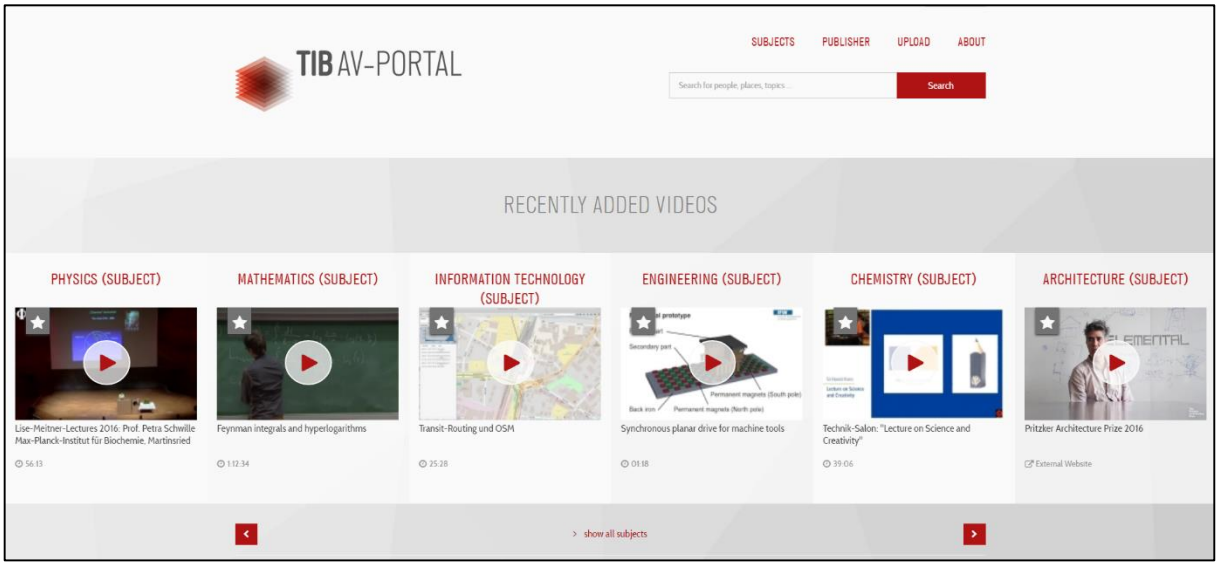


Figure 1: The home page of TIB's AV Portal (tib.av.eu)

[4] **https://www.tib.eu/de/forschung-entwicklung/nicht-textuelle-materialien/**

The TIB AV Portal

In a joint project with the Hasso Plattner Institute (HPI), the University Centre of Excellence in Software Systems Engineering affiliated with the University of Potsdam, TIB created a portal in this context that optimises access to scientific videos such as computer visualisations, learning materials, simulations, experiments, interviews, and recordings of lectures and conferences from the areas of science and technology. The key feature of the portal is its combination of state-of-the-art methods of multimedia retrieval and semantic analysis (Snoek et al, 2007). In 2010, the project team used focus groups and semi-structured interviews to ask researchers about what they needed most from a search system, the AV Portal is based on further analyses of that (Plank, 2012). These requirements included:

High-quality specialist content
Preferably free access and use
Long-term retrievability
Citability of films, preferably at the segment level
Good searchability
Links to additional research information

In 2011, a semi-functional prototype of the AV Portal was developed; in 2012-2013, the system was further developed and the beta version was created. Since spring 2014, the system has been fully operational at TIB. The page impressions increased from 78.936 in 2014 to 155.723 in 2016 (information as of October 2016).

TIB's AV Portal currently contains around 5,000 videos from the field of science and technology, as well as some 2,400 film credits with external links to other websites. The collection of the former IWF (Institute for Knowledge and Media) is also gradually becoming accessible online via the portal. The collection, which covers 100 years of scientific film history, was transferred to TIB in 2012[5]. A total of around 1,400 IWF films can already be accessed online; other titles are being added continuously as soon as the situation concerning rights can be clarified. In many cases, authors can be convinced to release their films under Open Access licences from the non-profit organisation Creative Commons[6] to make them freely accessible and usable for research and teaching.

[5] **https://www.tib.eu/de/recherchieren-entdecken/sondersammlungen/iwf-medienbestand/**

[6] **http://de.creativecommons.org/**

Process chain

A media asset management system (MAM) professionally captures the videos. The system has its own transcoder that handles all established codecs and creates statistics. The MAM system's underlying metadata schema on standardised registration of non-textual materials is based on the current DataCite Metadata Schema[7] and has been expanded by a few elements required for the detailed description of an AV medium. The entire metadata schema is made available to media providers online[8].

```
                    Mandatory Properties
  -->
- <creators>
  - <creator>
      <creatorName>name of the creator</creatorName>
      <nameIdentifier nameIdentifierScheme="String">any name identifier</nameIdentifier>
    </creator>
  </creators>
- <titles>
    <title>title of the video</title>
    <title titleType="Subtitle" language="eng">subtitle of the resource</title>
    <title titleType="AlternativeTitle" language="eng">alternative title of the resource</title>
  </titles>
- <publishers>
  - <publisher>
      <publisherName>name of the publisher</publisherName>
      <nameIdentifier nameIdentifierScheme="String">any name identifier</nameIdentifier>
    </publisher>
  </publishers>
  <publicationYear>2001</publicationYear>
  <language>ger</language>
```

Figure 2: Sample metadata record

To enable TIB to provide their users with videos via the AV Portal, media providers conclude a licence agreement with TIB and determine their terms of use. Simple licence agreements as well as various Creative Commons licences are available for selection. TIB explicitly recommends the Open Access "CC-Namensnennung – Deutschland 3.0" licence. This licence entails the fewest restrictions for use in research and teaching, and simultaneously guarantees that the author has to be mentioned. Standard licence agreements have been developed by TIB and made available online to media providers.[9]

Non-textual materials are digitally preserved if they are particularly important for science and teaching and of appropriate technical quality. TIB operates a professional digital preservation system called "Rosetta", which is jointly used by the German National Library of Medicine

[7] **https://www.datacite.org/**

[8] **https://av.tib.eu/about**

[9] **https://av.tib.eu/about**

(ZB MED)[10] and the Leibniz Information Centre for Economics (ZBW)[11]. If making a video available to users for viewing purposes or for downloading is permitted, that video is allocated a unique citation link (DOI name)[12]. DataCite registers the DOI via the API interface. In addition to carrying out DOI registrations of films, the AV Portal also offers a time-based citation link. Using the open standard Media Fragment Identifier (MFID)[13], a citable DOI is displayed for each film segment.

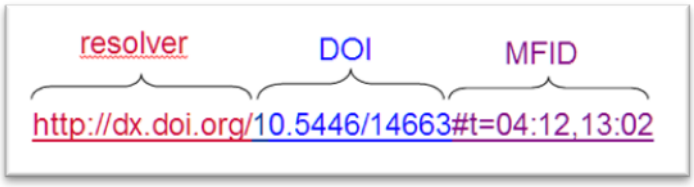


Figure 3: Digital Object Identifier allows for a precise citation of a videosegment

In light of the rapidly increasing number of digital AV media and the necessity to index them at the segment level, solutions for automatic indexing are very much needed, because this is not manageable manually (Neumann and Plank, 2013). The workflow for automatic video analysis in the TIB AV Portal includes the following steps:

First, the video is automatically segmented at the clipping boundaries on the basis of image characteristics. Key frames are extracted from the segments to create a visual index. After completion of this structural analysis, text overlays (e.g. on slides) are analysed using intelligent character recognition and stored in the form of a transcript. Likewise, a transcription is generated from spoken language using automatic speech recognition (Strobel and Plank, 2014). In the next step, visual concept detection classifies visual content by means of predefined specialised and generalised categories such as landscape, machine, drawing, animation and lecture (Blümel et al., 2012).

Within the TIB AV Portal, named entity recognition extracts terms listed in the **Integrated Authority File** (GND)[14] from audio transcripts and text overlays, meaning that the video is semantically tagged with keywords. The tags define entities of an ontology that are linked in semantic relations such as synonymy, hyperonymy and hyponymy. Videos from TIB's subjects – technology, architecture, chemistry, computer science, mathematics and physics – are automatically tagged at the segment level with the corresponding GND tag. Semantic searches can be conducted using entities (GND tags). Entities have main identifiers (e.g. Kernenergie), synonymous identifiers (Nuklearenergie, Atomenergie, etc.) and in some cases they have English identifiers (Nuclear Energy). Some entities are additionally associated with sub-categories. When searching for the term Kernenergie, for example, the system also searches for all other identifiers (synonyms, English translations) and any sub-categories of the entity

[10] **http://www.zbmed.de/**

[11] **http://www.zbw.eu/de/**

[12] TIB: DOI Service (see note 12).

[13] **https://www.w3.org/TR/media-frags/**

[14] **http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html**

Kernenergie. This way, the number of relevant video documents returned is expanded considerably.

The English identifiers were obtained by mapping GND entities onto data from other standards. These standards include DBpedia[15], Library of Congress Subject Headings (LCSH)[16], mappings from the Multilingual Access to Subjects (MACS) project[17] and the WTI "Technology and Management" thesaurus[18].

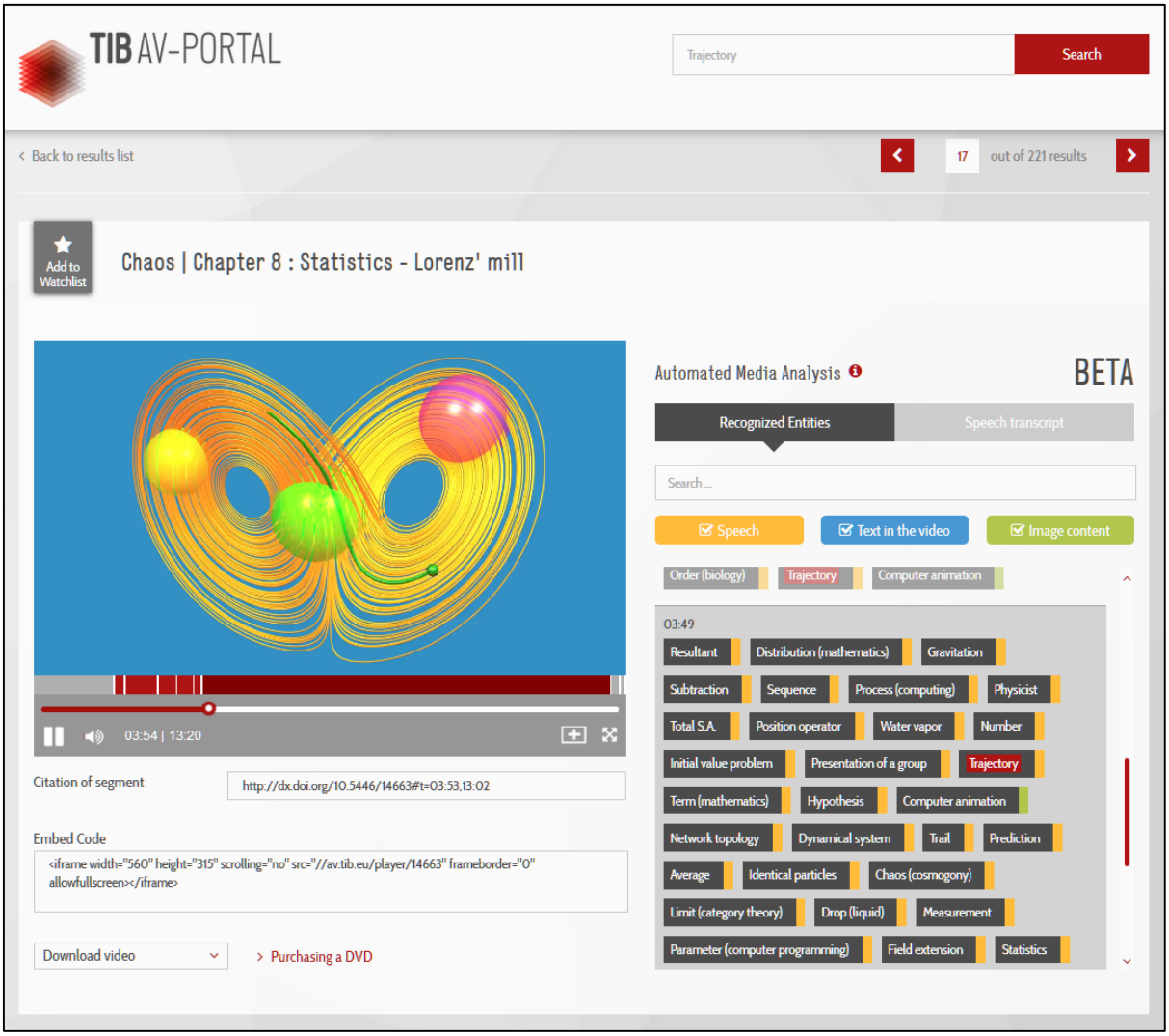


Figure 4: Detail page of TIB's AV Portal. The automatic, content-based indexing of individual videos at the segment level enables users to search in an accurate, content-based manner.

[15] http://wiki.dbpedia.org/

[16] http://id.loc.gov/authorities/subjects.html

[17] http://www.dnb.de/DE/Wir/Kooperation/MACS/macs_node.html

[18] https://www.wti-frankfurt.de/images/themenpakete/english/en-tema.pdf

Thanks to automated, semantic video analysis, TIB's AV Portal offers cross-lingual, content-based access at the segment level, improving keyword-based search by tagging materials with entities. In a traditional keyword-based search, only those documents that contain the particular search term entered are returned. In contrast, semantic search takes advantage of its knowledge basis and, beyond that, can return video documents containing, e.g. synonyms, hypernyms or sub-categories of the search term entered. As a result, the comprehensiveness of the relevant video documents will be reflected in the increased number of hits. The search results of the AV Portal can also be specified thanks to content-based faceted navigation. The AV Portal contains facets for subject area, publisher, year of publication, licence, terms found in the video, images and organisations. The user starts with a textual search in the AV Portal and then refines the search results continuously by means of facets.

Services for AV

Producers of scientific films can also simply upload their videos via an online form to the TIB AV Portal free of charge, or ask TIB for an FTP access (file transfer protocol). The quality of the video is evaluated, hosted, published in a legally sound manner, indexed according to international standards, semantically enhanced, transcribed, digitally preserved and finally given a DOI name – all this optimises the material's discoverability (Löwe et al, 2015).

To make audiovisual media available beyond the AV Portal itself, TIB has made the metadata and preview files it has licensed available to partners such as EUROPEANA[19], the Deutsche Digitale Bibliothek[20] and the Deutsches Filminstitut (German Film Institute)[21], as well as to many other institutions. Further expansion of cooperative activities is underway.

Recently, TIB has started to publish authoritative as well as time-based, automatically generated metadata and thumbnails of films, for which a use has been agreed under the **CC0 1.0 Universal** licence, as linked open data for further use in the standard RDF format.[22] In the future, the datasets will be updated quarterly. In addition, users can attend a tutorial at https://av.tib.eu/opendata, providing a brief overview of the structures of the datasets of TIB's AV Portal. The tutorial explains how datasets can be imported into an RDF database and searched via SPARQL (Marin Arraiza and Strobel 2015).

Finally, TIB provides advice on all matters concerning publishing media, including advice on technology, rights, metadata, digital preservation and DOI registration.

Conclusion

With its AV Portal, TIB has provided a reliable infrastructure for scientific videos, and thus a valuable service. This infrastructure includes hosting capabilities, metadata enhancement, and media-specific search and retrieval tools as well as digital preservation. A DOI name

[19] **http://www.europeana.eu/portal/**

[20] **https://www.deutsche-digitale-bibliothek.de/**

[21] **http://www.filmarchives-online.eu/**

[22] **https://av.tib.eu/opendata**

and/or a Media Fragment Identifier (MFID) guarantees the citability of entire videos, video abstracts, or even individual video segments. In this way, researchers are able to reference and share video content, for example in teaching or in social networks. Recordings of conferences can be specified in lists of publications and identified, further used and cited by other scientists. In this way, scientific videos become reliable sources and can be preserved as cultural heritage.

References

PLANK, M., 2012. *LIBER Blog: A portal for scientific audiovisual media: analyzing user needs.* Avaliable from: **http://libereurope.eu/blog/2012/10/03/a-portal-for-scientific-audiovisual-media-analysing-user-needs/.**

BLÜMEL, I., Ch. HENTSCHEL and H. SACK, 2013. *Automatic Annotation of Scientific Video Material based on Visual Concept Detection*. Proceedings of i-KNOW 2013, ACM, 2013, article 16. Avaliable from: http://**dx.doi.org/10.1145/2494188.2494213.**

KRAFT, A., P. LÖWE, M. PLANK, L. HELLER and B. DREYER, 2015. *Preserving the long tail in a big data world: Frameworks for e-infrastructures in research libraries*. Proceedings 2015 PV Conference, Darmstadt, Germany, 3-5 November 2015. (To be published)

LÖWE, P., M. PLANK and P. MARÍN-ARRAIZA, 2015. Acquisition of audiovisual Scientific Technical Information from OSGeo by TIB Hannover: A work in progress report. In: *Geomatics Workbooks n° 12,* FOSS4G Europe Como (Italy). Avaliable from: **http://geomatica.como.polimi.it/workbooks/n12/FOSS4G-eu15_submission_148.pdf**.

NEUMANN J. and M. Plank, 2013. TIB´s Portal for audiovisual media: New ways of indexing and retrieval. In: *IFLA WLIC 2013*. Available in: **http://library.ifla.org/92/1/124-neumann-en.pdf.**

SNOEK, C. G. M., B. HUURNINK, L. HOLLINK, M. DE RIJKE, G. SCHREIBER and M. WORRING, 2007. Adding Semantics to Detectors for Video Retrieval. In: *IEEE Trans. Multimedia*. **9**(5), pp. 975-986. Avaliable from: **http://dx.doi.org/ 10.1109/TMM.2007.900156**.

STROBEL, S. and M. Plank, 2014. Semantische Suche nach wissenschaftlichen Videos: Automatische Verschlagwortung durch Named Entity Recognition. *Zeitschrift für Bibliothekswesen und Bibliographie.* **9***(*4-5), pp. 255-259. Avaliable from: **http://dx.doi.org/10.3196/18642950146145154**.

MARÍN ARRAIZA, Paloma and Sven STROBEL, 2015. The TIB|AV Portal as a Future Linked Media Ecosystem. In: *Proceedings of the 24th International Conference on World Wide Web*. P. 733-734. Avaliable from: **http://dl.acm.org/citation.cfm?id=2742912.**

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

ONLINE SUBJECT SEARCHING OF DISSERTATIONS

Eva Bratková

brt@cuni.cz

Charles University, Institute of Information Studies and Librarianship, Czech Republic

Abstract

This paper evaluates searching for doctoral dissertations by subject in various online systems. The situation in the Czech Republic is introduced, including the problems involved in completing successful subject searches for dissertations – is it possible to find all relevant materials, or is it sufficient just to find something? The Czech situation is then compared with how systems abroad, particularly in the United States, are being implemented with new access routes for dissertations in the form of linked open data, in which controlled vocabularies of subject terms figure prominently. The paper also discusses how selected European systems whose dissertations are already presented in the WorldCat database will cope with a challenge: „... Over time, these references [for topic entities] will be replaced with persistent URIs to... Linked Data resources“?

Keywords

Dissertations, Bibliographic Records, Metadata, Subject Control, Abstracts, Keywords, Controlled Vocabularies, Information Retrieval, Digital Repositories, Linked Open Data

Introduction

Doctoral dissertations are some of the **most important information outputs** from universities. They contain valuable content and include the results of scientific research. Dissertations are a source that should be – like scientific books – properly bibliographically processed, including the subjects. As American librarians also often state, it is very much worth

preparing the widest and best possible access to dissertations, and so they prepare full catalog records of them, including subject terms, using controlled vocabularies (subject heading systems, classification schemes, etc.) (Middleton, 2015, p. 235; McCutcheon, 2008, p. 51). Well prepared dissertation records can then today be effectively linked through semantic web technologies in the form of open linked data. The linking of dissertation records with **subject authority records** is essential for today's users.

The bibliographic registration of dissertations at national or international level has a long tradition in many countries. Also typical for the new millennium is their effective storage in digital repositories to make them easily searchable online and, especially, accessible. The metadata that accompanies dissertations in local digital systems then also plays an important role in services at national and international level, into which they are easily harvested. The online searching for dissertations using formal bibliographic data (author, title, language, university, assigned degree name, etc.) is usually problem–free and, as a rule, also successful. There is, however, a problem with **searching for dissertations by subject**. If the search is dependent on an index composed solely of terms from the titles of the dissertations, from the authors' keywords and the abstract, or possibly also from the texts themselves, the **recall** of the search is usually low (while **precision** may also be high), as shown by the results of earlier well–known research [1]. Collections of dissertations and the records about them continue to grow, and searching by subject is hence becoming a more complicated problem, in particular from the perspective of professional searchers. The end user as a rule always finds "something" and then selects "something" from the offered list without realising that there are many more highly relevant dissertations in the repository.

For the reasons described above, this paper focuses exclusively on **dissertations** prepared within the framework of doctoral programmes. Other types of theses (bachelor's, master's, etc.) are not analysed or assessed, even though they also tend to be, in particular in low–output countries, registered in both local and national systems.

