conference·on·grey·literature·a
nd·repositories·conference·on·g
rey·literature·and·repositories
·**conference·on·grey·literature**·
and·repositories·conference·on·
grey·literature·and·repositorie
s·conference·on·grey·literature
·and·repositories·conference·on
·grey·literature·and·repositori
es·conference·on·grey·literatur
e·and·repositories·conference·o
n·grey·literature·and·repositor
ies·conference·on·grey·literatu
re·and·repositories·conference·
on·grey·literature·and·reposito
ries·conference·on·grey·literat
ure·and·repositories·conference
·on·grey·literature·and·reposit
ories·conference·on·grey·litera
ture·and·repositories·conferenc
e·on·grey·literature·and·reposi
tories·conference·on·grey·liter
ature·and·repositories·conferen
ce·on·grey·literature·and·repos
itories·conference·on·grey·lite
rature·and·repositories·confere
nce·on·grey·literature·and·repo
sitories·conference·on·grey·lit
erature·and·repositories·confer
ence·on·grey·literature·and·rep
ositories·conference·on·grey·li
terature·and·repositories·confe
rence·on·grey·literature·and·re
positories·conference·on·grey·l
iterature·and·repositories·conf
erence·on·grey·literature·and·r
epositories····**proceedings·2018**

CONFERENCE ON GREY

LITERATURE AND REPOSITORIES

Proceedings

National Library of Technology, 2018

English conference website

(**https://nrgl.techlib.cz/conference/11th-conference-on-grey-literature-and-repositories/**)

Czech conference website

(**https://nusl.techlib.cz/konference/11-rocnik-konference/**)

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

Editor: Mgr. Hana Vyčítalová

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

Bc. Petra Černohlávková, National Library of Technology

Mgr. Hana Vyčítalová, National Library of Technology

List of Reviewers:

Stefania Biagoni, Institute of Information Science and Technologies; National Research Council of Italy, ISTI-CNR

RNDr. Miroslav Bartošek, CSc., Masaryk University

PhDr. Eva Bratková, Ph.D., National Library of Technology

Libor Coufal, National Library of Australia

Dr. Jan Dvořák, Charles University

PhDr. Václava Horčáková, The Institute of History, Academy of Sciences of the Czech Republic

Mgr. Jan Hutař, Archives New Zealand

Ing. Martin Lhoták, Academy of Sciences Library

Mgr. Jindřich Marek, Ph.D., Charles University

doc. PhDr. Mgr. Richard Papík, Ph.D.

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

RNDr. Michal Růžička, Ph.D., Masaryk University

Małgorzata Rychlik, Adam Mickiewicz University in Poznań

Mgr. Václav Stupka, Masaryk University

Marcus Vaska, University of Calgary

Mgr. et. Mgr. Jan Vobořil, Iuridicum Remedium, z.s.

Mgr. Jan Zibner, Masaryk University

Obsah

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: https://nusl.techlib.cz/en/conference/conference-proceedings

ARE WE READY FOR THE FUTURE? IMPACT OF ARTIFICIAL INTELLIGENCE ON GREY LITERATURE MANAGEMENT

Dr. Dobrica Savić

linkedin.com/in/dobricasavic

Vienna, Austria

DISCLAIMER: Any views or opinions expressed here are those of the author and do not necessarily reflect those of his employer.

Abstract

Information management is one of many areas being affected by artificial intelligence (AI). From science fiction to Google's search algorithms, self-driven cars, chatbots, and factory robots, AI has become part of our daily reality. Many books, articles, and blogs have been written, elaborated, and debated in numerous fields and industries about the use of AI. Scientists like Stephen Hawking and many others, businessmen like Jeff Bezos and Elon Musk, politicians and managers have talked about AI from different perspectives and with different aims. Information technology developments impact the way we work, learn, communicate and go about our lives. This paper examines the potential impact of AI on grey literature (GL) management and is based on analysis of pertinent GL facets such as value, volume, variety, velocity, and veracity. The impact of AI on processing, sustainability and usability of GL management are given special attention. Examples of AI systems already implemented in similar fields or activities are offered. In conclusion, the paper presents possible solutions to challenges that GL managers could face in the near future.

Keywords
Grey literature, artificial intelligence, information technology, information management

Introduction

During the last few years, the term artificial intelligence (AI), has become omnipresent. It is being discussed in books, scientific articles, newspaper stories, government reports, parliament debates, court decisions, and ordinary conversations. From science fiction to Google's search algorithms, self-driven cars, chatbots, and factory robots, AI has become part of our daily reality. Scientists like Stephen Hawking and many others, businessmen like Jeff Bezos and Elon Musk, politicians and managers alike have made their own contributions. Information technology (IT) developments impact the way we work, learn, communicate and go about our lives. Information management is one of many areas being affected by various disruptive information technologies such as AI (Savic, 2017a).