Searching for doctoral dissertations by subject in the Czech Republic

In the Czech Republic today, it is not possible to search for records of all defended dissertations from a single location, nor those defended in the past 10 years (2007 to 2016), meaning the period when they are already mandatorily publicly accessible everywhere. Using existing information services, it is not even possible to establish how many dissertations were actually successfully defended in individual years. The national services *Theses.cz* (**http://theses.cz/**) and *NRGL* (the *National Repository of Grey Literature*, **http://nusl.cz**) report approximately the same annual additions – around 1,000 defended dissertations. Yet the real figure is probably higher, as not all dissertations are registered at national level, especially from large universities. There could be over 2,000 defended dissertations a year.

[1] BLAIR, David C., MARON, M. E. An evaluation of retrieval effectiveness for a full-text document-retrieval system. In: *Communications of the ACM*. 1985, **28**(3), 289-299. ISSN 0001-0782 (Print). ISSN 1557-7317 (Online). Available commercially from: **http://dx.doi.org/10.1145/3166.3197**. Freely available as a report from: **http://hdl.handle.net/2027.42/35415**.

Years of effort[2] to create complete national registration have still not been successful. The *NRGL* service has the highest number of dissertation records in its database: in a declared retrospective covering 1922 to 2016 it had – in September 2016 – a total of **18,304 records**, while for 2007 to 2016 it has **7,496**. The *Theses.cz* service has a slightly lower number of dissertation records in its database for the 2007 to 2016 period (**7,410** in September 2016), whereas the service has **7,716 records in total**. However, we can note that the two services do not carry out registration in the same way (see Table 1) – some dissertations are the same, some are different. Many dissertations are missing and can only be searched for in university databases. It is up to the individual universities as to which service they send their records to, or have them automatically harvested from the repositories. In the Czech Republic, there is currently no legislation for mandatory national registration, and voluntary reporting is not typical in this country.

Searching for dissertations by subject in the Theses.cz system

The national register of theses – *Theses.cz* – managed by Masaryk University, serves, inter alia, for searching for records of these works and the works themselves. The user interface has not changed greatly for some years, is not particularly user friendly either when formulating queries or when evaluating the resulting lists of records and the presentation of full records. It offers searches by several formal data types (year of defence, name of university, etc.). Searches by type of work is surprisingly missing, this being replaced by a view of assigned degree names (however, the results might not be accurate).

Figure 1: "Cloud" of alphabetically ordered common keywords in the Theses.cz system

[2] Efforts since 1990, intended to build upon the earlier complete bibliographic registration of dissertations in the Czech Republic and also in the Slovak Republic: *Bibliografický katalog ČSSR. České knihy. Zvláštní sešit. Československé disertace …* 1964-1978. Praha: Státní knihovna ČSSR, 1965-1981. ISSN 0323-1763. -- *Bibliografický katalog ČSSR. České knihy. Zvláštní sešit. České disertace …* 1979-1988. Praha: Státní knihovna ČSSR, 1983-1990. ISSN 0232-041X.

Searching for dissertations by subject however brings its own problems. The system basically offers a single option: search by **keywords** (simple terms or phrases). The user must enter their query into a field in a simple interface. An index of alphabetically ordered keywords is not available. Some compensation is the option to browse a small quantity of highly common keywords selected from the whole collection (however, all types of theses are included in this case). The set is presented in the form of an impressive "cloud" – see Figure 1 – however, it cannot be restricted to dissertations only. Clicking on a selected term produces an extensive list of resulting short records of all types of theses, in which it is difficult to identify and select highly relevant dissertations.

Ordinary users have perhaps over the years become used to a situation in which they enter a simple keyword the system uses to present a list of selected records, however definitely not all of which are relevant in terms of content. We can illustrate this using the example of the "digitální knihovny" subject (in Czech). Entering this query using various synonymous or variant terms (something an end user would however probably not usually do) led, in September 2016, to different results each time: "digitální knihovny" (466 records), "digitální knihovna" (up to 532), "elektronická knihovna" (184), "elektronické knihovny" (145), "digitální repozitář" (105), "digitální repozitáře" (20), "digitální archiv" (252), "digitální archivy" (27), "institucionální repozitář" (26) and so on. Assembling these together is difficult for a user – the manual processing of long lists of records is difficult, as is evaluating the level of content relevance for each record. An attempt to search for works on the subject of the actual *Theses.cz* system itself ("theses.cz") led, in September 2016, to 993 records, however most of these records were irrelevant.

As a bonus, the *Theses.cz* system offers searches for works on a "related topic" (related subject) in the form of a reference in searched short records. This is a computer calculation based on the presence of common keywords in records of works. The obtained set of records of related works is, however, also difficult to evaluate manually as regards content relevance – the user tends to get lost in the list, and if they click more than once, they are completely lost. It is difficult to discover highly relevant works.

The *Theses.cz* system, which is based, as regards subjects, only on author keywords and abstracts from university repositories or study systems, still has difficulty providing high quality and effective dissertation searches according to subject. Linking records to subject terms in any controlled vocabularies is not yet possible.

Searching for dissertations by subject in the NRGL system

The *NRGL* system is managed by the National Library of Technology (NTK). Up to 75% of the *NRGL* system is composed of records of all types of theses from Czech universities (Charvátová, 2016, p. 91). The central user interface (FAST software), built over the actual *NRGL* repository (Invenio software), is modern and user–friendly, and has text filters for comfortable navigation. As regards dissertations, this service also offers searching by several types of formal bibliographic data (author, name of university, language, type of work etc.) Limiting searches by year of defence is configurable using an impressive timeline. In addition to browsing and combining lists of data, it is also possible to directly enter queries into the system.

Searching for dissertations by subject is also in this case based only on author keywords, which are part of the records harvested from university repositories or databases of study systems of selected universities. The efficiency of the searches is again not optimal in this service. The terms representing the subjects are not controlled. A small check test performed by searching using the subject "digitální knihovny" (in Czech) by entering various synonymous or variant terms also led in this case to different results each time (in September 2016): "digitální knihovny" (346 records), "elektronická knihovna" (350), "digitální repozitář" (61), "digitální repozitáře" (22), "digitální archiv" (47) and "institucionální repozitář".

Efforts to improve the state of affairs in terms of the subject description and online searching for works by subject led a *NRGL* workplace to carry out a small survey of the state of author keywords with the objective of finding whether it would, for example, be possible to map them to the Polythematic Structured Subject Heading System (PSH)[3], which is managed and used for subject description at the NTK as a controlled vocabulary. In addition, it is also prepared in the form of linked open data. The results so far have shown that the mapping would be significantly ineffective (Charvátová, 2016, p. 91-95).

Unless the description of dissertations in source digital systems of universities – from which metadata is harvested for the NRGL – changes or improves, the effectiveness of searches by subject will not significantly change either. The question is whether it would be better to harvest bibliographic metadata from the catalog databases of universities instead, as these pay more attention to subject description in the majority of cases at professional level. What is preventing this?

Description of dissertation subjects in local systems of Czech universities

The unsatisfactory results of searches for Czech dissertations by subjects in both national services are caused by the fact that the source records for the dissertations come from digital repositories of universities or from databases of their study systems, while the description of the subjects was provided by the authors themselves (keywords and abstracts). This means that searches by subject in local repositories also show the same parameters described above for the national services. Repositories also offer users (in the DSpace application or elsewhere) under the label "Browse by subject" (or other label) sometimes longer lists of terms referring to the same thing – see the sample terms on the subject "informační systémy" (in Czech) from the otherwise great–looking repository of the Brno University of Technology (BUT).

[3] *Polytematický strukturovaný heslář (PSH)* [online]. Praha: Národní technická knihovna, 2000- [cit. 2016-09-26]. Available from (URI): **http://pshmanager.techlib.cz/**.

Figure 2: The index of "subject headings", or author keywords on the subject "informační systém" ("information system" in English) in the BUT repository (English equivalents are in a different part of the alphabet)

For the purposes of this paper, a small survey was performed of the **description of the subjects of dissertations** defended at selected Czech universities[4], including outside digital repositories. The objective was to establish if subject description is better elsewhere, if the searchability of dissertations by subject is better elsewhere, and whether there is a chance – in the near future – to get linked records about them in the context of the semantic web. As Table 1 shows, the description of dissertation subjects is performed in **institutional repositories (IR)** exclusively through author abstracts (AB) and keywords (KW). Yet both elements – or even one of them – are however not usually present in many records at all (the policies of the universities differ in this area). At two universities, repositories either do not exist or do not include dissertations (Czech University of Life Sciences (CULS) Prague and Palacky University (PU) Olomouc), or the records are only available in the databases of the study systems.

An interesting although disparate approach to author abstracts and keywords is, however, adopted by the **centralised catalogues of university libraries**[5]. 6 of the 14 catalogues take abstracts from the repositories, while up to 9 catalogues take keywords. Author keywords are, however, not localised uniformly into the MARC 21 format field, some catalogues are localised

[4] The 14 universities that have reported the highest numbers of defended dissertations in recent years. The numbers of dissertations were taken from both national services, and also from databases of local systems, including the central catalogues.

[5] Of the 14 universities, 3 do not catalogue dissertations at all, however rare records do sometimes appear (see Table 1).

into field 690, and the majority into field 653 (Index Term-Uncontrolled)[6]. The Masaryk University catalogue also uses the local field M530. Keywords are usually displayed in OPAC records, however are not, for understandable reasons, filed into the indexes of subject terms.

What was confirmed through the survey as pleasing and hopeful was that the majority of universities take great care when describing the subjects in their catalogues. Sometimes the processing is actually very detailed. 8 of the 14, respectively 11 universities, use one of the well–known domestic **controlled subject heading systems or thesauri** (see Table 1). This is primarily the Czech National Subject Authority File of the National Library of the Czech Republic (CZENAS), the Polythematic Structured Subject Heading System (PSH), and also the CZMESH subject heading system (for medicine). Other systems are also used (Czech theological thesaurus, CTT), and also local subject heading systems (CZ-BrMU, PHFFUK, MFF etc.) In the case of the PSH, there is the possibility of immediate linking of the catalog records to the PSH, which is already displayed in a structured form (the subject headings have URI identifiers), while with CZENAS there will probably be the possibility to display it in the form of linked open data in the near future, as can be anticipated from the conclusions of the INTERPI[7] research project. Up to 9 catalogues use the **UDC classification scheme** (Universal Decimal Classification) to describe dissertation subjects. In this case, there is definitely a possibility of an open linking of records to records of all UDC numbers. For now, there is a set of selected main UDC numbers available in a linked open data structure (**http://udcdata.info/**). For example, the class of medical sciences has a URI: **http://udcdata.info/037318**. Access to the whole UDC set (**http://cz.udc-hub.com/cs/login.php**) is available in Czech after registration. In the future, when record volumes will grow quickly, the orderliness of dissertations according to this universal classification will definitely be of benefit.

[6] In one case a curiosity was discovered: author keywords in catalog records of the University of Economics, Prague, are localised in field 690, yet in the NRGL database that collects these records, they are then offered for export in field 653.

[7] The headquarters of the INTERPI project is at the URL: **http://autority.nkp.cz/interpi**.

	IR		LIBRARIES CATALOGUE					THESES.CZ		NRGL		BASE			OpenAIRE		DART-EU	
	AB	KW	AB	KW	SH	UDC	conspectus	AB	KW	AB	KW	AB	KW	DDC	AB	KW	AB	KW
CULS, Prague			—	—	CZENAS	UDC	conspectus											
CTU, Prague	AB	KW	—	—	PSH	UDC	conspectus			*AB*	*KW*							
USB, Čes.Bud.			—	KW	CZENAS	UDC	conspectus	AB	KW	AB	KW				AB	KW		
MU, Brno	AB	KW	AB	KW	CZENAS CZ-BrMU CZMESH ET AL.	UDC	conspectus	AB	KW									
MENDELU	AB	KW	AB	KW	—	—	—	AB	KW	AB	KW				AB	KW		
TU Liberec	AB	KW	AB	KW	CZENAS	—	conspectus	AB	KW	*AB*	*KW*							
Charles Univ.	AB	KW	AB	—	CZENAS PHFFUK CTT, MFF CZMESH ET AL.	UDC	conspectus											
PU, Olomouc			AB	KW	—	UDC	—	AB	KW	*AB*	*KW*							
UPA	AB	KW								*AB*	*KW*	AB	KW	DDC				
TBU, Zlín	AB	KW	—	KW	CZENAS ECZENAS ET AL.	UDC	conspectus	AB	KW			AB	KW	—				
VŠB-TUO	AB	KW						AB	KW	*AB*	*KW*	AB	KW	DDC				
UE, Prague	AB	KW	AB	KW	CZENAS	UDC	—	AB	KW	AB	KW	AB	KW	DDC	AB	KW	AB	KW
BUT, Brno	AB	KW	—	KW	PSH	UDC	—			AB	KW	AB	KW	DDC	AB	KW		
UWB, Pilsen	AB	KW						AB	KW	*AB*	*KW*	AB	KW	—				

Table 1: Overview of data from the subject description of dissertations from selected universities in the Czech Republic in various information systems and selected national or international services [Legend: IR = institutional repository, AB = abstract, KW = keyword, SH = subject heading system]

7 catalogues also use the well–known CONSPECTUS categorisation scheme.

It is gratifying that Czech dissertations have the possibility of being better searchable in terms of subject, even if only through OPACs. Unfortunately, for now only separately in each catalogue. Records of Czech dissertations are not currently placed – unlike in other developed countries – into the **Czech national union catalogue (CASLIN)**. This is a great pity, as we could then have access to the whole collection of dissertation records from a single location. Another theme could be the idea of enriching the records in the repositories of universities with certain data from catalogues. The idea is logical, yet no practical solution can be anticipated at this time.

Dissertations and database of the Czech National Bibliography

In the Czech Republic, the nationwide registration of all defended dissertations was conducted in the past. The traditional system disappeared after the cancellation of the applicable legislation (after 1990). It is a pity that this registration could not continue, as similar systems successfully operate in other countries.

As part of the survey of the processing of dissertation subjects in catalogues, the database of the Czech National Bibliography (CNB, **http://aleph.nkp.cz/F/**) was also checked. It is true that the CNB today registers a number of dissertation autoabstracts, as they are officially published at some universities[8] (with ISBN identifiers). The CNB system gets them as legal deposit. They are subject to subsequent high quality processing in the CNB database,

[8] For example, Tomas Bata University in Zlín, Brno University of Technology, VŠB - Technical University of Ostrava, etc.

including detailed subject and systematic description [9]. For example, the record of a dissertation autoabstract by author Petr Maršálek entitled *"Únavové zkoušky ozubených kol = Gear fatigue tests"* (ISBN 978-80-248-2991-3) obtained up to 4 UDC numbers and, in addition to formal subject headings, also two subject headings from the CZENAS[10] controlled vocabulary. Thanks to cooperation between the National Library of the Czech Republic and OCLC, the record is already also in the WorldCat database (**http://www.worldcat.org/oclc/855464854**) and is also currently available on the web in the linked open data structure (in the schema.org and RDFa standards). The record of the autoabstract of the quoted dissertation as a work is freely available, and has a URI identifier: **http://worldcat.org/entity/work/id/1374607131**. The record of this work will be linked to the records of Czech controlled subject headings including their English equivalents (see the block "about" in Figure 3), as soon as the National Library of the Czech Republic issues them as linked data with the relevant URI identifiers. Currently, when you try to click on it the following English message appears: *"This is a placeholder reference for a Topic entity, related to a WorldCat Entity. Over time, these references will be replaced with persistent URIs to... and other Linked Data resources"*.

Figure 3: Part of the record of a Czech dissertation autoabstract as a work from an experimental OCLC linked open data server

[9] The cataloguing of published dissertation autoabstracts is also performed at the level of university catalogues.
[10] A complete record of the autoabstract of the given dissertation is in the CNB database, available at URL: **http://aleph.nkp.cz/F/?func=direct&doc_number=002466242&local_base=CNB**.

We can currently only regret that Czech dissertations are not processed in the system of national bibliography. High quality records of them could form part of important international systems, including the Google Books system, which takes records from the WorldCat catalogue. This would improve their searchability several fold, including from the mobile devices of end users.

Searching for doctoral dissertations by subject abroad

In the area of processing and searching for dissertations by subject, the situation abroad is substantially better than that in the Czech Republic. In many countries, the operation of various systems at both national and supranational level for their processing and online searching from a single location, including access to their texts, is a matter of course. The processing of subjects using controlled vocabularies of the subject headings schemes or classification schemes, which are in many cases already issued free of charge in the form of linked open data, is commonplace. Their searchability is undoubtedly at a high level, and also through the Google search engine. The situation in the United States and in Germany is concisely characterised below as an example.

United States

The United States (U.S.) has two systems today, and their databases have a similar number of dissertation records (Procious, 2014, p. 144). All doctoral dissertations are registered.

The first is the **PQDT (ProQuest Dissertation & Theses)** database operated by ProQuest. It holds around 3,500,000 dissertation records and also master's theses. All doctoral dissertations from the United States are registered (around 40,000 doctoral dissertations are defended annually). Around 700 universities are involved. The bibliographic records in the PQDT database are very detailed as regards their formal description, while ISBN identifiers are also assigned. Online searches for dissertations are provided by ProQuest in the FAST search system (a new one is under preparation). As regards subjects, the record is furnished with an author abstract, while ProQuest professionals classify dissertations into their own universal categorisation scheme (ProQuest Subject Categories). Lexical equivalents of the categorisation are in addition also localised in the subject headings field. The record also contains author keywords as a rule. Thanks to the cooperation with universities, it is possible to transfer records from the PQDT database (selected useful data) into university catalogues, where they are subject to high quality processing, in particular as regards the dissertation subject. Subsequently the records are transferred to the WorldCat database (OCLC). The searchability of dissertations by subject is further increased in some professional bibliographic databases (PsycINFO, ERIC etc.), into which ProQuest transfers its records by agreement. In them, the records are further supplemented with more detailed subject terms from thesauri and numbers from special classification schemes.

The second database in the United States that contains a large number of dissertation records is the **WorldCat** database. The records are sourced from university catalogues, are furnished with Library of Congress Subject Headings (LCSH) and automatically also headings from the new FAST system (OCLC) and also LCC or DDC classification numbers. Thanks to the technological maturity of the OCLC, the records of these dissertations are today issued in the form of linked open data, while the linking to subject heading systems is actually live.

For example, a dissertation by author N. R. Kale entitled *"A case study on robustness of Dynamic Time Warping for activity recognition using wearable computers"* (URI: **http://www.worldcat.org/oclc/846506856**), defended at the University of Texas in Dallas in 2012 has, in addition to the abstract, up to 8 subject headings linked from the record of the work (the URI of the dissertation in question as a work is **http://worldcat.org/entity/work/id/1356254544**) to the records of relevant subject heading systems available on the web: LCSH (for example the heading "Dynamic programming", URI: **http://id.loc.gov/authorities/subjects/sh85040313**) and FAST (for example the heading "Wireless sensor nodes", URI: **http://id.worldcat.org/fast/1750044**).

The searchability of dissertations from the United States by subject is thus at a high level, even under the conditions of a growing semantic web.

Germany

Searchability by subject for all dissertations defended at German universities is similarly favourable. Thanks to a law on legal deposit of dissertations, they are now completely registered in the German National Bibliography system. The records are accessible from the catalogue of the German National Library, meaning also from a single location. The publication of the records excels through its precision – all physical records of essential "things" (objects) already have unique URI identifiers, including records of subject headings and name authority data. The dissertation subject is represented primarily through the controlled Subject Headings Authority File (SWD – Schlagwortnormdatei), which is today part of the Integrated Authority File (GND – Gemeinsame Normdatei), also already issued in the linked open data structure. An example can be a dissertation by author Florian Klein entitled *"Metadaten-Verwaltung in einem verteilten RAM-basierten Speicherdienst"*, defended at the University of Düsseldorf in 2015 (**http://d-nb.info/1079652574**). The dissertation is assigned a URN identifier (access to the text is possible from two locations), the record is primarily linked to the national authority record of the author (**http://d-nb.info/gnd/187515727**), the records of two referees (**http://d-nb.info/gnd/1018114750** and **http://d-nb.info/gnd/122535316**) and a total of 6 records of controlled subject headings (in German): Speicher <Informatik> (**http://d-nb.info/gnd/4077653-0**), Metadaten (**http://d-nb.info/gnd/4410512-5**), Verwaltung (**http://d-nb.info/gnd/4063317-2**), Dienst <Informatik> (**http://d-nb.info/gnd/4835035-7**), Speicherverwaltung (**http://d-nb.info/gnd/4182146-4**), Anwendungssoftware **http://d-nb.info/gnd/4120906**). In addition, there is also the classification number of the German National Library classification scheme. See Figure 4.

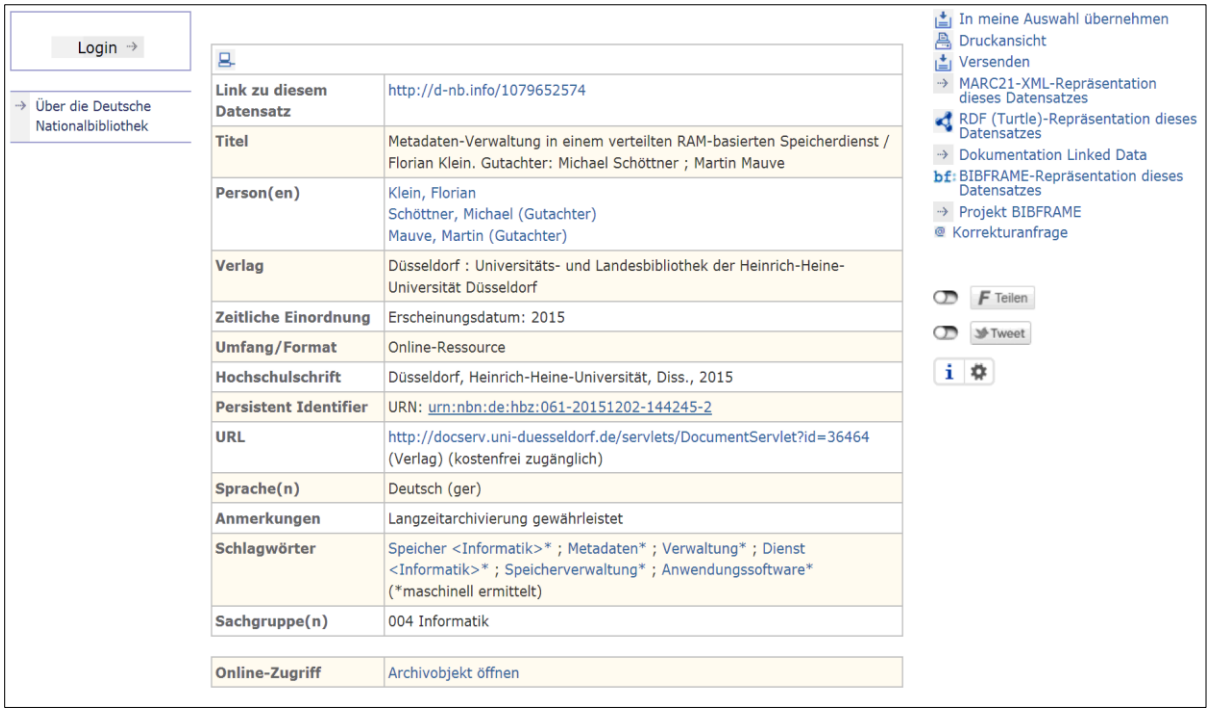


Figure 4: Record of a German dissertation in the catalogue of the German National Library with a URN identifier and also with hyperlinks to GND (Gemeinsame Normdatei) subject heading records

Thanks to cooperation with OCLC, records of German dissertations are also available in the WorldCat database, and are therefore also available in the open linked data form. Clicking on German national subject headings still results in the message quoted above in the text for the example description of a single Czech dissertation autoabstract. In fact, this message is no longer completely true, as German subject headings already have URIs, and it is anticipated that they will soon be correctly linked through the OCLC experiment.

Conclusion

The digital repositories at Czech universities, as well as the two national services that register and make dissertations available, are developing successfully and growing in terms of data volumes. The records of dissertations from some universities are also already transferring to international services (BASE, OpenAIRE and DART-Europe – see Table 1). Yet they are still lacking something fundamental, namely better and higher quality access to searches for dissertations by subject. With the growing volumes of records and dissertations themselves, there is the danger of serious losses in terms of quality and effective searching in the future. Just finding "something" from a specific quantity of content–relevant dissertations may possibly satisfy the layman (student etc.) at a given moment, yet will not satisfy a professional searcher, whose task it might be to search for all dissertations on a specific subject. A small survey in the Czech Republic showed that there is some hope for improved description of subjects, and this through the catalogues of university libraries. Many libraries (but not all) allocate subject headings, UDC classification numbers and the Conspectus category system. There is a chance that in the future dissertation records may be linked on the web to records of these subject heading systems and classifications. The exposure of some knowledge organization systems (KOS) in the linked open data structure is already a reality, while others are under preparation. This is a powerful technology that could help change the unfavourable Czech situation in the area of dissertation subject descriptions. Can we, for example, hope that it will

be possible – at least in some way – to integrate the descriptions of dissertation subjects in repositories and catalogues separately? Can we hope that records of Czech dissertations will also be entered into the Czech national union catalogue CASLIN? Yet these are questions for a different discussion.

References

CHARVÁTOVÁ, Michaela, 2016. *Sjednocování věcného popisu agregovaných záznamů v repozitáři NUŠL*. Praha. 121 p., 2 suppl. Diplomová práce (Mgr.). Univerzita Karlova, Filozofická fakulta. Available from: **https://is.cuni.cz/webapps/zzp/detail/161853/**.

McCUTCHEON, Sevim et al., 2008. Morphing Metadata: Maximizing Access to Electronic Theses and Dissertations. *Library Hi Tech*. **26**(1), 41-57. Also available commercially from (DOI): **http://dx.doi.org/10.1108/07378830810857799**.

MIDDLETON, Cedar C. et al., 2015. A Process for the Original Cataloging of Theses and Dissertations. *Cataloging & Classification Quarterly*. **53**(2), 234-246. Also available commercially from (DOI): **http://dx.doi.org/10.1080/01639374.2014.971997**.

PROCIOUS, Aaron W., 2014. WorldCat, the Other ETD Database: An Exploratory Study. *The Reference Librarian*. **55**(2), 144-150. Also available commercially from (DOI): **http://dx.doi.org/10.1080/02763877.2014.880276**.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

WAYS OF DISSEMINATING, TRACKING USAGE AND IMPACT OF ELECTRONIC THESES AND DISSERTATIONS (ETDS)

Meinhard Kettler

meinhard.kettler@proquest.com

ProQuest Information and Learning, Germany

Abstract

The digital transformation has had a tremendous impact on graduate research workflows and output. Most theses are submitted as ETDs, although the share varies by country and by subject. Universities worldwide are running institutional repositories to showcase new graduate research, as well as recently digitized material. The presentation will highlight studies on dissertation and theses usage in repositories as well as giving insights into ProQuest's unique dissertations and theses analytics.

Keywords

ETD, Thesis, PhD Thesis, Dissertation, Usage Analytics, Impact, ProQuest

Introduction

It is evident to everyone in and outside academia that theses and dissertations represent a valuable contribution to scholarly research. With their graduate works young researchers show their ability to examine and summarize existing knowledge as well as adding new and unique findings in a detailed study. Not all dissertations and theses mark the beginning of a researcher's career, but for those graduate students that choose to continue their professional life in research the PhD thesis or dissertation will often be the first entry in their personal publication list and therefore deserves special attention. Even if the thesis submission does not open up a professional life in research, it may very well contribute to the relevant literature in a given subject.

In this presentation I will try to summarize what I have learnt about the current situation and trends of dissertations' and theses' visibility and impact, having been active in the field of dissertations dissemination for ProQuest for a bit more than a year, and also draw some personal conclusions. Hence the view comes from outside the academic community, yet from the perspective of a renowned commercial player in the field, collecting graduate research for dissemination for more than 70 years.

Availability and Discoverability

From the user's perspective there are a number of ways to retrieve content in dissertations and theses. As with other communication areas, the transformation from print to digital that began in the 1990s had a disruptive effect on dissertation and theses. Since then, policies and workflows in graduate schools have been amended, so that today the original versions of many graduate works consist of 'born-digital' PDFs and no longer bound copies that can only be found on the shelves of the institution's library.

Today a huge number of theses can easily be found by any interested user on the web. However, if unexperienced users conclude that all the world's graduate output is available online and easily discoverable by Google and other search engines, they might get frustrated when trying to retrieve specific information. Discoverability and accessibility is granted only for a part of the existing output. Furthermore, it can turn out to be extremely cumbersome, if not impossible, to find out how big the invisible part of the iceberg is. When a certain PhD thesis does not show up in the result list of a search engine, nor in the institutional repository (IR), nor at the numerous dedicated aggregator access points for theses, there might be a variety of reasons:

- Author's decision (accepted by the institution)
- Publication embargo requested by the author or due to third party copyrights
- Monograph published under a different title
- Submission in print without electronic copy
- Institutional repository or catalogue not in place
- Limited discoverability due to language
- ...and more

The above cases illustrate that it is impossible to get a full overview of global graduate works. A single international registry or authority for theses simply does not exist and most likely will not be launched in the near or mid-term future. Policies to submit and publish the works linked

to an academic degree vary not only from country to country, but also between universities in the same country and, quite often, even between faculties and departments of one university.

As to print submission, electronic full text theses and dissertations (ETDs) have only been available for the last twenty years. Before that time, submission in electronic form was not possible or at least not supported by the institutional workflows. However, when looking at today's figures, it might surprise that in some developed countries without technical barriers print submissions still play a significant role.

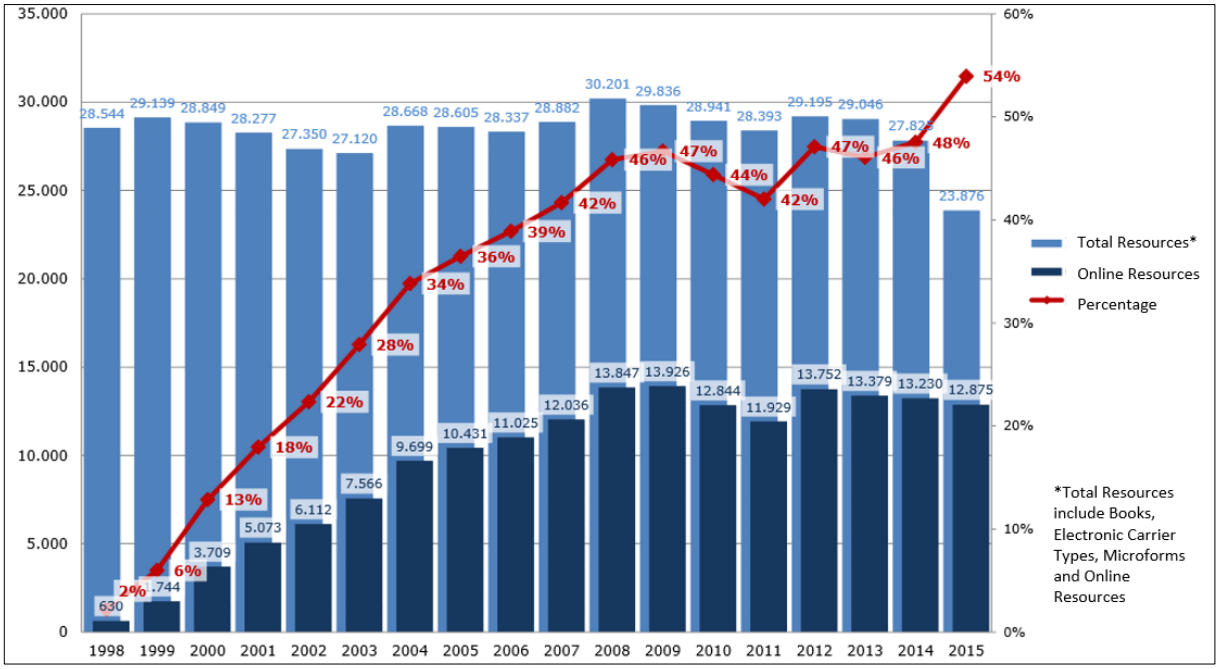


Figure 1: Online Resources Share of Total PhD Dissertations and Habilitations in German National Library Collection by Publication Year (updated 2 March, 2016).
By Deutsche Nationalbibliothek 2016,
http://www.dnb.de/DE/Wir/Kooperation/dissonline/dissonlineStatistik.html

The German National Library publishes statistics of PhD theses (Dissertationen) on its German website (Figure 1). According to the available graph from March 2016, covering the years from 1998 to 2015, the number of PhD theses per year captured in the German National Library catalogue remains stable between 25,000 and 30,000, whereas the percentage of online available items rose from 2 to 54%. Although this looks like a remarkable increase, the remaining 46% "non-online" formats, including print, electronic carrier types and microform, in 2015 still feels very high. As explained on the website, the latest numbers may not yet be exact and complete as a result of technical changes in the harvesting interface, so that files from some universities are missing and will be ingested in the German National Library database only later. Nevertheless, many of us would have expected a much higher percentage for online availability.

For France estimates can be found on the poster *French Electronic Theses and Dissertations in Europe* presented by Hélène Prost and co-authors during the 19th International Symposium on ETDs in Lille, France (Prost, 2016). From the national platform *theses.fr* the authors retrieved the following numbers for French theses production: between 8,000 and 13,000 theses per year were deposited from 2007 to 2015 with an online share of 60% for the last two years. Not only is the percentage of online formats slightly higher than in Germany, but there

is also a clear commitment to an e-only policy. The authors mention a French decree issued in 2016 defining deposit of the digital version of a thesis as mandatory, with the aim to bring print deposit to an end in 2018.

At the same time ETD submission rates from US schools is higher. Acting as official offsite dissertation and theses repository for the U.S. Library of Congress, ProQuest received 93% of US submissions for its database *ProQuest Dissertations and Theses (PQDT)* in electronic form in 2014 (McLean, 2016).

Clearly, the trend goes in the direction of online publishing, especially in the natural sciences. Therefore, it comes as a surprise that in German universities, where students are obliged to publish their theses, doctorate students can choose how to submit and publish their theses. In the long run, the freedom regarding deposits and the lack of standards in the policy-defining faculties might turn into a competitive disadvantage for authors and institutions, negatively affecting visibility and impact of the output.

Dissertations and Theses - Meaningful Content to Showcase

How do theses submitted in digital formats reach their readers? While a global and complex discussion about the possible transition from paid to open access business models for scholarly journal articles and about new ways of funding publishing continues in many countries ETDs have been available on Open Access institutional repositories for a number of years. This offering implies benefits not only for users, but also for authors, whose graduate works have the chance to be more easily discovered, read and perhaps even cited by fellow researchers around the world. The dissemination of research increases the potential for international research collaboration as well as recognition.

Electronic theses represent a high share of content in many institutional repositories. In fact, many repositories were built for the purpose of making ETDs available, and some universities still maintain dedicated repositories or document servers only for ETDs.

In Finland, it was demonstrated that electronic theses represent valuable enrichments to repositories and enjoy surprisingly high usage by public consumers. The Finnish science community has not only managed the transition from print theses to ETDs successfully, but also extended the online availability to (master and bachelor) theses from 25 universities of applied science. More than 100,000 theses are publicly available on the portal *Theseus*, hosted by the National Library of Finland, with 15,000 new items added per year. In 2014 there were almost 18 million full text downloads - more than the total amount of downloads from all other Finnish repositories (Ilva, 2015).

Another example can be found in the repository of the prestigious and research-intensive University College London (UCL). In 2014, of the top 50 documents by downloads, 28 are doctoral theses, including 7 of the top 10.[1] Given these figures there is no reason to underestimate theses as content of "lower" relevance and meaning. For graduate schools, as

[1] **http://discovery.ucl.ac.uk/past_stats/annual-2014.html**

well as for the authors, it is important to ensure visibility and dissemination of their graduate treasures.

Usage and Impact - the Need to Measure and Analyse

In September 2016, the Directory for Open Access Repositories (OpenDOAR) counted 1,790 out of 3,220 repositories that include theses and dissertations as a content type.[2] Both numbers are still growing. Regarding the share of repositories with theses and dissertations we have to keep in mind that not all repositories in this registry are linked to degree-granting institutions and consequently some have no PhD theses to display.

On the other hand, the overwhelming and growing number of individual access points holds new challenges: how can users find their way to relevant theses content when it is distributed over hundreds or thousands of web servers with a variety of user interfaces? Google Scholar as well as aggregator platforms like the global ETD search NDLTD (Networked Digital Library of Theses and Dissertations, **http://www.ndltd.org/**), or the DART Europe E-theses portal (**http://www.dart-europe.eu/About/info.php**), respond to this need by indexing available web content or harvesting metadata from repositories, making use of available standard protocols like OAI PMH. Although not visible to the end-user, there are complex tasks to complete in the back-end, sometimes meaning portals might return incomplete results from their sources. Constantly, administrators are forced to fight data quality and transmission issues from the large variety of dynamically changing content providers.

Next to data quality, getting accurate usage statistics for Open Access repositories also provides challenges for librarians. In most cases the metrics cannot be interpreted 'as is', as usage figures are often skewed or inflated by the activity of robots crawling the open web and not always easily identified. This means the data cannot be taken as 'human' usage and compared to other analytics unless it is filtered and edited.

Some entities have undertaken the effort to create standards and services for comparable usage statistics, such as the British Jisc-funded national aggregation service IRUS-UK (**http://irus.mimas.ac.uk/**) or the initiative DINI in Germany (**https://dini.de/english/**). Most initiatives currently act on a national level.

In the context of theses collection and aggregation, ProQuest, as the first global provider of graduate works, continues to play a crucial role. The US company started microfilming PhD theses under its former name UMI in 1938, mainly for the purpose of long-term preservation. Today *ProQuest Dissertations and Theses Global* covers 99% of US output and output from thousands of universities outside the US, including nearly two million PhD and masters theses in searchable full text. Therefore, it is regarded as one of the premier one-stop-shops with a high share of unique content over all subject fields. Part of the database is openly accessible on the platform PQDT Open (**http://pqdtopen.proquest.com/about.html**).

As mentioned before, research libraries all over the world still have long shelves with enormous amounts of printed theses. How can institutions uncover the ideas in those works and preserve them for the future? In April 2016 Dimity Flanagan and Linda Bennett presented a case study

[2] **http://www.opendoar.org/find.php?format=charts**

on usage and impact of 2,000 theses recently digitized at the London School of Economics in co-operation with ProQuest (Bennett, 2016). Here is a summary of some of their observations:

- Overall download numbers nearly tripled from 2014 to 2015 with the added digitized content.
- Shortly after inclusion of 'new' content downloads per item reach the same level as before.
- Significantly increased overall traffic from social media platforms, including for some of the digitized theses.
- Seeking individual author permission too labour-intensive; a take down policy introduced instead.
- Correlation of usage and citations could not be proved.

The effect of increased usage for old material can also be seen in *ProQuest Dissertations and Theses Global*. Part of the LSE project was to make the newly digitized theses from the LSE available in PQDT in order to gain even more global visibility. As a consequence, in July 2016 an LSE PhD thesis from 1971 digitized as part of this project reached the top 25 dissertation and theses ranking published by ProQuest.[3]

Overall, there seem to be only few current studies that look at the development of usage of ETDS (and print theses) in particular. The impact of theses for future research measured by citations is even more complex to analyse. In order to help reveal connections between research topics as well as between authors, PQDT displays links to cited resources, as well as 'cited by' links. Features of this kind allow for more meaningful analysis of graduate works impact in the future.

ETDs as a Text and Data Mining Resource

The content of individual dissertations and theses can serve as an extremely useful resource for other academics or PhD students in the same field, as it often offers more detail and a more comprehensive literature review than journal articles. In recent times, ETDs as whole sets of big data have received increasing attention as objects of research with new text and data mining methods. At ProQuest our dissertation database grows by 130,000+ items per year, representing a critical mass of valuable academic output covering a wide time range, especially for the United States. This data is an interesting resource for researchers investigating the development and the volume of graduate works in certain subject fields, career paths of graduate students or the general development of language.

As an example may serve Benjamin Miller's doctoral dissertation at CUNY (City University New York) on composition and rhetoric, examining keywords and subject terms of other dissertations. The author analyzed thousands of selected dissertation records including their full texts provided to him by ProQuest (Miller, 2015).

In 2016 a study from the University of Kansas using ProQuest dissertation data as well found in the field of Psychology 80% of the dissertations (PhD theses) could not be linked to peer-

[3] **http://www.proquest.com/products-services/dissertations/ProQuest-Most-Accessed-Dissertations-and-Theses-July-2016.html**

review articles, and remain unpublished after seven years. First of all, this study demonstrates the suitability and importance of dissertation data for research of this kind. Secondly, the study's result tells us again, how important it is not to forget dissertation and theses content when researching specific scientific topics, as journal articles might not give the full picture in all subject fields.

Joachim Schöpfel and his colleagues from Lille examined "dissertations as data" in general and presented their findings at the 19th International Symposium on ETDs in Lille, France. They looked at both the dissertations' texts and the corresponding supplementary material consisting of other data types. All this can be defined as integrated dissertation data and be turned into an object of research. At the same time, the topic highlights the need of best practices regarding deposit and publishing policies for research data, not only in the area of dissertation and theses, but also in scholarly communication in general.

Regarding text and data mining of scientific content, existing grey areas and country specific inconsistencies concerning the legal basis and possible conflict with copyrights should be addressed in order to create a solid environment for better research opportunities.

Conclusion

In the examples and case studies mentioned, I have tried to highlight how relevant graduate works as part of the grey literature are for the researcher as reader and end-user. Nevertheless, with an ever growing number of online available documents, it can turn out to be a tedious task to discover the needed piece of research in the giant information 'haystack'.

For authors of dissertations or theses and for their corresponding institutions multiple channels should be explored and utilized in order to ensure maximum visibility and discoverability of the academic output. Tracking of usage and impact will probably get more attention in the future with new tools and, hopefully, standardized methods.

Peter Suber summarized in *Open Access*: "If a university requires theses and dissertations to be new and significant works of scholarship, then it ought to expect them to be made public, just as it expects new and significant scholarship by faculty to be made public. Sharing theses and dissertations that meet the school’s high standard reflects well on the institution and benefits other researchers in the field. The university mission to advance research by young scholars has two steps, not one. First, help students produce good work, and then help others find, use, and build on that good work." (Suber, 2012).