This paper examines the potential impact of AI on information management (IM), specifically on grey literature (GL), starting from the analysis of pertinent GL facets such as volume, variety, velocity, veracity, and value. It continues with a review of the potential impact of AI on GL processing, sustainability, and usability challenges. It also offers parallel examples of AI systems already being implemented in similar fields or activities.

In conclusion, the paper will present possible solutions to challenges that grey literature managers will most likely face while trying to accommodate and benefit from new AI technologies. By increasing our knowledge about AI and other potentially disruptive technologies, we improve our chances to increase their relevance and potential benefits to our work.

Definitions

Two basic concepts considered in this paper which need clarification and explanation are *grey literature* and *artificial intelligence*.

Grey literature has been defined differently by a number of researchers, justifying Schöpfel's (2011) theory that GL is much easier to describe than to define. The most widely accepted and used definition is from the 12th International Conference on Grey Literature (GL12), held in Prague in 2010.

> *"Grey literature stands for manifold document types produced on all levels of government, academics, business and industry in print and electronic formats that are protected by intellectual property rights, of sufficient quality to be collected and preserved by library holdings or institutional repositories, but not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body".* (Farace and Schöpfel, 2010).

Although this definition focuses on important aspects of GL, it might need to be expanded to take into consideration new challenges brought about by new disruptive technologies, such as AI. In 2017, I proposed a new definition, which might help meet some of these challenges. According to this revised definition,

> *GL represents any recorded, referable and sustainable data or information resource of current or future value, made publicly available without a traditional peer-review process.* (Savic, 2017b).

Artificial intelligence (AI) and related machine learning (ML) applications[1] are systems that can think and act rationally, almost like humans. They are usually very costly and complex to develop maintain and deploy. Their power comes from a combination of many technologies and techniques, such as powerful parallel computer processing, deep learning, neural networks, and natural language processing (NLP). Initially, they appeared as rule-based or expert systems, but today's algorithms can understand, learn, predict, adapt and potentially operate autonomously. They are often built into physical devices (e. g., robots, cars, consumer electronics, and security systems); and into apps and services (e.g. virtual personal assistants, smart advisors, voice recognition, computer vision, translation, and finance). Applied in the area of information and knowledge management they become a powerful help in processing, organizing and disseminating data and information. Incorporated in some web apps, AI enhances the user experience by offering new smart and adaptive user interfaces.

One of the earliest yet still popular definitions of AI is that offered by Marvin Minsky[2] back in 1968. He defined artificial intelligence as “the science of making machines do things that would require intelligence if done by men”. Another definition that AI researchers believe will be valid for many years states that “artificial intelligence is the study of how to make computers do things at which, at the moment, people are better.” (Rich, 2010).

On the practical side, AI started with simple test applications, such as the famous Turing test, continued with games, like checkers, and later with expert systems which represented knowledge through rules. A good example of this was Deep Blue, an IBM developed system for playing chess that defeated world champion Garry Kasparov in 1997.

Machine learning, as part of AI, came into focus in the 80s. It allowed computers to learn how to recognize patterns and make predictions. This was a revolutionary move from the traditional hard-coding software programs to performing specific instructions to complete a task. Systems became dynamic and the need for programmers to make certain changes was eliminated.

The latest stage of AI is ‘deep learning’. It typically requires a lot of processing power and a large set of “training data”, through which the use of neural networks allows ‘intelligent behavior’. The three most popular forms of training are: unsupervised, supervised, and

[1] AI is the broader concept of machines being able to carry out tasks in a way that we would consider “smart”. Machine Learning is a current application of AI based around the idea that we should really just be able to give machines access to data and let them learn for themselves. (Marr, 2016)

[2] Marvin Lee Minsky (August 9, 1927 - January 24, 2016) was an American scientist in the field of artificial intelligence (AI), co-founder of MIT's AI laboratory, author of several texts on AI and philosophy, and winner of the 1969 Turing Award.

reinforcement training. Good examples of deep learning applications are those used for face and speech recognition, robotics, chatbots, self-driving cars, computer games, etc.

Regarding the level of AI complexity three types are generally acknowledged (Dickson, 2017):

- Artificial narrow intelligence
  (a system developed and trained for a particular single task within a limited domain)
- Artificial general intelligence
  (a system that can understand and reason about its environment as a human would)
- Artificial super intelligence
  (a system much smarter than the best human brains in practically every field).