References

BENNETT, Linda and Dimity FLANAGAN, 2016. Measuring the impact of digitized theses: a case study from the London School of Economics. *Insights*. **29**(2): 111–119. DOI: **http://dx.doi.org/10.1629/uksg.300**.

CORBETT, Hillary, 2016. Out of the Archives and Into the World: ETDs and the Consequences of Openness. In: SMITH, Kevin L. and Katherine A. DICKSON, eds. *Open Access and the Future of Scholarly Communication: Implementation*. Available from **http://hdl.handle.net/2047/D20216388**.

EVANS, Spencer C. et al, 2016. *The Large Majority of Dissertation Research in Psychology Goes Unpublished*. Poster presented at the 28th Annual Convention of the Association for Psychological Science, Chicago, IL, May 2016.

ILVA, Jyrki, 2015. *Repositories and ETDs – a success story from Finland*. Presentation at Open Repositories, Indianapolis. Available from **http://urn.fi/URN:NBN:fi-fe2015061110223**.

MCLEAN, Austin, 2016. *Current Usage of Dissertations: A Global Perspective*. Presentation at the Council of Graduate Schools - Future of the Dissertation Workshop, Washington D.C., January 2016. Available from **http://cgsnet.org/sites/default/files/DissFwd_Print%20All%20Papers.pdf**.

MILLER, Benjamin M., 2015. *The Making of Knowledge-Makers in Composition: A Distant Reading of Dissertations*. CUNY Academic Works. Available from **http://academicworks.cuny.edu/gc_etds/1056**.

PROST, Hélène, Amélie BUIRETTE and Amélie HALIPRÉ, 2016. *French Electronic Theses and Dissertations in Europe – A Scientometric Approach*. Poster Presentation at the 19th International Symposium on Electronic Theses and Dissertations, Lille, July 12, 2016. Available from **https://etd2016.sciencesconf.org/98998**.

SCHOEPFEL, Joachim, Eric KERGOSIEN, Stéphane CHAUDIRON and Bernard JACQUEMIN, 2016. *Dissertations as Data*. Presentation at the 19th International Symposium on Electronic Theses and Dissertations, Lille, July 13, 2016. Abstract available from **https://etd2016.sciencesconf.org/92328/**.

SUBER, Peter, 2012. *Open Access*. Cambridge: The MIT Press. The MIT Press Essential Knowledge Series. ISBN 978-0-262-51763-8. Available from: **https://mitpress.mit.edu/sites/default/files/9780262517638_Open_Access_PDF_Version.pdf**.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

TEXT AND DATA MINING OF GREY LITERATURE FOR THE PURPOSE OF SCIENTIFIC RESEARCH

Matěj Myška[1]

matej.myska@law.muni.cz

Masaryk University, Faculty of Law, Institute of Law and Technology, Czech Republic

Abstract

This paper explores the legal possibilities of users to text and data mine repositories of grey literature for scientific research without the consent of the grey literature repository operator or the holders of rights to the content stored therein. In the first part of this short paper we briefly introduce the relevant intellectual property rights. In the second part, the current exceptions to these exclusive rights are discussed and evaluated. The third part critically analyses the suggested mandatory exception for text and data mining in the European Commission's proposal for a new directive on copyright in the Digital Single Market.

Keywords

Text and Data Mining, Exceptions, Sui Generis Database Rights, Copyright

The publication of this paper is supported by the Czech Science Foundation - project Legal Framework for Collecting, Processing, Storing and Utilizing of Research Data - registration no. GA15-20763S.

[1] I would like to thank Jakub Harašta for critical revision of this paper. However, all mistakes and omissions are mine.

Introduction

As observed by Floridi in 2014, the development of ICTs brought us into the zettabyte era, where tsunami of bytes submerge our environments (2014, p. 13). The amount of data and information produced is constantly growing, as is the number of scientific research papers (i.e. white literature).[2] Even though there are no similar exact data on the rising amount of grey literature, given the fact that GL is usually available online without the traditional constraints of white literature publishing (Banks, 2006, p. 5), it can be reasonably assumed that the number is not declining. Jeffery and Asserson even go as far as proclaiming that the vast majority of research output is grey (2014, p. 223).

On the other hand, this technological development not only brought us the age of data deluge (Borgman, 2012, p. 1059), but also offered researchers various new and innovative tools and techniques[3] to effectively process this vast amount of data and information (including GL)[4] in an automated way. These are commonly labelled "text and data mining".[5] In fact, the whole process of academic research has been reshaped with the technological development, in that TDM can be employed in all of the stages thereof. Technology can thus free up the time that would otherwise be spent on *"finding ideas for a research paper, literature review and formulation, data and methodology and analysis of results"* (Filippov, Hofheinz, 2016, p. 5).

However, the legality of TDM remains unclear under European intellectual property laws[6] and subsequently the national law of Member States. Potential infringement of exclusive rights is presented in the first part of this paper. The second part discusses the current exceptions that are potentially applicable to TDM. The third part introduces and critically analyses the proposed mandatory exception for TDM in the proposal for a new directive on copyright in the Digital Single Market (hereafter the "Proposal").[7]

[2] The number of scientific papers published was estimated at around 50 million in 2010 (Larsen, von Ins 2010). The number continues to grow by 2.5 million a year and is steadily on the rise (Ware, Mabe 2015, p. 6), See also (Bornmann, Mutz 2015).

[3] For an overview of various types of data mining methods and applications from the technological point of view see (Colonna 2013). Colonna also critically notes that this "buzzword" is so semantically obfuscated that it is starting to lose its meaning (2013, p. 309). For further technical details of TDM see, e.g., (Larose 2014).

[4] Due to its variety, GL is an ideal raw material for TDM (Schöpfel, 2010, p. 29).

[5] However, automated processing might also include protected or unprotected content other than text and data and so "content mining" would probably be a more accurate term (especially in the context of GL) (Murray-Rust, Molloy, Cabell 2014, p. 11).

[6] In this paper, we deal mainly with the two most relevant directives, namely Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (hereafter "ISD") and Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases (hereafter "DD"). For further details on European copyright law see, e.g., (Walter, Lewinski 2010; Stamatoudi, Torremans 2014). For treatment of TDM outside Europe see, e.g., (European Commission, Directorate-General for Research and Innovation 2014, pp. 44–48).

[7] Proposal for DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on copyright in the Digital Single Market of the European Commission. Brussels, 14.9.2016, COM(2016) 593 final, 2016/0280 (COD).

TDM as possible infringement of intellectual property rights

The principle of legal license, i.e. that *"everything which is not forbidden is allowed"*[8], forms the basis for further inquiries. Consequently, text and data mining is only relevant from the legal point of view when it encroaches upon protected subject-matter. We further focus only[9] on 1) copyright protection[10] (and specifically the right of reproduction) and 2) the sui generis database right (specifically extraction right).[11] Such choice is justified by the fact that any processing of digital content[12], and implicitly TDM, includes making (at least a temporary) copy[13] of the subject-matter being processed. (Triaille et al., 2014, p. 28). Such copying would involve copyrighted works[14] (contained in, e.g., the GL repository). Moreover, when the structure/arrangement of the database is copied, the copyright of the database creator is infringed. Finally, reproduction of whole/substantial parts of (otherwise unprotected) content of a database[15] is prohibited by the sui generis right of extraction. In order to avoid liability for copyright/sui generis database rights infringement, the miner must rely either on the consent of the holder of rights or on existing exceptions to these exclusive rights.

Potentially applicable exceptions

The first potentially applicable copyright exception is the one that allows for a **temporary copy to be made (Art. 5(1) ISD).** Such a copy must be transient and incidental (not permanent) and for the duration of TDM only (i.e. the protected subject matter being mined must be automatically deleted afterwards). Furthermore, it must be an integral and essential part of the technological process, which should not pose a problem as TDM is essentially a technology built on making copies. Next, the copy must enable lawful use – lawfulness might be based on the application of another exception[16] or "using" the works in a way that is lawful, e.g. mining for new information. Finally, the allowed copy must be without independent economic significance.[17] This exception, with its extensive pre-requisites, is applicable in only a few

[8] As expressed, e.g., in Art. 2(4) of the Constitution of the Czech Republic (Constitutional Act No. 1/1993 Sb., as amended).

[9] For a detailed discussion of legal aspects of TDM see (Triaille et al. 2014; Truyens, Van Eecke 2014); for a concise analysis of EU and Czech Law see (Myška, Harašta 2015).

[10] Including copyright protection for the selection and arrangement of the contents and structure of the database.

[11] TDM techniques could also infringe the rights to privacy and protection of personal data of a natural person. However, these issues are not discussed in this paper. For discussion of this topic see, e.g., (Rubinstein 2013).

[12] As noted by (Truyens, Van Eecke 2014, p. 158).

[13] This copy is usually done in the process of creating corpora for subsequent analysis (Truyens, Van Eecke 2014, pp. 154–155).

[14] Provided that they are protectable subject-matter, i.e. that they are original as stated by the CJEU, e.g., in Infopaq I, C-5/08, ECLI:EU:C:2009:465. For details on originality in EU copyright law see (Rosati 2013).

[15] A database is protected by the sui generis rights if the database owner made substantial investment in obtaining, verifying and presenting the contents of the database (Art. 7 DD).

[16] E.g. the "scientific research" exception specified hereunder.

[17] CJEU stated in Infopaq II, C-301/10, ECLI:EU:C:2012:16, para. 54, that this condition precludes modification of the used protected subject-matter under this exception.

cases (especially due to the non-permanent character of the copies) and does not provide enough legal certainty.

The exception for the purpose of **scientific research** (Art. 5(3)(a) ISD) allows for the reproduction of protected subject-matter solely for that purpose, to the extent justified by the non-commercial purpose to be achieved and under the condition of indicating the source (including author's name), if it is not impossible.[18] This exception provides prima facie solid foundations for performing TDM. However, the non-commercial criterion significantly reduces the scope of applicability, especially in borderline cases where commercial entities participate in the research. Further, TDM relies on the quantity of the mined content – the scientific research exception, however, suggests a rather restrictive approach as reproductions should be made only to the above-mentioned *"extent justified by the non-commercial purpose".*

Apart from the temporary copy exception, the other potentially applicable copyright exceptions in ISD are only facultative "blueprints" and may not be transposed into the respective national law. However, the assessment of the legality of TDM (including TDM of GL) will be dependent on the applicable national law.[19]

Regarding the protection of **databases,** DD foresees a facultative exception for scientific research (Art. 6(2) DD) allowing for the reproduction of copyrightable structure/arrangement thereof. As these pre-requirements do not differ in further details, the short remarks made to the copyright exception for scientific research are also applicable here. Moreover, DD also includes a counterpart of the "temporary copy" exception in Art. 6(1) DD which is also mandatory. This exception allows the lawful user to reproduce the arrangement/structure of the database *"which is necessary for the purposes of access to the contents of the databases and normal use"* thereof. The sui generis right of extraction might also be limited for scientific research under same preconditions (Art. 9(b) DD), but only for the "lawful user".[20] Such subject might also extract unsubstantial parts of the content from the database made available to the public for any purpose (Art. 8 DD). However, the user may not do so in a repeated and systematic manner in a way that would imply *"acts which conflict with a normal exploitation of that database or which unreasonably prejudice the legitimate interests of the maker of the database"* (Art. 7(5) DD).[21] Such an exception is thus practically irrelevant to TDM as it relies on extraction of all or at least substantial parts of the content. Similarly, as far as copyright exceptions are concerned, the national transpositions of the facultative exceptions (including the one for scientific research) differ, which consequently leads to undesired legal uncertainty (Triaille et al., 2014, pp. 80–81).

[18] As the number of works used in TDM is usually fairly high, this safeguard clause could be used.

[19] For an overview of different implementations of ISD see in detail, e.g., (Hugenholtz, Eechoud, Gompel, Helberger 2006; Westkamp, Guibault, Rieber-Mohn 2007; Eechoud, Hugenholtz, van Gompel, Guibault, Helberger 2009; Triaille, Dusollier, Depreeuw, Hubin, Coppens, Francquen 2013). The transpositions differ, e.g., in scope. In BE, LU and IT, only parts of the works might be used, which poses a practical problem for TDM. (Triaille et al. 2014, p. 56).

[20] This concept itself is rather opaque. For a detailed discussion see, e.g., (Derclaye 2008, pp. 120–126). However, in the context of GL mining this condition will be fulfilled rather easily as the GL repositories are usually made available online to all users without restrictions.

[21] For the complicated interplay between Art. 8 and 7(5) DD, see in detail (Triaille et al. 2014, pp. 78–79) and sources and case law of CJEU cited therein.

It could thus be concluded that the current system of exceptions does not provide enough legal certainty for TDM,[22] which could, as Renda et al. suggest, be consequently *"detrimental to the development of new offers and services, which in turn limits benefits to society through a direct negative impact on so-called 'dynamic efficiency' (e.g. innovation and the development of new welfare-enhancing products and services)."* (2015, p. 131). Handke et al. (2015, p. 21) even showed that as far as data mining research is concerned, copyright seems to have a negative net effect on innovation.

The proposed text and data mining exception

To eliminate the problems described above, a specific TDM exception was proposed by the European Commission in the recently introduced reform of European copyright law.[23]

TDM is defined in the Proposal as: *"any automated analytical technique aiming to analyse text and data in digital form in order to generate information such as patterns, trends and correlations".* According to Art. 3 of the Proposal, the beneficiary of this exception should be, without the consent of the respective right holder, allowed to 1) directly or indirectly, temporarily or permanently reproduce by any means and in any form, in whole or in part the copyrighted works; [24] (2) temporarily or permanently reproduce the structure of the database by any means and in any form, in whole or in part and (3) extract and/or re-utilize of the whole or of a substantial part of the contents of the database.[25] These activities could in the end-effect lead to commercialisation of the results, as the exception does not restrict the nature thereof. The pre-requisite invariably required for application of the exception is lawful access to the protected subject-matter that is to be mined and the scientific research purpose to be achieved. In accordance with the suggestions of various expert groups,[26] such an exception shall not be overridden by contract. Contractual arrangements to the contrary shall be deemed unenforceable (Art. 3(2) Proposal). In order to avoid potential overloading and security/integrity breaches of the networks and databases, the right holders shall be allowed to implement proportionate measures (Art 3(3) Proposal). Lastly, the overprotective application of such preventive measures shall be avoided by defining *"commonly-agreed best practices"* between the right holders and researchers (Art. 3(4) Proposal). No fair compensation for the use allowed under this exception that is comparable to, e.g., the private copying levies, is foreseen for the TDM exception. It is also subject to the three-step test, which again limits its scope of application (Art. 6 Proposal). The complicated provision of Art. 6(4) ISD, regulating the relation between technological measures of protection (DRM) and exceptions to exclusive rights, shall also apply. Consequently, if the works/repository is

[22] This view was also expressed by the respondents to the European Commission consultation on review of EU copyright rules (*Report on the responses to the Public Consultation on the Review of the EU Copyright Rules* 2014, p. 63).

[23] Article 3(1) of the Proposal: "Member States shall provide for an exception to the rights provided for in Article 2 of Directive 2001/29/EC, Articles 5(a) and 7(1) of Directive 96/9/EC and Article 11(1) of this Directive for reproductions and extractions made by research organisations in order to carry out text and data mining of works or other subject- matter to which they have lawful access for the purposes of scientific research."

[24] Art. 2 ISD.

[25] Furthermore, the "press publishers right" according to Art. 11(1) of the Proposal shall not be infringed by acts under the TDM exception.

[26] See, e.g., (European Commission, Directorate-General for Research and Innovation 2014, p. 52).

protected by technological measures, the right holders should voluntarily allow the beneficiaries to make use of the TDM exception. In the absence thereof, Member States should provide for appropriate measures so that the beneficiaries may actually profit from the TDM exception.

A fundamental flaw of the proposed TDM exception lies in the restricted personal scope. The beneficiary must be a public research institution (Art. 2(1), i.e. "*a university, a research institute or any other organisation the primary goal of which is to conduct scientific research or to conduct scientific research and provide educational services".* Furthermore, these activities must be done *"(a) on a non-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognised by a Member State"* and in *"such a way that the access to the results generated by the scientific research cannot be enjoyed on a preferential basis by an undertaking exercising a decisive influence upon such organisation;"*

This relatively complicated precondition should eliminate business as beneficiaries, but should enable Public-Private Partnerships.[27] As the League of European Research Universities remarks, this limits significantly the practical value of TDM (League of European Research Universities, 2016). It could be argued that such an exception is discriminatory in its nature. The NGO Communia Association even goes as far as to say that this regulation creates a *"privileged class of data miners"* by pointing out that the proposed article could be interpreted in such a way that any other subject (i.e. not research organisations) must specifically ask for the consent of the right holders to mine (Vollmer, 2016). Cocoru and Boehm see this condition as *"just another way of reflecting the commercial vs. non-commercial debate"* (2016, p. 8). The uncertainties raised would, according to them, only benefit those *"who can afford to capitalise on legal uncertainty"* (Cocoru, Boehm, 2016, p. 8). A broader approach to this exception would lead to *"more competition, broader services, and more creation and dissemination of knowledge"* (Cocoru, Boehm, 2016, p. 18). A logical solution to this problem is the elimination of this restrictive criterion.

In the context of mining **GL repositories**, the proposed solution would seem to be favourable for miners. Unlike the traditional databases of white literature, GL and its repositories are usually not hidden behind a paywall and are made available online without further contractual restrictions. Consequently, the prerequisite of lawful access should be fulfilled as standard. By contrast, for the operators of GL repositories, the exception, when adopted in its current format, would potentially lead to heightened traffic and usage of their repositories. They could, however, employ the required adequate measures to prevent the collapse of these. Logically, criticism regarding the scope of beneficiaries, as provided above, remains the same and is even more substantiated due to the value and importance of GL.[28]

[27] COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT on the modernisation of EU copyright rules. Accompanying the document Proposal for Directive of the European Parliament and of the Council on copyright in the Digital Single Market and Proposal for a Regulation of the European Parliament and of the Council laying down rules on the exercise of copyright and related rights applicable to certain online transmissions of broadcasting organisations and retransmissions of television and radio programmes. Brussels, 14.9.2016 SWD(2016) 301 final. Part 1/3, p. 109.

[28] See more on this: e.g., (Myška, Šavelka 2013, para. 3). GL also comprises the full spectrum of content to be mined, not only text and data. What is more, GL comprises detailed data and information not available in standard distribution channels (Castro, Salinetti 2004, p. 4).

Conclusion

Due to the legal uncertainty and technical complexity of TDM, the status quo in terms of the possibility of mining content without the consent of the respective right holders is far from optimal. The users of TDM must rely on the national implementations of ISD exceptions, which are relatively restrictive, do not actually enable the use of TDM to its full potential and, in the case of the scientific research exception (Art. Art. 5(3)(a) ISD and 9(b) DD), may differ at a national level. Consequently, the content miner could well end up infringing both the rights of the GL repository operator (extraction, copyright) and of the right holders of the works contained therein (copyright).

A possible solution to the status quo, which must, however, be developed and investigated in further research, could be found in even broader opening of the repository by licensing it and the works contained therein openly (e.g. under the Creative Commons 4.0 Attribution International license,[29] which allows for TDM). The "all-in" solution of broadest possible license/waiver (where legally possible) of rights without any further obligations imposed on the potential miner is even more advisable. However, this might not always be so easily achieved because of, e.g., the obligations of the GL repository operator or the author of the mined works related to receiving public funding.[30]

This less-than-ideal state of affairs should be improved by the newly introduced TDM exception in the Proposal of the Commission, which seems to be *prima facie* beneficial to the content miner. Due to its limited personal scope, however, it does not fulfil its promising potential. These critical remarks also remain valid for the mining of GL.

De lege ferenda, an amendment to the current wording of the proposed TDM exception, as eliminating the remaining ambiguities of the Proposal, is advisable. Content mining should be allowed for *"anybody that has lawful access"* (European Commission 2016, p. 109) to the mined subject-matter without further conditions.[31] Furthermore, the delicate issues of technical protection of repositories/works should be regulated in more detail – ideally in such a way that the right holders cannot technically block activities that would be legally permitted under the exception. Yet again, this rather simplistic claim must be developed and investigated in further research. On the other hand, an attribution (information) obligation could be imposed on the beneficiary of the exception. Naming the works which have been mined (including their source, i.e. the repository from which they have been acquired) is not technically unfeasible, e.g. in the form of a link to a dedicated website where the sources would be mentioned.[32]

[29] Creative Commons 4.0 Attribution International License [online]. [Accessed 28 September 2016]. Available from: **https://creativecommons.org/licenses/by/4.0/**.

[30] As is the case in H2020 Framework Programme for Research and Innovation – Art. 43(4) Regulation (EU) No 1290/2013 of the European Parliament and of the Council of 11 December 2013 laying down the rules for the participation and dissemination in Horizon 2020 – the Framework Programme for Research and Innovation (2014-2020).

[31] Such wording of the TDM exception was initially considered as one possible solution, but was ultimately not chosen as it would allegedly have *"a considerable negative effect on publisher's TDM licensing market"* (European Commission 2016, p. 109).

[32] The author would like to express his thanks to prof. Christian Handke for suggesting this idea.

The abovementioned revisions would ultimately lead to achieving the Open Mining Manifesto,[33] would support the deployment of TDM and would consequently help to *"drive science, competitiveness and innovation"* in the EU (Association of European Research Libraries 2015, p. 1) by increasing legal certainty.

References

ASSOCIATION OF EUROPEAN RESEARCH LIBRARIES, 2015. *A Copyright Exception for Text and Data Mining* [online]. 9 December 2015, [Accessed 28 September 2016]. Available from: **http://libereurope.eu/wp-content/uploads/2015/11/TDM-Copyright-Exception.pdf**.

BANKS, Marcus, 2006. Towards a Continuum of Scholarship: The Eventual Collapse of the Distinction between Grey and Non-Grey Literature. *Publishing Research Quarterly*. Spring 2006, **22**(1), 4–11.

BORGMAN, Christine L., 2012. The Conundrum of Sharing Research Data. *Journal of the American Society for Information Science and Technology*. 2012, **63**(6), 1059–1078. DOI 10.1002/asi.22634.

BORNMANN, Lutz and Rüdiger MUTZ, 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*. 1 November 2015, **66**(11), 2215–2222. DOI 10.1002/asi.23329.

CASTRO, Paola De and Sandra SALINETTI, 2004. Quality of grey literature in the open access era: Privilege and responsibility. *Publishing Research Quarterly*. **20**(1), 4–12. DOI 10.1007/BF02910856.

COCORU, Diane and Mirko BOEHM, 2016. *An analytical review of text and data mining practices and approaches in Europe* [online]. 1 May 2016. OpenForum Europe. [Accessed 28 September 2016]. Available from: **http://www.openforumeurope.org/wp-content/uploads/2016/05/TDM-Paper-Diana-Cocoru-and-Mirko-Boehm.pdf**.

COLONNA, Liane, 2013, A Taxonomy and Classification of Data Mining. *SMU Science & Technology Law Review*. 1 October 2013, **16**, p. 309.

DERCLAYE, Estelle, 2008. *The Legal Protection of Databases A Comparative Analysis*. Cheltenham, UK; Northampton, MA: Edward Elgar. ISBN 978-1-84720-133-1.

EECHOUD, Mireille M. M. van, HUGENHOLTZ, P. Bernt, VAN GOMPEL, Stef, GUIBAULT, Lucie M. C. R. and Natali HELBERGER, 2009. *Harmonizing European Copyright Law: the Challenges of Better Lawmaking*. Alphen aan den Rijn: Kluwer Law International. Information law series, Vol. 19. ISBN 978-90-411-3130-0.

EUROPEAN COMMISSION and DIRECTORATE-GENERAL FOR RESEARCH AND INNOVATION, 2014. *Standardisation in the area of innovation and technological*

[33] "The right to read is the right to mine"; "Users and providers should encourage machine processing"; "Facts don't belong to anyone" (Murray-Rust et. al 2014, pp. 26-29).

development, notably in the field of text and data mining: report from the Expert Group. [online]. Luxembourg: Publications Office [Accessed 28 September 2016]. ISBN 978-92-79-36743-4. Available from: **http://bookshop.europa.eu/uri?target=EUB:NOTICE:KI0114289:EN:HTML**.

EUROPEAN COMMISSION, 2016, COMMISSION STAFF WORKING DOCUMENT IMPACT ASSESSMENT on the modernisation of EU copyright rules. Accompanying the document Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market and Proposal for a Regulation of the European Parliament and of the Council laying down rules on the exercise of copyright and related rights applicable to certain online transmissions of broadcasting organisations and retransmissions of television and radio programmes. Brussels, 14.9.2016, SWD(2016) 301 final. Part 1/3.

FILIPPOV, Sergey and Paul HOFHEINZ, 2016. *Text and Data Mining for Research and Innovation What Europe Must Do Next. Interactive Policy Brief*. No. 20.

FLORIDI, Luciano, 2014. *The 4th revolution: how the infosphere is reshaping human reality*. First edition. New York ; Oxford: Oxford University Press. ISBN 978-0-19-960672-6.

HANDKE, Christian, GUIBAULT, Lucie and Joan-Josep VALLBÉ, 2015. *Is Europe Falling Behind in Data Mining? Copyright's Impact on Data Mining in Academic Research* [online]. SSRN Scholarly Paper. Rochester, NY: Social Science Research Network, [Accessed 28 September 2016]. Available from: **http://papers.ssrn.com/abstract=2608513**.

HUGENHOLTZ, P. Bernt, EECHOUD, Mireille M. M. van, GOMPEL, Stef Van and HELBERGER, Natali, 2006. *EUROPEAN COMMISSION DG INTERNAL MARKET STUDY CONTRACT NO. ETD/2005/IM/D1/95: The Recasting of Copyright & Related Rights for the Knowledge Economy.* Amsterdam: Institute for Information Law, University of Amsterdam.

JEFFERY, Keith G. and Anne ASSERSON, 2014. Data Intensive Science: Shades of Grey. Procedia Computer Science. *Procedia Computer Science* [online]. **33**, 223–230. DOI 10.1016/j.procs.2014.06.036. ISSN 18770509.

LAROSE, Daniel T., 2014. *Discovering Knowledge in Data: An Introduction to Data Mining*. Second edition. Hoboken: Wiley. Wiley Series on Methods and Applications in Data Mining. ISBN 978-0-470-90874-7.

LARSEN, Peder Olesen and Markus VON INS, 2010. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*. **84**(3), 575–603. DOI 10.1007/s11192-010-0202-z. ISSN 0138-9130.

LEAGUE OF EUROPEAN RESEARCH UNIVERSITIES, 2016, EU copyright reform and TDM : potentially good for research but certainly not (yet) for innovation! [online]. LERU : League of European Research Universities. LERU [online]. 14 September 2016. [Accessed 28 September 2016]. Available from: **http://www.leru.org/index.php/public/news/eu-copyright-reform-and-tdm-potentially-good-for-research-but-certainly-not-yet-for-innovation-/**.

MURRAY-RUST, Peter, MOLLOY, Jennifer C. and Diane CABELL, 2014. Open Content Mining. In: *Issues in Open Research Data* [online]. London: Ubiquity Press. p. 11–30. ISBN 1-909188-30-1. Available from: **http://dx.doi.org/10.5334/ban**.

MYŠKA, Matěj and Jakub HARAŠTA, 2015. Omezení autorského práva a zvláštních práv pořizovatele databáze v případě datové analýzy. *Časopis pro právní vědu a praxi.* **23**(4), 375–384.

MYŠKA, Matěj and Jaromír ŠAVELKA, 2013. A Model Framework for publishing Grey Literature in Open Access. *jipitec* [online]. 25 August 2013, **4**(2), 104-115 [Accessed 28 September 2016]. Available from: **http://www.jipitec.eu/issues/jipitec-4-2-2013/3744**.

RENDA, Andrea, SIMONELLI, Felice, MAZZIOTTI, Giuseppe, BOLOGNINI, Alberto and Giacomo LUCHETTA, 2015. *The implementation, application and effects of the EU Directive on Copyright in the information society* [online]. Brussels: Centre for European Policy Studies, [Accessed 28 September 2016]. CEPS Special Report, 120/2015. ISBN 978-94-6138-487-4. Available from: **https://www.ceps.eu/system/files/SR120_0.pdf**.

Report on the responses to the Public Consultation on the Review of the EU Copyright Rules [online], 2014. Brussels: European Commission Directorate General Internal Market and Services, [Accessed 28 September 2016]. Available from: **http://ec.europa.eu/internal_market/consultations/2013/copyright-rules/docs/contributions/consultation-report_en.pdf**.

ROSATI, Eleonora, 2013. *Originality In EU Copyright: Full Harmonization through Case Law*. Cheltenham, UK ; Northampton, MA: Edward Elgar. ISBN 978-1-78254-893-5.

RUBINSTEIN, Ira S., 2013. Big Data: The End of Privacy or a New Beginning? *International Data Privacy Law*. **3**(2), 74–87. DOI 10.1093/idpl/ips036.

SCHÖPFEL, Joachim, 2010. Access to European Grey Literature. In: *Grey Literature Repositories*. Zlín: VeRBuM. p. 20–33. ISBN 978-80-904273-6-5.

STAMATOUDI, Irini A. and Paul TORREMANS (eds.), 2014. *EU Copyright Law: A Commentary*. Cheltenham: Edward Elgar. Elgar Commentaries. ISBN 978-1-78195-242-9.

TRIAILLE, Jean-Paul, DUSOLLIER, Séverine, DEPREEUW, Sari, HUBIN, Jean-Benoit, COPPENS, François and Amélie de FRANCQUEN, 2013. *Study on the Application of Directive 2001/29/EC on Copyright and Related Rights in the Information Society* [online]. Brussels: European Commission, [Accessed 28 September 2016]. ISBN 978-92-79-29918-6. DOI DOI:10.2780/90141.

TRIAILLE, Jean-Paul, MEEÛS D'ARGENTEUIL, Jérôme de and Amélie de FRANCQUEN, 2014. *Study on the legal framework of text and data mining (TDM)* [online]. Luxembourg: Publications Office, [Accessed 28 September 2016]. ISBN 978-92-79-31976-1. Available from: **http://bookshop.europa.eu/uri?target=EUB:NOTICE:KM0313426:EN:HTML**.

TRUYENS, Maarten and Patrick VAN EECKE, 2014. Legal aspects of text mining. *Computer Law & Security Review*. April 2014, **30**(2), 153–170. DOI 10.1016/j.clsr.2014.01.009.

VOLLMER, Timothy, 2016. *Commission proposes to limit text and data mining in Europe. International Communia Association* [online]. 6 September 2016 [Accessed 28 September 2016]. Available from: **https://www.communia-association.org/2016/09/06/commission-proposes-limit-text-data-mining-europe/**.

WALTER, Michel M. and Silke von LEWINSKI (eds.), 2010. *European Copyright Law: A Commentary*. Oxford ; New York: Oxford University Press. ISBN 978-0-19-922732-7.

WARE, Mark and Michael MABE, 2015. *The STM report An overview of scientific and scholarly journal publishing* [online]. STM: International Association of Scientific, Technical and Medical Publishers, [Accessed 28 September 2016]. Available from: **http://www.stm-assoc.org/2015_02_20_STM_Report_2015.pdf**.

WESTKAMP, Guido, GUIBAULT, Lucie and Thomas RIEBER-MOHN, 2007. *MARKT/2005/07/D: Study on the Implementation and Effect in Member States' Laws of Directive 2001/29/EC on the Harmonisation of Certain Aspects of Copyright and Related Rights in the Information Society* [online]. Amsterdam: Institute for Information Law, University of Amsterdam, [Accessed 28 September 2016]. Available from: **http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2006358**.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

BUT DIGITAL LIBRARY – INTRODUCTION OF INSTITUTIONAL REPOSITORY

Jan Skůpa

skupa@lib.vutbr.cz

Brno University of Technology, Central Library, Czech Republic

Martin Fasura

fasura@lib.vutbr.cz

Brno University of Technology, Central Library, Czech Republic

Vojtěch Bartoš

bartos@fbm.vutbr.cz

Brno University of Technology, Czech Republic

Abstract

The Brno University of Technology (BUT) Digital Library is an institutional repository running of DSpace (version 5.3) initially created in 2007 as a repository for theses. The Digital Library's range of services is gradually growing and this paper deals specifically with repository services, the process for publishing theses, and support for open access. The repository enables authors to store the research results in open access mode. Full text is integrated into the national Czech systems for reporting publications (RIV).

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

Keywords
DSpace, Institutional Repository, The Brno University of Technology, Open Access, Theses

Introduction
Brno University of Technology was founded in 1899, is one of the oldest universities in the Czech Republic and the first university in Moravia. Its library – to which the Central Library is still indirectly linked today – was founded in the same year as the university. Not only the new époque, but also legislative conditions have led to a repository named the Digital Library (DL) taking shape at the Brno University of Technology.

Perhaps the greatest motivation for the establishment of the DL was the issuing of Rector Directive No 9 of March 2007 "Editing, submission and publication of theses and dissertations at Brno University of Technology". This directive was a reaction to changes in the Higher Education Act (Act No 111/1998, as amended by Act No 552/2005) (Česká republika, 2005) and for the first time introduced the mandatory submission of works in electronic form.

The objectives of this paper are to summarise the history of the Digital Library in Brno (**https://dspace.vutbr.cz)**, and to describe its contents and other activities closely related to its operation.

The first system – DigiTool
The Central Library at Brno University of Technology reacted to the need for a repository by submitting the University Development Fund project "Building the Digital Library of Brno University of Technology", adopted in 2007 and implemented the following year. The objectives of the project were to acquire and put into operation a server, the DigiTool system, and to fill the Digital Library with full-text records. The DigiTool system was mainly selected to secure full support by the supplier. The system also had a similar foundation as the Aleph library system, and thus facilitated easier connection of the two systems. The DigiTool system was operated between 2008 and 2010 and an internal process for the submission and publication of electronic theses and dissertations (eVŠKP) was set up at Brno University of Technology. In 2011 there was an evaluation of the course of the project. The DigiTool system was shown to have poor future prospects – in particular due to the fact that the producer had stopped developing the system, there was practically no supportive user base, and due to the pricing policy. This unsustainable state meant the need to select a new system.

The DSpace system was finally selected from several possible options – it is one of the most widely used systems for institutional repositories (including in in the Czech Republic) and thus provides extensive user support. One great advantage of this system is its free distribution method (open source), which represents savings on maintenance and licencing costs.

DSpace

The DSpace system was installed and put into operation in 2012 with the help of employees from the Computer and Information Services Centre (CVIS) at Brno University of Technology. Initially, DSpace was operated in version 1.7. It was upgraded to DSpace 4.0 in 2014, and to DSpace 5 in 2015 together with the new, responsive Mirage2 design. The system has also undergone significant technical changes – after long consideration the database was changed from Oracle to PostgreSQL. This solved some problems that it had not been possible to reliably rectify, and will perhaps also prevent future problems. Postgre is used by most DSpace installations – problems and errors are corrected faster, and support is also much more comprehensive.

Content of the Digital Library at Brno University of Technology

The Digital Library at Brno University of Technology fulfils the function of an institutional repository – it collects, manages and provides access to works created at the university. It actually only retains so-called born-digital materials. There is no sophisticated digitisation workplace with the task of systematically digitising historic materials at the Central Library. The content of the repository is divided into collections and sub-collections (according to the DSpace taxonomy, these are communities, sub-communities and collections) depending on the type of stored record. Root collections (communities) preserve and make available:

- Magazines
- Certified methodologies
- Publication activity of employees at Brno University of Technology
- Proceedings from conferences
- Central Library
- Scientific writings
- Brno University of Technology annual reports
- Theses

If appropriate, the collections will be further divided by faculty, and also by institute and department for the publication activity of Brno University of Technology staff.

Undergraduate theses

Theses represent the largest collection in the repository. All eight faculties and one university institute (of Forensic Engineering) contribute to this collection of theses. These are undergraduate theses and dissertations dating back - for some faculties - to 2006. The first records differ between the faculties. It always depends when the faculty began using the unified workflow in the university's information system. In September 2016, the repository contained 43,603 records.

Theses are taken from the university information system and are available at least five days before defence. The linking of the DSpace system and the information system is performed by a synchronisation program primarily using the REST API model. Each record always has a metadata description and reference attached in an HTML file. The metadata transferred from

the information system are not manually revised, and the students themselves or the supervisors who confirm the entered data are responsible for formal correctness.

Publication activities of Brno University of Technology staff

The collection of publication activities by Brno University of Technology staff forms an important part of the repository. Brno University of Technology employees may publish the full texts of articles from specialist magazines, papers from conferences and chapters from books through the Digital Library. A special section for entering full texts has been created in the Apollo university information system in the module for reporting publication activity. The author may decide whether or not they want to publish the stored text in the repository (if not, the space serves an archive function for the author). After registration of the application, the full text is checked by the repository manager. This is primarily an analysis of whether it is possible to publish the full text and whether it is properly described through its metadata. The full text is placed into the repository and becomes freely available only after approval. It is also possible to set a time embargo according to conditions imposed by publishers – the publication of the whole record in the Digital Library is then postponed until a specific date. The entry process is described in the instructions at **https://www.vutbr.cz/uk/digitalni-knihovna/jak-zverejnit**.

The obligation to enter full text is not regulated by any internal regulation, with the exception of articles supported from the Fond Open Access grant programme. The self-archiving of own texts is a procedure that is used more frequently, yet is not sufficient. At the present time, therefore, we are focusing on analysing ways to increase the number of entered records. The plan includes both contacting authors (e.g. by preparing a summary of publication activity and cooperation with storage in the repository) as well as a possible policy change at the university to introduce the mandatory submission of full text.

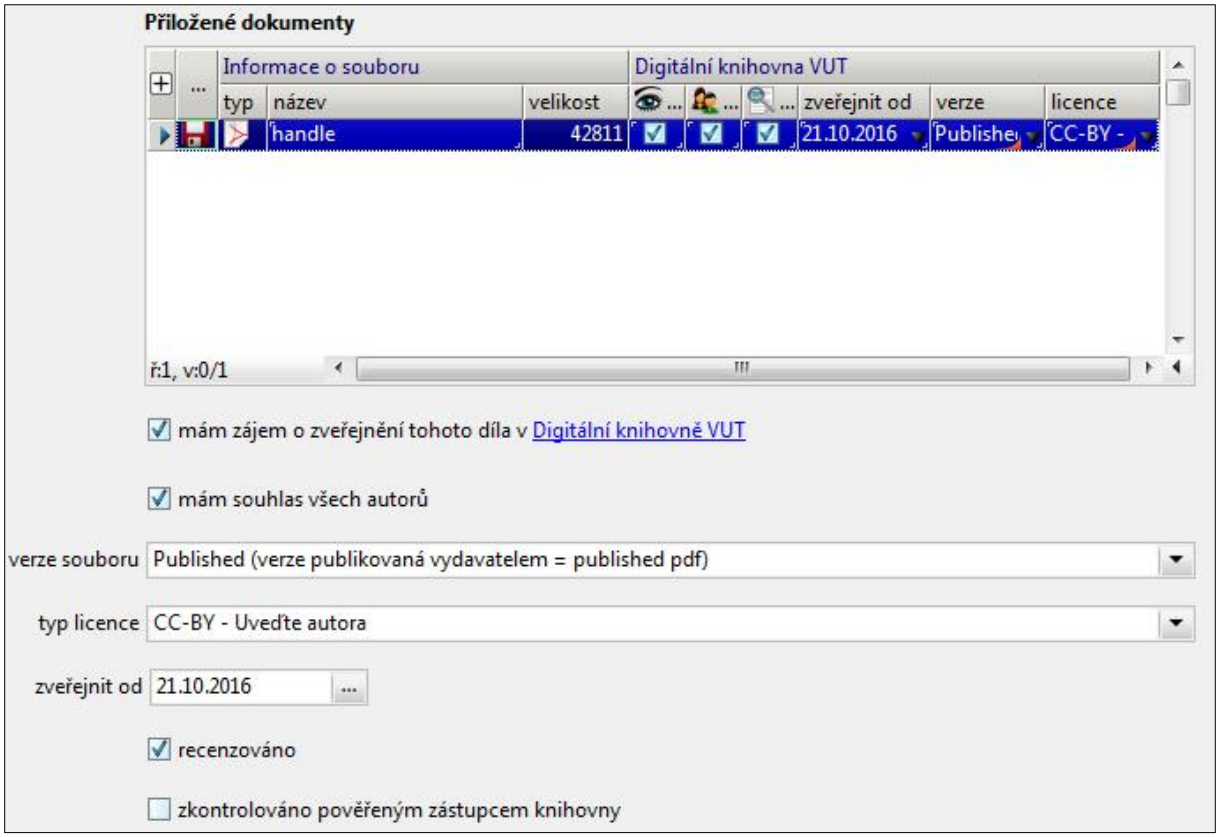


Figure 1: Form for entering an article via the Apollo information system

Magazines

Magazines produced at Brno University of Technology are systematically stored in the Digital Library. The first archived magazine since the Digital Library was put into operation was the magazine Události, issued by the Nakladatelství VUTIUM publisher. This is a magazine providing information about events at the university. The magazine editors are individually contacted with an offer to archive the magazine in the Digital Library. This was expanded to include the magazines Mathematics for Application, Kvaternion (both from the Faculty of Mechanical Engineering) and Trendy ekonomiky a managementu [Trends in Economics and Management] (Faculty of Business and Management). The magazine with the largest number of articles stored in the DL is Radioengineering (Faculty of Electrical Engineering and Communication). The Digital Library currently holds a total of 1,640 articles from 94 issues of this magazine. The articles were placed into the Digital Library retrospectively from 1992. A magazine makes its content available under the CC-BY licence and is also indexed in Web of Science and in Scopus.

Conference proceedings

Papers from conference proceedings form an equally important part of the Digital Library. Often the full texts of the proceedings are only provided to conference participants (on CD or flash disk) and, in the better case, the whole proceedings are placed on the conference website as a PDF, which however does not have a long lifespan. The Central Library is working to contact faculties with an offer to archive such proceedings. Currently, 18 proceedings from six conferences have been placed in the Digital Library.

Certified methodologies

Mainly due to requirements from the Ministry of Culture, we have also incorporated into the repository our own collection of certified methodologies. Both the record and the full text of a certified methodology are entered manually into the repository. At the same time, a special set for OAI-PMH has been created, which enables the National Repository of Grey Literature to harvest certified methodologies with full text for its repository.