All of the AI systems currently in use, or currently being developed, are at the initial, narrow level of AI. More powerful programs and computers, potentially quantum ones, would be needed to achieve the second, general level of intelligence. Some researchers believe that once that level is achieved, it will not be much longer until the level of super intelligence is reached.

The general impact of artificial intelligence

Artificial intelligence is here to stay and its impact is irreversible. The benefits and the magnitude of its potential power have been experienced by many people. The history of its progress is exemplified by a collection of remarkable achievements. In 1956, Arthur Samuels Checker program, developed for play on IBM’s 701, was introduced and in 1962 beat the checkers master. In 1966, ELIZA, one of the first chatbots, carried out conversations with people in natural language. In 1976, an expert system called MYCIN made a successful diagnosis of infectious diseases, while dealing with uncertainty. In 1986, an artificial neural network system NETtalk[3] read written English texts aloud. In 1977, IBM’s super computer Deep Blue defeated world chess champion, Gary Kasparov. 2009 brought the first Google self-driven car to the California freeway. In 2011, IBM’s Watson system beat two human experts at the Jeopardy TV game show. Just four years later Daimler-Benz demonstrated its first autonomous big rig truck on Germany’s autobahn, while Google self-driving cars drove over one million miles. In 2016, Google’s DeepMind AlphaGo computer program beat the world’s best Go player. (Ertel, 2017). These are just some of AI researchers many successes. Various countries, universities and research centers have placed their current priorities on AI projects. It is, therefore, reasonable to expect exponential future developments in this interesting area of research.

Even today, although at the initial or somewhat basic level of complexity, a number of operational AI applications in classic office environments strongly indicate the trends of their future impacts. Big companies such as Microsoft, Google, Facebook, Instagram, Twitter, and Disney are heavily involved with AI applications. Here are some examples of lesser known applications which demonstrate variety, depth, and the force of change to come[4].

[3] **https://www.nytimes.com/1988/08/16/science/learning-then-talking.html**

[4] Examples adapted from Dom Nicastro, 8 Examples of Artificial Intelligence (AI) in the Workplace‘, Dec 7 2017. CMSWire. https://goo.gl/s6bWGH and Bernard Marr, 27 Incredible Examples Of AI And Machine Learning In Practice. April 30 2018. Forbes. https://goo.gl/YX52TK

SAP CoPilot: Digital Assistant for the enterprise – using their phones, users can ask business related questions and the system offers an answer.

Deloitte: Automated document review with natural language processing - quickly reads thousands of complex documents, extracting and structuring textual information for better analysis.

AISense: Call, Meeting Transcriptions – records voice conversations and makes them searchable and easily accessible using automatic speech recognition, speaker identification, speech-and-text synchronization, and natural language processing.

WalkMe: AI for Software Training - enables business software to learn about the user's individual roles, habits, and actions.

ServiceChannel: Restaurant Facilities Management Aid - helps automate the repair and maintenance process, cut repair and maintenance costs, maintain compliance, and minimize risk. It manages contractors, work orders, preventative maintenance, assets, proposals, and invoices.

Niles: Learning Slack Conversations - listens and records conversations that happen within the Slack collaboration platform. Every time someone sends a message, it learns, so users can ask questions such as, "What products do we sell? What sizes? How much do we charge? Who's in charge of this department?" If it fails at an answer, users can keep the system up-to-date by providing the right answer.

Acculation, Inc.: AI Meets Social Media - uses data-driven processes to make decisions about content for social media. It can actually create the content.

BBC Talking with Machines - an audio drama that allows listeners to join in and have a two-way conversation via their smart speaker. They are prompted to answer questions and insert their own lines into the story.

UK Press Association RADAR (Reporters and Data and Robots) - robots write 30,000 local news stories each month fed with a variety of data from government, public services, and local authorities. The machine uses natural language generation technology to write local news stories that are not covered by humans.

American Express - relies heavily on data analytics and machine learning algorithms to help detect fraud in near real time, therefore saving millions in losses.

Impact of artificial intelligence on grey literature creation

There are various ways to look at the impact artificial intelligence will have on the creation of grey literature. This review takes an analytical approach by looking at each of the main facets of grey literature, taken from wider studies in the field of information and data management. As Figure 1 shows, five Vs are taken into consideration. They are variety, volume, veracity, velocity, and value.