Documents for SSP

In cooperation with the Alfons university counselling centre (formerly Přes bloky), full texts (both digital and digitised by the counselling centre) of books and textbooks are placed into the Digital Library for the needs of students with specific needs. The records of these documents are accessible for all users, while the downloading of the full text is only available for students who demonstrate eligibility according to the conditions of the counselling centre (Skůpa, 2016a) The Digital Library is also connected to the Library Gateway for the visually impaired (available from **https://www.teiresias.muni.cz/knihovni-brana/**). However, the connection is made through the Aleph library system primarily due to the absence of the Z39.50 interface in DSpace. Records are thus catalogued twice – to the Digital Library and to the library catalogue. In the library catalogue, however, the records are not visible and exist only for the purposes of transfer via the Z39.50 protocol into the Library Gateway.

Linking the Digital Library

The Digital Library at Brno University of Technology cooperates through the OAI-PMH protocol with other repositories that could increase the use of the DL. It is connected to the OpenAIRE portal (**www.openaire.eu**), where all openly available documents are placed. In 2016, the DL was connected to the National Repository of Grey Literature, where the metadata records of theses and conference proceedings are placed. Only certified methodologies are transferred in full text.

The OAI-PMH protocol is also used by other tools. The tool Citace PRO was implemented in 2015, which generated a bibliographic record for all records in the detail view according to the ČSN ISO 690 standard. This means that it is easier for the user to save the record on a personal account in the Citace PRO manager (available at **http://citace.lib.vutbr.cz**). The newest implemented tool is the alternative PlumX metrics, which serve on the one hand to make the display of statistics in the DSpace space more attractive, and on the other hand to record the real impact of the individual links – use, mentions on social networks, or in citation managers.

The Digital Library is also indexed and therefore searchable via the Primo discovery search engine (**https://primo.vutbr.cz**). In the search engine environment, users can then search both in all areas (library catalogue, Digital Library, online resources), and also only in the Digital Library.

Open access

The Central Library actively promotes open access at the university. It was already actively participating in the Open Access Week international event from 2010. In 2013, Brno University of Technology became a signatory to the Berlin Declaration. At the same time, through Rector Decision No 21/2013, the "Declaration of the Institutional Open Access Policy at Brno University of Technology" was adopted (available from **https://www.vutbr.cz/uk/openaccess/rozh21-pdf-p84132**). This policy expresses the university's proactive stance, however does not determine any obligation related to publishing or saving scientific articles and other works.

In connection with these changes, it has also been possible to push through real support for open publishing – Fond Open Access – through which CZK 1 million is allocated every year. The fund defines strict rules for awarding grants (Skůpa, 2016b). An article must be published in a fully open magazine evaluated using the R&D Methodology, which in real terms means its indexing in the Web of Science or in Scopus. At the same time, the payment for publication (APC) must not be over USD 2,000 or EUR 3,000. The fund was first launched in mid-2014 and six articles were supported from it. In 2015, the fund operated under the same conditions and subsidised 17 articles. 11 articles have so far been supported in the current year, and around CZK 500,000 remains to be drawn. One of the obligations is that a subsidised article must be included in the Digital Library at Brno University of Technology. This has contributed to completing the collection of the Publication Activity of Brno University of Technology staff.

DOI

The Central Library allocates a DOI identifier to magazines and proceedings produced at Brno University of Technology. At the present time, a DOI is regularly allocated to the magazines Mathematics for Application, Radioengineering and Trendy ekonomiky a managementu. A DOI was also allocated to several conference proceedings and to the book Otevřený přístup k vědeckým informacím: současný stav v České republice a ve světě [Open Access to Scientific Information: Current State in the Czech Republic and Abroad]. A DOI is allocated free of charge, the only condition is the inclusion of the full text in the Digital Library.

Conclusion

Over the years, the Digital Library at Brno University of Technology has become an important part of the services provided by the Central Library at Brno University of Technology. According to visitor statistics, this is the most-visited library website (as of 25 September, it had 12,000 visitors in that month). The Central Library is always trying to improve its operation and supplement it with additional functionality. Objectives for the near future include increasing the quantity of published full texts (magazines, proceedings and single articles) and increasing support for open access, in particular its green path.

References

ČESKÁ REPUBLIKA, 2005. Zákon č. 552/2005 Sb.: Zákon, kterým se mění zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů, a některé další zákony. In: *Sbírka zákonů ČR*. 30. 12. 2005. ISSN 1211-1244. Also available from: **http://www.zakonyprolidi.cz/cs/2005-552**.

SKŮPA, Jan, 2016a. Dokumenty pro studenty se specifickými potřebami. *Ústřední knihovna VUT v Brně* [online] Brno, 2016 [cit. 2012-09-20]. Available from: **https://www.vutbr.cz/uk/digitalni-knihovna/ssp**.

SKŮPA, Jan, 2016b. Podmínky fondu Open Access. *Ústřední knihovna VUT v Brně* [online]. Brno, 2016 [cit. 2012-09-22]. Available from: **https://www.vutbr.cz/knihovny/openaccess/dotace/fond-oa/podminky**.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

THE IMPACT OF GENERAL DATA PROTECTION REGULATION ON THE GREY LITERATURE

Michal Koščík

koscik@med.muni.cz

Masaryk University, Czech Republic

Abstract

On 27 April 2016, the European Commission adopted new "general data protection regulation" that will be effective in all Member States from 25 May 2018. This article focuses on the impact of such regulation on the operators of grey literature, especially with regard to the administrative requirements introduced by new "privacy by design rules". The article also assesses the statutory licenses of a public repository to process and make available personal information even without the consent of a data subject.

Keywords

Data Protection, Privacy, GDPR, GDPR Compliance

Introduction

Compliance with data protection rules and the privacy awareness of the operators of grey literature repositories have gradually improved over the past decade. Operators in the Central European region have invested significant amounts of time, resources and effort to achieve compliance and train their employees in data protection issues. They have also been adapting

in recent years to developments in Case-Law of the European Court of Justice (hereinafter "CJEU") in order to comply with the newly-formulated right to be forgotten.

On 27 April 2016, the European Commission adopted new "general data protection regulation [1] **"** (hereinafter "GDPR"). In contrast to the previous data protection directive (95/46/EC), the regulation does not have to be transposed into the legal systems of individual Member States. The rules contained in GDPR have direct effect and will be effective in all Member States from 25 May 2018. GDPR is a rather extensive piece of legislation. Its recital has 173 points, the normative part has 99 Articles and the whole directive altogether takes up 88 pages of the official European Journal. Moreover, it is accompanied by the directive on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences, which was adopted on the same day.

This article outlines the impact of new data protection rules on grey repositories and assesses the steps that need to be taken in advance to be prepared for compliance with the new regulation.

Continuity with the principles of the current directive

The adoption of GDPR was not brought about by any need to significantly change the fundamental principles of existing data protection rules, but rather by an acknowledgement that data processing technologies have changed significantly since the adoption of the current directive in 1995. As the recitals of GDPR state, the objectives and principles of Directive 95/46/EC remain sound (see recital 9. GDPR), but the legislation needs to address current conditions such as Rapid technological developments and globalization (see recital 6 GDPR), cross-border flows of personal data (recital 5 GDPR) and significant risks to the protection of natural persons, in particular with regard to online activity.

The definition of personal data remains extremely broad[2], as does the definition of personal data processing[3]. Basically every systematic work with any kind of information that contains references to individuals falls under the scope of GDPR unless it is done in the course of purely personal or household activity (Art. 2(2) GDPR) or by authorities responsible for forensic and security tasks.

GDPR preserves the concept of distinction between the "controller of data", i.e. the person who determines the purpose of data processing, and the "processor of data", who performs certain activities at the controller's request. The **purpose** of data processing remains the central concept and starting point for any further considerations. Once the purpose has been defined, the controller and processor have to process data in accordance with fundamental principles of lawfulness, fairness, transparency, purpose limitation, data minimization, storage limitation,

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation); OJ L 119, 4.5.2016, p. 1–88.

[2] 'Personal data' means any information relating to an identified or identifiable natural person.

[3] "Processing' means any operation or set of operations which is performed on personal data or on sets of personal data.

accuracy, integrity and confidentiality[4]. It should be noted, however, that these concepts are also in line with the current data protection directive and case-law at the Court of Justice (CJEU). **The change is more a matter of more precise formulation of these principles than any fundamental change in basic concepts**. The exact formulation of the rules arising from these principles is more detailed and offers explicit solutions for cases which have been open to interpretation until now.

The principle of storage limitation and exception for archiving in the public interest

The principle of storage limitation is of great relevance to grey repositories. GDPR states that personal data must not be kept in a form which permits identification of data subjects for any longer than is necessary for the purposes for which the personal data are processed (Art. 5(1)(e) GDPR). This means that the stored documents must be made anonymous at a certain point in time. GDPR, however, allows one major exception, i.e. long term processing "solely for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes". Hence, public repositories are generally entitled to collect, process, store and make available certain personal data, even if that information was not originally created or collected for the purposes of archiving in a repository.

The exception is not without restrictions. The use of every exception has to be balanced with the rights of the data subject. Learning how to balance public interest and the rights of an individual is the most important and most difficult legal question for any public repository.

Data protection by design and default – liability begins even before processing takes place

The concept of data protection by design is a certain form of good practice on the part of the data controller to design its processes and systems in order to minimize the risks of data protection breaches. GDPR introduces the obligation of the controller to take appropriate technical and organizational measures to ensure and be able to demonstrate that processing is performed in accordance with GDPR (Art. 24). The controller is explicitly required to assess the risks and make plans for the security of data, both at the time of determining the means of processing and at the time of processing itself (Art. 25(1) GDPR)[5]. We believe that an explicit formulation of "privacy of design" principles will have little real impact since these principles are applied by most repositories even now. The obligation to be able to demonstrate such compliance at any point in time will, however, increase the paperwork and legal costs at every institution that stores and processes virtually any kind of documents. The paperwork that needs to be done before, during and after processing and the records that have to be kept are defined

[4] These concepts are explained in detail in Articles 5-11 GDPR.

[5] For further reference see Allen & Overy. *The EU General Data Protection Regulation - A new data protection landscape* [online]. 2016 [cit. 1.20.2016]. Available from: **http://www.allenovery.com/SiteCollectionDocuments/Radical%20changes%20to%20European%20data%20protection%20legislation.pdf**.

in extensor to Article 30 GDPR. Fortunately, this extensive list is not applicable to institutions that employ less than 250 employees.

Processing data with and without consent

The processing of personal data is lawful if the person has given consent[6] to process his or her personal data. All consent must be "specific, informed and unambiguous" (Art. 4(11) GDPR) and has to be "made by a statement or by a clear affirmative action". GDPR hence rules out the possibility of so called "opt-out consents", i.e. schemes where a repository makes the individual aware that his data are collected and processed and presumes his consent unless the individual indicates otherwise. Hence, the request for consent must be given in an intelligible and easily accessible form and using clear and plain language and it must be as easy to withdraw consent as it is to give it[7].

Art. 6 GDPR sets forth five explicit exemptions when the controller may process documents with personal data without the consent of the data subjects. The operators of repositories will rely mainly on the exemption of "compliance with a legal obligation of a controller" (where certain documents have to be archived by law), "performance of tasks in the public interest" (especially repositories operated by public libraries) or "other legitimate interests pursued by the controller", providing that such interests are not such interests that are "overridden by the interests or fundamental rights and freedoms of the data subject".

GDPR does not specify which purposes are legitimate in justifying processing without consent and which are not. However, the recitals of the directive set forth that it should be for the controller to demonstrate that its compelling legitimate interest overrides the interests or the fundamental rights and freedoms of the data subject. We can conclude by reading these provisions together with the exception to the principle of storage limitation in Art. 5 (see above) that public repositories which gather and process grey literature for archiving, scientific, historical research or statistical purposes do not necessarily require consent from any individual mentioned in the documents being processed. This does not mean, however, that their statutory license is not unrestricted. Operators will have to bear in mind the purpose of every single activity performed with a document that contains personal data.

Each activity and process that involves the document would have to be assessed in light of whether such activity is truly necessary for the public purpose. For example, if the repository concludes that archiving certain documents is within the public interest, it has to then ask whether posting such documents online is also within the public interest.

Right to object and right to erasure (right to be forgotten)

A data subject has the specific right to object (see Art. 21 GDPR) to any form of processing of his/her personal data, even if the repository (as a data controller or processor) is a public

[6] It is widely discussed whether processing based on consent is the most appropriate way to approach data processing, see, e.g., MÍŠEK, Jakub. Consent to Personal Data Processing – The Panacea or The Dead End? *Masaryk University Journal of Law and Technology*. **8**(1), 69-83. ISSN 1802-5943.

[7] See EU GDPR Portal - **http://www.eugdpr.org/key-changes.html**.

institution which processes such data to fulfil its statutory tasks. The repository must invariably prove that the interest in processing such information overrides the interests or the fundamental rights and freedoms of the data subject.

Apart from the right to object, GDPR introduces special rules about information that is not only processed, but published as well. Grey repositories already have already had to adapt their policies to comply with the "right to be forgotten" rule which was formulated by the Court of Justice of the European Union in the "Google Spain"[8] case[9] from 2013. GDPR follows the ideological path outlined by the court in the Google Spain case and makes the lives of repositories marginally easier by providing clearer and explicit rules in a piece of legislation. The rules on "the right to be forgotten" or "the right to erasure" (these two terms have to be read as synonyms) are found in Art. 17. If the repository archives and/or publishes a document which contains personal data, the individual concerned can demand that these data be erased if such data are no longer necessary in relation to the purposes for which they were collected or otherwise processed. These rules are in line with the Google Spain ruling. However, a new rule that goes much further to protect the interests of the data subject is introduced in Art. 17(2). A controller who has granted the right of erasure and erased personal information is obliged to inform other controllers that are processing such personal data to erase any links to, or copies or replications of, those personal data. In other words, if a repository publishes a copy of a document from a third party, it has to inform such party that a request to erase personal data has been made.

GDPR also articulates exceptions where a repository can justify the processing of personal data even against the request of a data subject to erase such information. The repository is entitled to keep its documents intact (even published) if the processing serves a task carried out in the public interest or in the exercise of official authority vested in the controller, on the grounds of public interest in the area of public health, for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes[10].

Anonymization, pseudonymization and profiling

The anonymization or pseudonymization of a document is a technique which is broadly used to manage risks of privacy or data protection claims. GDPR acknowledges these techniques as legitimate and states that anonymized data fall outside the scope of data protection regulation.

However, GDPR distinguishes between anonymized and pseudonymized information, which is any information that can be de-cyphered, whereby the individual can be tracked and identified (even if the key to decipher pseudonyms is not in the possession of the entity that processes pseudonymized data). Hence, pseudonymized data are personal data and fall within the scope of the regulation. Pseudonymization is acknowledged as a technique that can

[8] Case C-131/12, ECLI:EU:C:2014:317

[9] A detailed analysis was published in 2014: KOŠČÍK, Michal. Privacy and anonymization in repositories of grey literature. *The Grey Journal. 2015*, **11**, 47-51. ISSN 1574-1796. MÍŠEK, Jakub a Jakub HARAŠTA. Analýza praktických dopadů rozhodnutí Soudního dvora EU ve věci Google Spain. *Bulletin advokacie*. 2015, (1-2), 30-34. ISSN 1210-6348.

[10] See recital 65 GDPR – the exact rules are formulated in the Art. 17(3) GDPR.

reduce the risks to the data subjects concerned and help controllers and processors meet their data-protection obligations" (see recital 28 GDPR); it is emphasized, however, that pseudonymization cannot be the only technique deployed by the repository in order to comply with data protection rules.

Conclusion

It can be concluded that a repository operator that has taken data protection and privacy issues seriously will not have many problems in complying with the standards of GDPR. The positive side of GDPR is that it extensively formulates rules that have been open to interpretation by doctrine and the case law of European courts. It can be said that the European Commission does not stray from the widely accepted interpretations of the current Data Protection Directive, a fact which adds to the legal certainty of both data subjects and data controllers alike. Therefore, the main change, and main negative impact, of the directive is the increased requirements on paperwork and record-keeping, especially for institutions that employ more than 250 employees.

References

ALLEN & OVERY. *The EU General Data Protection Regulation - A new data protection landscape* [online]. 2016, [cit. 1.20.2016]. Available from: **http://www.allenovery.com/SiteCollectionDocuments/Radical%20changes%20to%20European%20data%20protection%20legislation.pdf**.

KOŠČÍK, Michal. Privacy and anonymization in repositories of grey literature. *The Grey Journal*. 2015, **11**, p. 47-51. ISSN 1574-1796.

MÍŠEK, Jakub. Consent to personal data processing – The Panacea or The dead end? *Masaryk University Journal of Law and Technology.* 2014, **8**(1), p. 69-83. ISSN 1802-5943.

MÍŠEK, Jakub a Jakub HARAŠTA. Analýza praktických dopadů rozhodnutí Soudního dvora EU ve věci Google Spain. *Bulletin advokacie.* 2015, 1-2, 30-34. ISSN 1210-6348.

POLČÁK, Radim. Getting European data protection off the ground. *International Data Privacy Law.* 2014, **4**(4), 282-289. DOI 10.1093/idpl/ipu019.

SCHAAR, Peter. Privacy by design. *Identity in the Information Society.* 2010, **3**(2), 267-274. DOI **10.1007/s12394-010-0055-x**.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

ARE THERE ANY DIGITAL CURATORS IN CZECH LIBRARIES?

Radka Římanová

radka.rimanova@ff.cuni.cz

Charles University, Institute of Information Studies and Librarianship, Czech Republic

Marek Melichar

marek.melichar@ruk.cuni.cz

Charles University, Computer Science Center, Czech Republic

Abstract

The development of digital repositories reveals the need for new Czech terminologies for library positions. One such job title or specialization is "digital curator," a term which refers to a specialist responsible for digital collections. Competencies for digital curators include technical, organizational, communication, and marketing skills as well as some level of expertise in computer science as well as in information and library science. A survey of several administrators of Czech digital libraries asks how to best professionally label this group of library staff with respect to the actual activities they perform in their jobs - digital curators, administrators of digital libraries, or digital librarians? Job titles have a strong impact on higher education curricula, on the management of libraries, and also on the coordinated system of remuneration of employees at libraries, academic, and research institutions.

Keywords

Digital Data, Digital Libraries, Professional Competence, Digital Curator, Librarian as Profession

Introduction

The English term "*digital curation*" was initially used in archives of scientific data and data from space research ("*data curation*"). The collections of the first large archives of scientific data in the 1990s, such as ***NASA Planetary Data Systems***, contained unique digital data collected over several preceding decades on many types of carrier. Maintaining the continuous availability of such collections was complex, and so the concept of "*digital preservation*" appeared in the community of archives of scientific data at the end of the 1990s, followed by talk of "*digital curation*". In the early 21st Century, collecting, protecting, managing and accessing, and various methods of increasing the value of, digital collections [1] became standard functions of many libraries and information centres (Higgins, 2011; Prom, 2011). The issue of the management, protection and accessing of content in digital form gradually began to affect academic libraries and cultural heritage institutions (Ray, 2009). The importance of repositories and digital collections in all types of institution is growing. Little attention was initially paid to questions of the professional preparation of the managers of digital libraries and repositories in comparison with the level of investments into technology (Ray, 2009; Pomerantz, 2006). Only projects such as ***DigCCurr I and II*** between 2006 and 2013[2] defined the education profiles needed to create a *digital curator curriculum*.

Currently (Madrid, 2013; Kim, 2015) academic institutions in the field of library and information science (LIS) routinely offer courses on digital curatorship. Czech LIS schools place this topic into the gradual innovation of subjects (Jilečková, 2015), while an independent accredited study programe ocusing exclusively on digital curatorship does not yet exist in the Czech Republic. Not even the thesis of Michal Konečný (2016) considers the introduction of a new field of study, but rather proposes a syllabus for a single-semester subject entitled "Introduction to Digital Curatorship". The practical needs of Czech libraries have resulted in the definition of the *profession of librarian – digital library manager*[3], which is part of the National System of Occupations and the National Qualifications Framework (Houšková, 2012). In the ***Catalogue of Work*** (Czech Republic, 2010), issues associated with the management of a digital library have been incorporated in the description of the profession of *Librarian* in a very non-conceptual manner.

Survey among Czech managers of digital libraries/collections

In the Czech Republic, there is no empirical data about how digital data is managed in libraries, or who does it. SDRUK (The Association of Libraries of the Czech Republic) attempted to map related issues in 2007 [4], there is some relevant information available in the results of

[1] For example the definition of DCC **http://www.dcc.ac.uk/digital-curation/what-digital-curation**.

[2] The project outcomes are available at: **https://ils.unc.edu/digccurr/index.html**.

[3] The profession of librarian – digital library manager – is addressed at three competency levels - librarian, independent librarian (V, specialist librarian (expert)). To understand the descriptions of the librarian specialisations (where there is less IT competence than might be necessary) it is important to realise that pure IT specialists may also work in a library, for whom there are competency profiles in the "Information Technology" group, see **http://katalog.nsp.cz/poziceOdbornySmer.aspx?kod_sm1=5&kod_smeru=5**.

[4] **http://sdruk.mlp.cz/data/xinha/sdruk/zmapovani_situace_digitalizace_v_cr.pdf**

the Europe-wide project focused on the collection of statistical indicators of the digitisation of cultural heritage called “Enumerate”[5], yet summary current information is lacking in this field.

The following text summarises the results of a survey conducted between May and July 2016. We questioned the employees of six libraries and one archive, a total of fifteen respondents[6]. The survey was conducted in the form of a semi-standardised interview. We questioned people who actually work with digital data in libraries. We did not record the interviews, but made two parallel records of them, and coded and evaluated the texts of the interviews. The results cannot be related to any defined target population. Our objectives were to better understand the situation that has naturally arisen, and to support discussion that could bring about changes in undergraduate training and lifelong learning programmes.

The organisation of the management of digital data in individual institutions

As expected, the experiences of the respondents differed depending on the size of the institutions. In smaller institutions, the management of digital data includes more types of activity, and digital data managers have more responsibility. The technology management (system administration) and content management departments are reflected in the organisational structure as well as practice in the majority of institutions. Yet in smaller institutions, part of the role of the IT department (system administration, backups, installation and testing new versions, OS management, connecting HW and networks, etc.) is performed by employees of the department also responsible for the digital content (digital library content management). The scope of activities performed by digital collection content managers differs greatly in different institutions.

Larger institutions may have separate departments for digital libraries, whose employees look after data and metadata (production, metadata creation, access) at content management level. The activities connected with application management – application servers, databases and server operating systems, or backing up or automating data transfers and conversion – are typically performed by employees in the IT or infrastructure department. The institutional culture then determines how effective or ineffective cooperation between IT and the digital collection content managers is.

The content management departments of digital libraries are gradually drawing in people who work with digital data elsewhere in the institutions, often in informal roles. In some places, “everything digital” originally fell under IT, yet with time it was shown that IT cannot address “content” or “XML”. A search was thus undertaken to see who would take over content management. Elsewhere, the management of digital data was more tied to production or acquisition, or with the management of library applications in general. It was only the growing volumes of data that applied pressure for the release of employees directly for digital library

[5] **http://www.enumerate.eu/**

[6] The selection of respondents was based on a circle of known people working in the given field, and was understandably not representative. During the seventh interview, it was clear that similar answers were being obtained for important groups of questions from multiple respondents, and so in this pilot phase the survey can be concluded.

content management[7]. In terms of the organisational structure, the majority of the respondents are relatively close to top management, while sometimes they report directly to the director. Almost all the respondents had the possibility of direct communication within the organisation, however not always completely without problems. Digital library managers would welcome more interest from the founders and better financial support for their work.

The link to digitisation was the common theme of the interviews. Digitisation remains an important topic for libraries, although "technology past its prime" is still of great concern for most libraries. Digitised data are not the first digital data in libraries, but provide great added value for users. These digital data are the daily bread for digital library managers, and many have experience with work on scanners or digitising lines. The managers do not usually create the descriptive metadata of digitised data themselves, but rather are expected to provide orientation in standards, the principles of cataloguing work, and the ability to work with XML. The importance of the role of cataloguer does not disappear even in the context of digital collections.

Daily activities and the competencies of digital data managers

The management of digital data in libraries is performed by people with diverse professional histories and education. Only some of them have specialist librarian or informatics education. Almost all the respondents had at some point worked and studied concurrently, many perhaps unsuccessfully, and studied information science and librarianship. The management of digital data in libraries is performed by - in addition to librarians - historians, Bohemists, philosophers or high school students from technical colleges[8]. It is pleasing that they are all trying to understand librarianship, or are actively acquiring competency in other fields (programming, system administration, data analysis, law, and management skills). Although the respondents differed from one another, they were connected by a passion for what they do – they do it "from the heart". Everyone is trying to expand their skills, and are being supported in this by the institutions directly (they can attend organized classes, study while employed etc.)[9] or indirectly (they have space for self-study, and are acquiring experience from their colleagues).

So what do people actually do when managing digital data? As expected, they add content to digital libraries, import data after checking them, and validate, assemble and convert them. In the environment of the digital library they then maintain the data and metadata, and modify certain metadata. Many of them are connected with data production through digitisation (they directly scan, create data packages, or propose approaches for the processing of digitised data, help write projects or operate scanners, manage the people doing the scanning, and manage the processed data) or with acquiring content from originators (they conclude agreements with content authors and owners, acquire and modify content, transfer it for further processing in the library, and sometimes even organise editorial processing for re-publication

[7] The interviews confirmed the hypothesis that an insignificant percentage of employees manage the digital libraries compared to traditional library activities.

[8] Two of the respondents, with very humanities-oriented university education, mentioned high quality teaching of computer science at grammar school as the cause of their understanding of "programming" being a normal part of daily work.

[9] This support has its limits, and most respondents would welcome the possibility to participate in high quality commercial courses, for example relating to the management of operating systems, yet noted that such training is very costly and of a long-term nature and is not something their employers usually offer.

or e-books). They manage digital library applications at a technical level (they install and maintain servers and operating systems, monitor and configure backups, etc.), even though they are not formally part of the IT department. Many of them participate in system development (they analyse and propose SW modifications, test new versions, manage supplyer development, prepare materials for tenders), while the roles of other respondents are shifting more towards promotion and strategic management[10].

It is noteworthy that all the responsibilities mentioned above are sometimes accumulated. A mismatch between original qualifications, field of study, and the activities that they people do, is the rule. A trained librarian programs, writes scripts, installs applications, and maintains server operating systems. A Bohemist or historian checks and enters data into a digital library, tests new SW application versions, and participates in analytical work for development. A trained philosopher manages SW implementation, etc. Competencies are acquired and transferred informally – through contact with more experienced colleagues, self-study or practice. Everybody has some type of foundation from which they derive their current confidence (training in librarianship, IT - even if only partial, contact with colleagues, previous work with scanners or in cataloguing), but all of them also feel that they are lacking something (a systematic and comprehensive course in a specific programming language, more experience with the Linux environment, more technical experience, etc.)

As can be seen from the above, the differences between “non-IT specialists” and “IT specialists” are hazy. IT activities are stealthily penetrating the daily routines of ever more people, even outside librarianship. In addition to purely librarian competencies (knowledge of MARC21, cataloguing rules, RDA etc.) the management of digital data also anticipates competency in work with XML (validation and conversion of metadata, linking metadata, automatic XML processing), work in the Linux environment (processing automation, bulk reports/exports, document searches, data transfers, md5 data validation, etc.), sometimes programming, work with digital formats, scanners, etc.

Conversion takes place mainly in the "non-IT specialist" > "IT specialist" direction. The interest of pure "IT specialists" in library issues usually ends with XML validation. Content standards and descriptive rules are not of much interest to "IT specialists", and if they decide to supplement their education with a specific level of librarianship training, they tend to fail because they feel that they are “just repeating what they already know”, and in the end give up at “history of libraries”, because an “IT specialist” cannot surely be expected to learn any lists by heart. Libraries logically have a larger percentage of people with librarian training available, and hence they are in turn available to the “IT specialist”. On the other hand, there are very few people in libraries with purely informatics training[11].

Users of digital libraries

How do the managers of digital libraries communicate with users? Digital library users surprisingly include the actual librarians, for example during retrospective cataloguing. In some institutions, queries reach digital library managers via a filter of reference librarians, in other

[10] Surprisingly, the respondents did not consider their own roles as technicians but as managers.

[11] In almost every interview we heard the complaint that the wage situation at state organisations prevents the creation of a stable high quality team. Benefits of the “benevolent treatment” type can help this to some extent.

cases users ask questions directly to the managers. Communication frequency is variable. Users point out errors (metadata, bad scans), are interested in exporting specific works, and are interested whether and when a specific work will be digitised. There are queries about the rights to the re-publication of works. Users also use data from digital collections for data-driven research or lay genealogical research. The systematic improvement of digital libraries using "user experience" methods was not mentioned by the respondents. Or even the participation in systematic marketing. If users find a digital collection themselves, have a problem and make contact, then our respondents are happy to help.

The roles of the respondents within the framework of the competency framework of digital curatorship (by Michal Konečný, 2016)

In his thesis, Konečný (2016) addressed the competency framework of digital curatorship. The importance of the roles[12] of eight Czech experts, inter alia, were evaluated during the preparation of this model. Thanks to this initial analysis, Konečný (2016) fine-tuned the terminology used and developed his own proposal for a competency model at basic, advanced and expert levels. In accordance with the competency model, he then developed a detailed proposal for the syllabus of a master's course **Introduction into Digital Curatorship**.

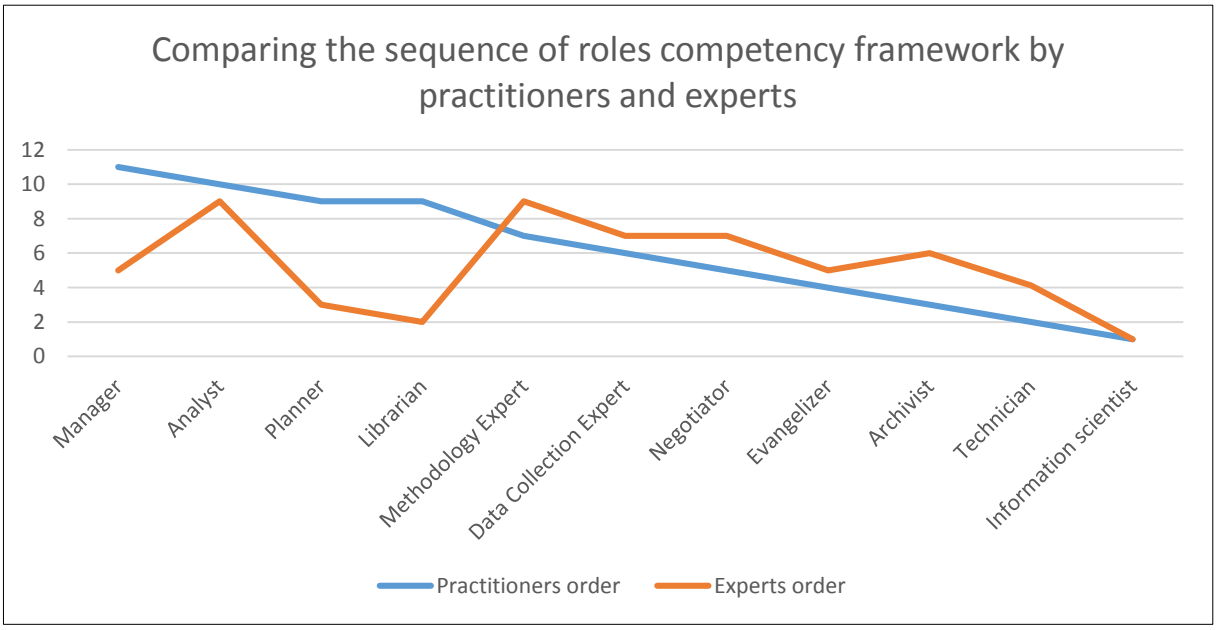


Figure 1: Comparing the sequence of roles comptency framework by practitioners and expert. Resource of the terms of role typology: Konečný, 2016.

The table of roles was also submitted to the fifteen respondents who manage digital libraries daily. Both practitioners and experts see the competencies of digital curatorship fairly

[12] The characteristics of the individual roles are available in Konečný's work.

consistently, while the sequence of the roles differs slightly. The experts considered the role of *methodology expert*[13] to be the most important, while practitioners the role of *manager*[14].

(Konečný, 2016) realises that his competency framework is too broad, something also noted by the Slovak study by Androvič, Ciglan, Matúšková (2016). Most respondents did not exclude any of the roles. There are differences between experts and practitioners in the evaluation of the importance of the roles of manager, planner and librarian, which is probably related to their differing assignments. Experts assessed the importance of competency for a digital curator in general, while practitioners did so considering their personal roles and experience.

Conclusion

So, is digital curatorship practiced in Czech libraries? Our survey did not unequivocally answer this question. Konečný (2016) defines digital curatorship primarily with regard to *long-term preservation (LTP)*, while our respondents primarily address the management of digital data. The creation of digital libraries in the Czech Republic is methodologically centralised as regards LTP, with small institutions expecting the big players – for example, the National Library of the Czech Republic – to address this issue. The management of digital data in libraries is only now being formalised. Not all libraries have optimal organisational structures to enable them to effectively employ their personnel capacities, or sufficient employees with the right competencies. The people who manage digital data in libraries deserve our admiration and support. They should have the possibility to participate in further systematic education.

The effective management of digital data today primarily requires the creation of conditions in the library environment enabling several active and enthusiastic people to work. An emphasis on sustainability, documentation, and continuous development and financing is desirable. Digital data require continuous attention - one cannot stop for a year or two, unlike with retrospective cataloguing or the fund revision. The use of all the experience of the digital pioneers will bring many valuable findings to theory, training and practice in librarianship. We heard from almost all the respondents that they would need more confidence when working with data on Linux servers. Let this then be the concrete conclusion of the article – we would be glad if all digital curators and librarians were granted this.

References

ANDROVIČ, A., Ivan CIGLAN, and J. MATÚŠKOVÁ, 2016. Digitálne pramene – webharvesting a archivácia e-Born obsahu. *IT lib* [online]. (2), 5-12 [cit. 2016-09-06]. ISSN 1336-0779. Available from: **http://itlib.cvtisr.sk/archiv/2016/2/digitalne-pramene-webharvesting-a-archivacia-e-born-obsahu-digital-resources-webharvesting-and-e-born-content-archiving.html?page_id=3177**.

[13] Focuses on current digital curatorship trends and standards. Prepares conceptual proposals for the introduction of a curatorial approach to long-term preservation. Reviews and evaluates current practices in their institutions from the perspective of good practice.

[14] Organises and manages activities connected with the lifecycle of digitally stored information in accordance with the needs of the institution. Selects, checks and evaluates practices that guarantee the applicability and sustainability of information. Communicates with stakeholders.

ČESKÁ REPUBLIKA, 2010. *Nařízení vlády č. 222/2010 Sb., o katalogu prací ve veřejných službách a správě. Účinnost od 1. 10. 2010* [online]. [cit. 2016-09-06]. Available from: **http://www.mpsv.cz/files/clanky/8980/Katalog_praci_UZ_1_10_2010.pdf**.

HIGGINS, S., 2011. Digital curation: the emergence of a new discipline. *International Journal of Digital Curation* [online]. **6**(2), 78-88 [cit. 2016-09-06]. DOI **10.2218/ijdc.v6i2.191**.

HOUŠKOVÁ, Z., 2012. Knihovnické profese v Národní soustavě povolání a v Národní soustavě kvalifikací. *Knihovna plus* [online]. **8**(2) [cit. 2016-09-06]. ISSN 1801-5948. Available from: **http://knihovna.nkp.cz/knihovnaplus122/housko.htm**.

JILEČKOVÁ, Š., 2015. *Vzdělávací programy pro oblast digitálních knihoven a digitalizace na školách informační vědy a knihovnictví v USA* [online]. Praha, [cit. 2016-01-27. Available from: **https://is.cuni.cz/webapps/zzp/detail/122206**. Diplomová práce. Univerzita Karlova. Vedoucí práce Barbora Drobíková.

KIM, J., 2015. Competency-based curriculum: an effective approach to digital curation education. *Journal of Education for Library and Information Science* [online]. **56**(4), 283-297 [cit. 2016-09-06]. DOI: 10.12783/issn.2328-2967/56/4/2. ISSN 0748-5786. Available from: **http://dpi-journals.com/index.php/JELIS/article/view/1573/1388**.

KONEČNÝ, M., 2016. *Návrh kompetečního modelu a kurikula digitálního kurátorství* [online]. Brno, [cit. 2016-01-27]. Available from: **http://is.muni.cz/th/426710/ff_m/**. Diplomová práce. Masarykova univerzita. Vedoucí práce Miroslav Bartošek.

MADRID, M. M., 2013. A study of digital curator competences: a survey of experts. *The International Information and Library Review* [online]. **45**(3-4), 149-156 [cit. 2016-09-06]. DOI: 10.1016/j.iilr.2013.09.001. Available from: **http://linkinghub.elsevier.com/retrieve/pii/S1057231713000106**

Návrh národní koncepce dlouhodobé ochrany digitálních dat pro knihovny, 2014. *Ústřední knihovnická rada ČR* [online]. [cit. 2015-03-31]. Available from: **http://files.u-k-r.webnode.cz/200000139-36130370e3/2_Navrh_koncepceLTP2014_verze_12-11.docx**.

POMERANTZ , J., S. OH, S. YANG, E.A. FOX and B.M. WILDERMUTH, 2006. The core: digital library education in library and information science programs. *D-Lib Magazine* [online]. **12**(11) [cit. 2016-09-06]. ISSN 1082-9873. Available from: **http://www.dlib.org/dlib/november06/pomerantz/11pomerantz.html**

PROM, C., 2011. Making digital curation a systematic institutional function. *International Journal of Digital Curation* [online]. **6**(1), 139-152 [cit. 2016-08-01]. DOI: 10.2218/ijdc.v6i1.178.

RAY, J., 2009. Sharks, digital curation, and the education of information professionals. *Museum Management and Curatorship* [online]. **24**(4), 357-368. [2016-09-06]. ISSN 0964-7775. DOI:10.1080/09647770903314720.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

THE CATALOGUES OF RECORDS COMPANIES OF EARLY 20TH CENTURY

Filip Šír

filip_sir@nm.cz

National Museum, Czech Republic

Gabriel Gössel

goessel@volny.cz

National Museum, Czech Republic

Abstract

Grey literature often contains documents such as publisher or corporate catalogues. Included among these are catalogues of record label companies, which provide evidence of all the sound documents issued for sale. Recordings are part of the collections of many institutions, but few of them also own these secondary information resources. Therefore, they are not able to contribute to the overall view of the cultural heritage of the era when the sound industry began and developed rapidly.

Keywords

Record Label Companies, Publishers´ Catalogues, 20th Century, Discography, Audio Documents

This work was financially supported by Ministry of Culture of the Czech Republic (DKRVO 2016/47, National Museum, 00023272).

Introduction

The introduction to an article on grey literature would - in most cases - begin in a similar fashion, namely with an explanation of the term 'grey literature' itself. Therefore, before the actual text of our article, we have decided to highlight three perspectives associated with the term 'grey literature'. We will intentionally sort them by creation date. We will deliberately try to put these three resources in a single location, as it is important to understand that even though we know or come across the term 'grey literature', in the 20th and 21st Centuries there is no institution here with any connection to the issue of audio documents, respectively the catalogues of gramophone companies.

1993 – Czech Terminology Database of Library and Information Science (TDKIV)

Documents that are not published in the usual manner and are therefore not available on the regular book market (e.g. theses and dissertations, research reports, internal documents, official publications, etc.) There are specialised information systems (e.g. the SIGLE database)[1] for searches and distribution of grey literature.

2006 - Grey literature in the Internet age

Both the concept and the content of the theme of 'grey literature' are rather ambiguous. Exhaustively defining grey literature is not easy to do, and a general description of this literature is preferred. Most commonly it is described using the statement that it is "publications that are not available through normal bookstore sources and methods". It is also often labelled 'unconventional' or 'half-published' literature.[2]

2008 – Grey literature

Grey literature, or unpublished or half-published literature, is information produced at all levels of government, academic, business and industrial institutions in both electronic and printed form, not having undergone the standard publishing process and not distributed in the standard sales networks, i.e. issued by institutions whose main activity is not publishing.[3]

Regarding the above terms, for our purpose it is important to mention that they do not explicitly define texts related to audio documents, meaning they do not include the catalogues of gramophone companies, sources of discographic data that should be retained just like university works, etc. If we combine this with the fact that here there is no institution that manages these historic documents, we come to the question of whether we should be drawing attention to this problem through the only possible solution: namely that we explain in detail what these catalogues were used for, what is found in them, and localise their incidence

[1] MATUŠÍK, Zdeněk. Grey literature. In: *KTD: Czech Terminology Database of Library and Information Science (TDKIV)* [online]. Prague: National Library of the Czech Republic, 2003- [cit. 2016-08-15]. Available from: **http://aleph.nkp.cz/F/?func=direct&doc_number=000001056&local_base=KTD**.

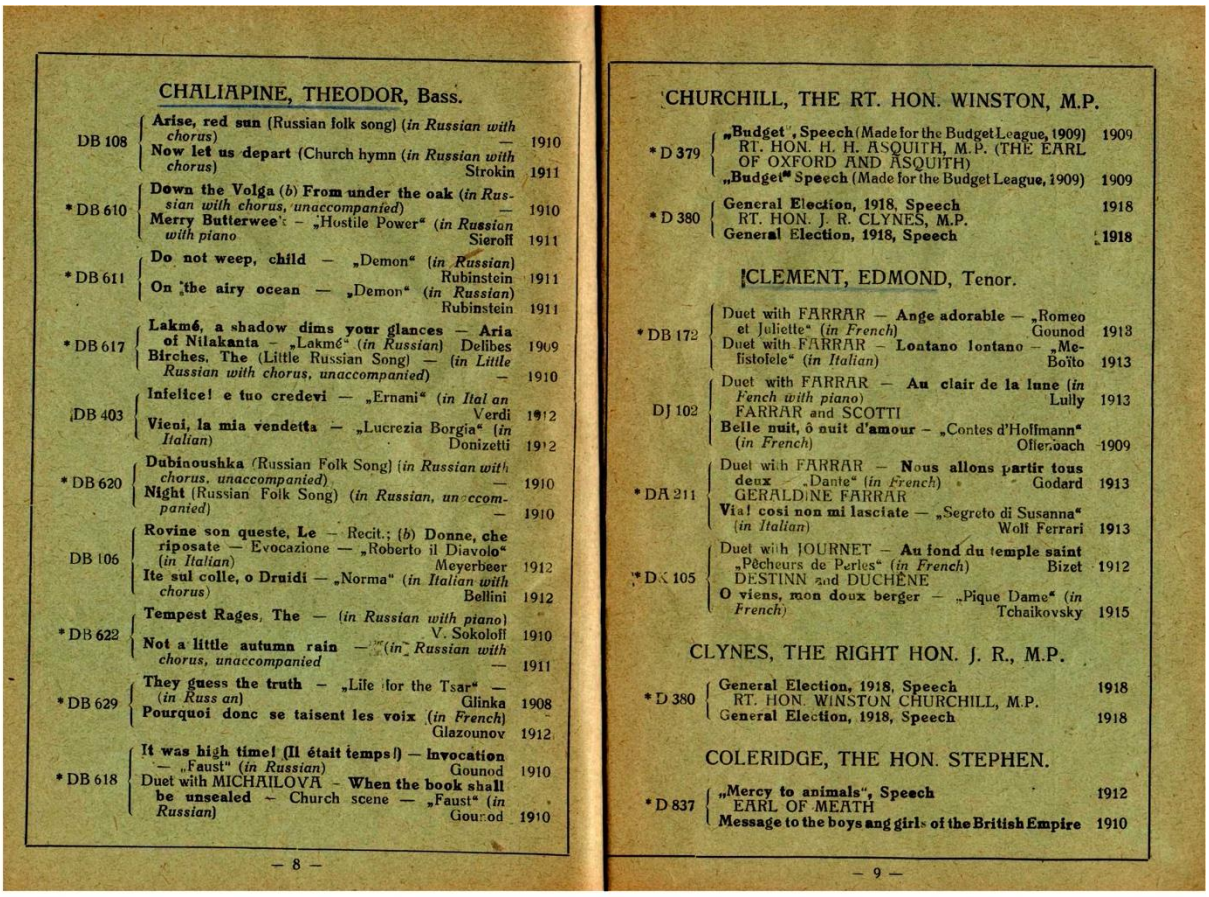
[2] MYŠKOVÁ, Petra. Grey literature in the Internet age. In: *Contemporary Libraries 2006*. Brno: Czech Association of Libraries, 2006. p. 279-284. ISBN 80-86249-41-7. Available from: **http://www.sdruk.cz/sdruk/publikacni-cinnost/clanek/knihovny-soucasnosti-2006-sbornik**.

[3] Šedá literatura. *Národní technická knihovna* [online]. Prague: NTK, 2008 [cit. 2016-09-13]. Available from: **https://www.techlib.cz/cs/2947-seda-literatura**.

primarily outside cultural heritage institutions. In our case, this means contacting private collectors and attempting to rescue these often even primary and unique resources for the creation of discographic works.

When examining the history of an audio recording and completing data about existing audio recordings, researchers ideally want all the data carriers in question physically available, in our case phonograph cylinders and shellac discs. Such an ideal case cannot of course ever exist, as the release of audio recordings – especially in the early history of audio recording – was an activity that lacked any global, let alone regional, coordination, nor could it have been coordinated. The activities took place exclusively according to the law of supply and demand, the financial strength of the individual producers and, last but not least, the potential purchasing power of the target market for which the carriers in question were intended. The record books of sometimes already long-disappeared gramophone companies were usually not preserved, whether because of changes in political regime, wars, or frequent changes of ownership in the firms themselves. Similarly, the sales catalogues of former gramophone companies were rarely preserved, even if they are practically the only sources – and in addition often rather unreliable – enabling knowledge of our cultural past, as reflected – even if often in a markedly distorted fashion – precisely through historical audio recordings. Catalogues and lists of historic phonograph cylinders and gramophone records, once a worthless single-use commodity, are today an indispensable tool for compiling discographies of individual performers or compiling an inventory of the overall production of a given gramophone company in a given time period.

The strengths and weaknesses of using company catalogues and other promotional materials can be illustrated through the example of compiling the first comprehensive discographic publication on recordings on gramophone records of the Czechoslovak gramophone company Esta, active from 1930 to 1946. The two most important pieces of data needed to determine the exact date of a recording are provided primarily by the matrix and secondarily by the order number. Both these numbers are usually furnished with various prefixes and suffixes, knowledge of which also helps us identify the recording technique, whether it is an initial or subsequent recording, the number of the pressing machine, and the price category for the resulting recordings. If we do not have physical moulding of the record in question, from which these data can be read, we can only work with the order number under which the recording was placed in the catalogue – if, of course, it has been preserved. The order numbers of records were usually published in ascending order. In the case of Esta records, however, they were added at random, due inter alia to irrational decisions taken by the various managers responsible for the creation of the repertoire at different times. In order to successfully compile a discography of this company, we were therefore forced to work both with as many actually existing records as possible, and also with all available catalogues and other promotional materials from Esta, while at the same time gradually eliminating the numerous demonstrable errors that these materials contain.



CHALIAPINE, THEODOR, Bass.

DB 108 — Arise, red sun (Russian folk song) (*in Russian with chorus*) — 1910
Now let us depart (Church hymn (*in Russian with chorus*) Strokin 1911

*DB 610 — Down the Volga (*b*) From under the oak (*in Russian with chorus, unaccompanied*) — 1910
Merry Butterwee's – „Hostile Power" (*in Russian with piano* Sieroff 1911

*DB 611 — Do not weep, child — „Demon" (*in Russian*) Rubinstein 1911
On the airy ocean — „Demon" (*in Russian*) Rubinstein 1911

*DB 617 — Lakmé, a shadow dims your glances — Aria of Nilakanta – „Lakmé" (*in Russian*) Delibes 1909
Birches, The (Little Russian Song) — (*in Little Russian with chorus, unaccompanied*) — 1910

DB 403 — Infelice! e tuo credevi — „Ernani" (*in Ital an* Verdi 1912
Vieni, la mia vendetta — „Lucrezia Borgia" (*in Italian*) Donizetti 1912

*DB 620 — Dubinoushka (Russian Folk Song) (*in Russian with chorus, unaccompanied*) — 1910
Night (Russian Folk Song) (*in Russian, unaccompanied*) — 1910

DB 106 — Rovine son queste, Le — Recit.; (*b*) Donne, che riposate — Evocazione — „Roberto il Diavolo" (*in Italian*) Meyerbeer 1912
Ite sul colle, o Druidi — „Norma" (*in Italian with chorus*) Bellini 1912

*DB 622 — Tempest Rages, The — (*in Russian with piano*) V. Sokoloff 1910
Not a little autumn rain — (*in Russian with chorus, unaccompanied* — 1911

*DB 629 — They guess the truth – „Life for the Tsar" — (*in Russ an*) Glinka 1908
Pourquoi donc se taisent les voix (*in French*) Glazounov 1912

*DB 618 — It was high time! (Il était temps!) — Invocation — „Faust" (*in Russian*) Gounod 1910
Duet with MICHAILOVA – When the book shall be unsealed – Church scene — „Faust" (*in Russian*) Gounod 1910

— 8 —

CHURCHILL, THE RT. HON. WINSTON, M.P.

*D 379 — „Budget", Speech (Made for the Budget League, 1909) 1909
RT. HON. H. H. ASQUITH, M.P. (THE EARL OF OXFORD AND ASQUITH)
„Budget" Speech (Made for the Budget League, 1909) 1909

*D 380 — General Election, 1918, Speech 1918
RT. HON. J. R. CLYNES, M.P.
General Election, 1918, Speech 1918

CLEMENT, EDMOND, Tenor.

*DB 172 — Duet with FARRAR — Ange adorable — „Romeo et Juliette" (*in French*) Gounod 1913
Duet with FARRAR – Lontano lontano — „Mefistofele" (*in Italian*) Boïto 1913

DJ 102 — Duet with FARRAR — Au clair de la lune (*in French with piano*) Lully 1913
FARRAR and SCOTTI
Belle nuit, ô nuit d'amour – „Contes d'Hoffmann" (*in French*) Offenbach 1909

*DA 211 — Duet with FARRAR — Nous allons partir tous deux — „Dante" (*in French*) Godard 1913
GERALDINE FARRAR
Via! così non mi lasciate — „Segreto di Susanna" (*in Italian*) Wolf Ferrari 1913

*DA 105 — Duet with JOURNET — Au fond du temple saint „Pêcheurs de Perles" (*in French*) Bizet 1912
DESTINN and DUCHÊNE
O viens, mon doux berger — „Pique Dame" (*in French*) Tchaikovsky 1915

CLYNES, THE RIGHT HON. J. R., M.P.

*D 380 — General Election, 1918, Speech 1918
RT. HON. WINSTON CHURCHILL, M.P.
General Election, 1918, Speech 1918

COLERIDGE, THE HON. STEPHEN.

*D 837 — „Mercy to animals", Speech 1912
EARL OF MEATH
Message to the boys and girls of the British Empire 1910

— 9 —

Figure 1: Example of His Master's Voices records catalogue from 1927

So much for the example of using company catalogues in current discographic practice. Information about the first commercial audio carriers appeared through such catalogues concurrently with the placement of these audio carriers on the market. The large American company Columbia Phonograph Co. published its first catalogue of phonograph cylinders in 1891 – it contained a list of about 200 recordings. Two years later, this company's catalogue already had 32 pages. The first catalogue of recordings on gramophone records was issued around 1892 by the London-based company Parkins & Gotto, which imported, inter alia, the first gramophones and records of their inventor Emile Berliner to England. This catalogue comprised a strip of paper measuring around 20x12 cm – one side depicted gramophones with handles or "talking machines" accompanied by the following text: "This cheerful curiosity can bring endless joy to children of all ages. We have the pleasure to introduce to you a recital of the poem Shine, Star, Shine in a voice so comic that all you can do is laugh." If, however, the listener was actually induced to laugh at anything, it was primarily the strong German accent of the English spoken by Mr Berliner, who personally recited favourite American children's rhymes on his invention – meaning on gramophone records. Descriptive colour catalogues were issued in the USA at the end of the 19th Century by Gianni Bettini, a producer of today very rare phonograph cylinders with recordings of international opera stars. Every gramophone company that wanted to remain on the market paid great attention to the publication of catalogues of their recordings from the start of the 20th Century. These catalogues took different forms – from a simple sheet of paper with a list of several of the most popular songs through to carefully coloured hardback publications, presenting – in addition to an exhaustive list of all the company's released recordings – their other products as well, such as gramophones and their parts, but also many other, often curious additions: brushes for

cleaning the records, small oil containers with a special oil for lubricating the spring movements of the gramophones, petroleum solutions for removing the recording from wax cylinders (so they could be recorded on again), various scissors or tools for sharpening bamboo or steel needles, several types of needle packaging for differing reproduction volume, tools for establishing the gramophone turntable speed, etc. Even if such catalogues were issued mainly in the years before World War One, and many in huge quantities, only very few of the oldest ones have been preserved.

And what was the situation like in this country? The first ever discovered list of Czech recordings on phonograph cylinders and gramophone records is contained in a relatively extensive catalogue from the wholesaler Bial & Freund based in Wroclaw (then Breslau), issued in the autumn of 1901. In addition to hundreds of recordings from its international repertoire, it also presented over two dozen "Böhmische Gesänge" [Bohemian Songs] performed by anonymous interpreters. Discographic research has so far succeeded in identifying around half of these recordings – they were phonograph cylinders from a German branch of the American Columbia recorded in Berlin by tenor Otakar Mařák and soprano Josefina Krausová. A certain Siegfried Adler has been identified as another Czech singer, of whom unfortunately history leaves us no trace.

The oldest discovered catalogue of Czech recordings on gramophone records from the then largest gramophone company in the world, The Gramophone Co. Ltd., boasting the colour picture of Nipper the dog on the cover, dates from 1909, and in 30 pages provides primarily a list of recordings from Czech opera singers, including their photographs. More Czech catalogues from this company were released four times a year from 1911. They usually had around one hundred A5-format pages and presented, in addition to hundreds of Czech songs, also thousands of recordings from its international repertoire, which anybody interested could order from one of its authorised importers of records from the English parent company. Until 1915, moreover, two-page leaflets with lists of the latest news from Czech interpreters on records from the parent company were released every month as a rule. This practice was stopped in 1915, as during World War One the British Gramophone Company – as well as other gramophone companies of the time – did not record any new Czech songs. In the post-war period, the activity of the company was renewed in 1921, when Karel Hašler was appointed director of its Czechoslovak branch. From that year, relatively extensive catalogues of the records from this brand were released in Czech every year, showing thousands of songs from its international repertoire. New releases usually contained a warning that the current list cancelled all earlier releases. In 1930, for example, such a A5-format list contained over 300 pages. Moreover, until 1939, the Czechoslovak branch of the Gramophone Company was engaged in such publishing, releasing in this country primarily records from the His Master's Voice and Columbia brands, and also as a rule monthly leaflets with several pages informing about news in the repertoires of this brand's recordings.

From 1905, the large French company Pathé produced a Czech programme on phonograph cylinders, and later also on records. Pathé recorded such songs until 1908 in Vienna with Czech artists with long-term engagements at the local theatres or opera houses. This is one of the reasons why some of them no longer recorded any songs for other gramophone companies present on the Czech markets, which did not record Czech programmes outside the Czech lands. The Pathé phonograph cylinder catalogue from 1906 also shows several Czech songs performed by a certain Bronislawa Wolske, a soprano probably of Polish origin. The last discovered catalogue of recordings from this company dates from 1912, while after

the war Pathé did not reopen its offices in the Czechoslovak Republic. Another large foreign company with an extensive Czech repertoire was the Anglo-German-Italian International Talking Machine Record, which in 1904 was the first company in the world to release double-sided records under the Odeon brand. The catalogues of Czech recordings on such records brought new items four times a year, while for distributors of records and gramophones of this and other brands of the German parent Lindström, a company quarterly entitled "Lindström zpravodaj" was moreover already published in Czech from 1902.

Various periodicals or irregularly issued printed products from diverse wholesalers, usually posing as "exclusive" or "general" representatives of some of the large foreign companies present on the Czech or Moravian market, are also of interest to researchers. Prague wholesaler Josef Vrba issued relatively informative catalogues every year on glossy paper, featuring many photographs. In addition to a list of phonograph cylinders and gramophone records from various producers, they also provided an extensive range of parts for gramophones and phonographs, including a range of other accessories for storing or transporting phonograph cylinders and gramophone records. The Brno-based company Jarušek & spol. even issued the large-format periodical Jaruškovy besedy containing a series of educational articles on the theme of reproduced music, and lists of phonograph cylinders and gramophone records with recordings by foreign interpreters, which this company exported all the way to Bosnia, for example. Such printed materials often contained false advertising, incorrect data and denigration of the competition. Hence, for example, the company Josef Kukla asserted in advertisement that its "new double-sided Mozart disks last twice as long as other disks" – in fact they were fire-sale disks of the Lyrophon brand with a neutral label to which Mr Kukla merely attached his own paper label. Often, for example, the fact that the store in question was "Czech and Christian" (Landiš, Prague) was emphasised on such leaflets, or they encouraged potential interested parties to consider why they should "support Jewish traders from Vienna or Kraków when we have a good domestic factory?" (Jarušek, Brno).

The catalogues and lists that the producers delivered directly to their sellers, or that were printed out by the domestic representatives of the company in question, are in principle the most useful research tools. The most important data they contained was the order numbers of the records, data about the interpreter, and the names of the songs. Until the end of the era of mechanical disk recording – meaning until around the end of the 1920s – Czech catalogues of records of the German company Homokord also presented the matrix numbers of the individual recordings. These are something like the birth certificate number of each recording, using which it is also usually possible to deduce the exact date of the recording itself. The opening pages of the company catalogues also usually presented the principles of the labelling of the records – in the case of HMV labels, for example, the prefix AN for the order number was used for records 30 cm in diameter, and AM for records 25 cm in diameter. The form of the prefix was also determined by the retail price of the record, and in addition the colour of the label usually reflected the price category of the record. The retail price of HMV records 30 cm in diameter with a regular garnet colour label was CZK 22.50 in the mid-1930s, while more valuable recordings by foreign interpreters used a black label (CZK 28), and the most expensive records had a white label and sold for CZK 80.

As regards the arrangement of the individual musical genres in the record catalogues, from the very start the practice was basic division into orchestral and song pieces. The orchestral pieces were further divided into marches, waltzes, polkas, mazurkas, hymns, concert and characterful pieces, folk dances, symphonic music, operas, operettas and modern dance.

These were further split, for example, into foxtrots, tangos, Charlestons and other kinds of dances. The songs were split into recordings of tenors, sopranos, duets and choirs. Instrumentals were classified by instruments, and usually began with solo violins and piano pieces through to the once popular zithers, xylophones, piccolos and horns. Other sections presented humorous pieces, fairy tales and Christmas pieces.

The catalogues of gramophone records were usually offered to regular customers of a company free of charge at stores. In the case of the sending of records by post with cash on delivery, which was common at the time, each consignment contained a certain minimum number of records supplemented with the latest company catalogue and other promotional materials. From the end of the 1920s, gramophone companies began – in addition to regularly published catalogues of new items and accessories – to also issue various thematic publications focusing, for example, on recordings of opera and symphonic music, film music, fairy tales for children, etc. The frequency of the release of promotional materials also increased: for example, from the mid-1930s, Esta was already publishing multiple-page leaflets presenting new items every month. As regards the print runs of such materials, this depended mainly on the number of distribution stores for the gramophone records of the company in question, which understandably had to have a reasonable amount of them available for their customers. The gramophone records of some brands were sold exclusively in a few stores in the republic – for example, in 1933, Dixi records were only sold in the Czechoslovak Republic through less than ten Je-Pa stores, and between 1935 and 1937, Pallas records could only be purchased in Prague at two company stores. At the end of the 1930s, on the other hand, Esta records were sold through at least one hundred stores throughout the republic, and we can assume that Ultraphon, the largest gramophone company at the time, distributed its records to a much larger number of sales outlets.

More serious wholesalers usually retained the original order numbers for offered records of the respective labels in their catalogues, and so the reconstruction of the whole output of the company is thus relatively easy. Usually, however, in their catalogues, dealers presented their own order numbers under which they offered records from various producers. Such catalogues are unusable for research, as they usually do not allow the identification of a recording. For example, the largest Prague wholesaler, Jan Kettner, whose catalogues are paradoxically quite common, was known for this practice. Other complications for researchers are the frequent (at the time) changes in ownership, capital transfers and mergers of record producers into concerns, when a new owner began to re-press older recordings from the output of a company that had since disappeared on a newly introduced label, and even sometimes under different song titles, and/or with different names of interpreters or composers, different names of the accompanying orchestras and, in addition, in different combinations on both sides of the record. In its new promotional materials, a new publisher also sometimes used the company figurative mark of the original publisher, in which it replaced for example only the name of the original label with a new one. The catalogues of such records are then a highly dubious and confusing contribution to the work of a researcher, or for compiling the discographies of the individual gramophone companies.

The authors of the offered songs were only indicated in catalogues in relatively rare cases in the years before World War One. They began to be indicated more consistently only after 1918, when Czechoslovakia acceded to the principles of copyright law. The payment of royalties for the use of musical works by the press, or their publication on an audio carrier, began to be monitored in the 1920s by various organisations like OSA and AMMRE. From the

new editions of the catalogues of some companies, we can also see adjustments resulting from the changes in the political situation: so, for example, the same recordings of military wind music labelled in 1914 as "Imperial and Royal Infantry Regiment No 28" were after 1918 labelled as recordings of the "Music of Czechoslovakian Infantry Regiment No 28". The First Republic's "Music of the 5th T. G. Masaryk Regiment" became during the protectorate "Music of the Government Troops" and the names had to disappear from the titles of various compositions dedicated to First Republic politicians when re-released. In addition, after 1945, the catalogues of gramophone records usually contained forewords justifying the exclusion of "unsuitable" older recordings that had been "subject to revision from artistic and technical points of view and, if needed, replaced with new pieces…" – meaning undoubtedly "more politically correct", as we would say today.

Conclusion

In conclusion, let us only add that today, the catalogues of historic gramophone records, similarly to brochures and other informational materials from the former gramophone companies, represent not only valuable sources of information for every researcher, but also amusing insights into the past. The mentioned catalogues, as part of grey literature, rank among the important sources of information and attention should be paid to them, as they are primarily publications providing information about recordings that make up part of our musical cultural heritage. The effective registration of audio grey literature and its comprehensive monitoring are important – the question thus arises as to whether the solution is the systematic collection of historic documents and their protection and preservation in digital form in the National Repository of Grey Literature (NRGL).

References

MATUŠÍK, Zdeněk. Grey literature. In: *KTD: Czech Terminology Database of Library and Information Science (TDKIV)* [online]. Prague: National Library of the Czech Republic, 2003- [cit. 2016-08-15]. Available from: **http://aleph.nkp.cz/F/?func=direct&doc_number=000001056&local_base=KTD**.

MYŠKOVÁ, Petra. Grey literature in the Internet age. In: *Contemporary Libraries 2006*. Brno: Czech Association of Libraries, 2006. p. 279-284. ISBN 80-86249-41-7. Available from: **http://www.sdruk.cz/sdruk/publikacni-cinnost/clanek/knihovny-soucasnosti-2006-sbornik**.

Šedá literatura. *Národní technická knihovna* [online]. Prague: NTK, 2008 [cit. 2016-09-13]. Available from: **https://www.techlib.cz/cs/2947-seda-literatura**.

9th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2016 [cit. 2016-12-5]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**. ISSN 2336-5021.

National Library of Technology, 2016