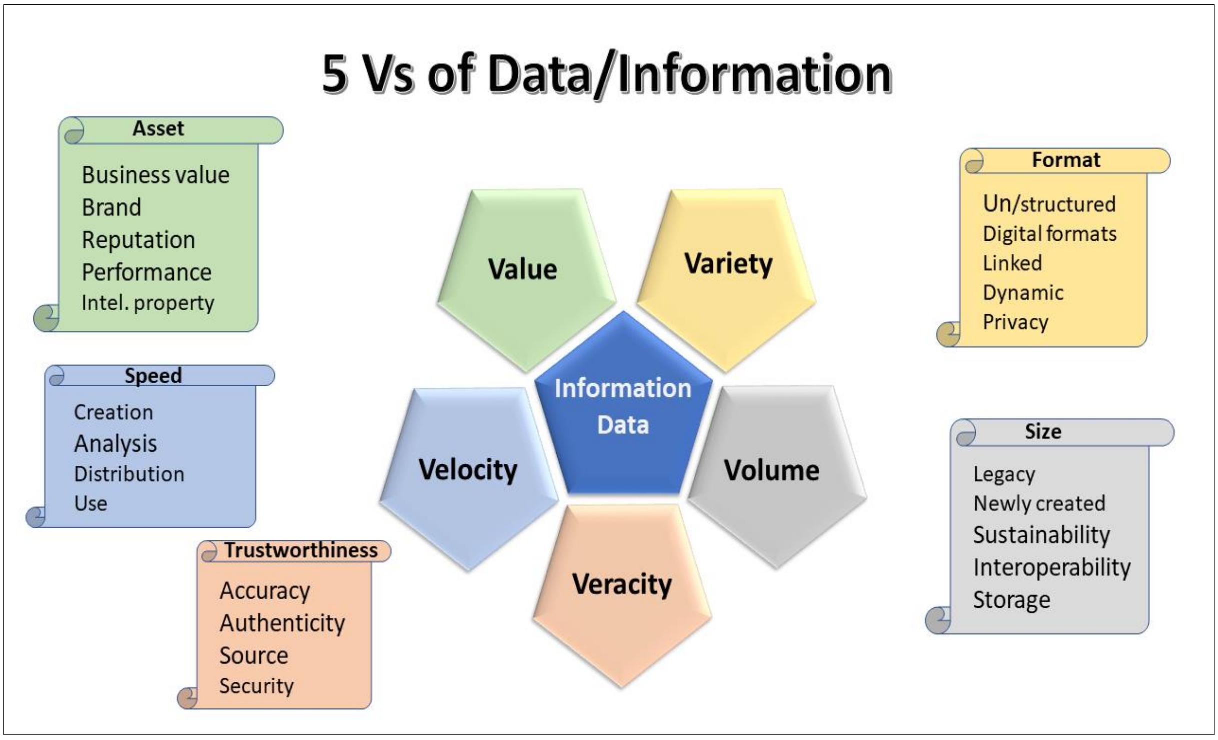
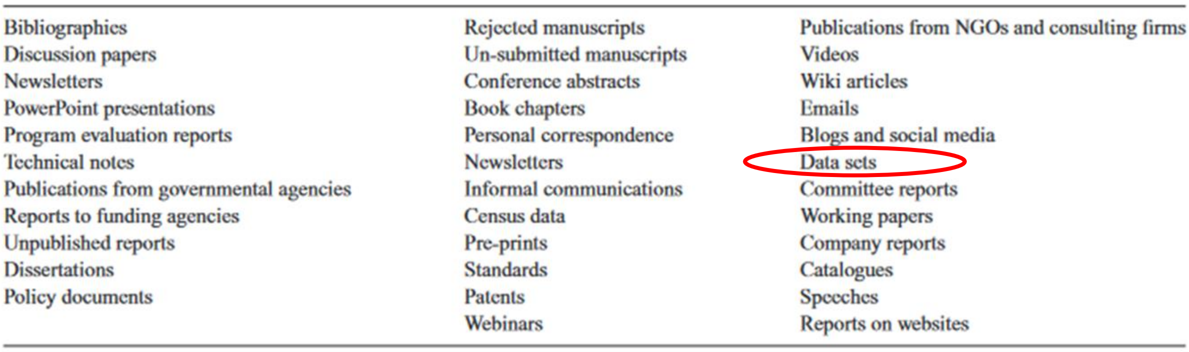


Figure 1: 5Vs of data/information

The variety of grey literature formats could experience a considerable impact from AI use, although as Figure 2 indicates, there are already a large number of identified formats. A more complete list is available at the GreyNet International website[5]. It lists over 150 document types specific to GL.



Bibliographies
Discussion papers
Newsletters
PowerPoint presentations
Program evaluation reports
Technical notes
Publications from governmental agencies
Reports to funding agencies
Unpublished reports
Dissertations
Policy documents

Rejected manuscripts
Un-submitted manuscripts
Conference abstracts
Book chapters
Personal correspondence
Newsletters
Informal communications
Census data
Pre-prints
Standards
Patents
Webinars

Publications from NGOs and consulting firms
Videos
Wiki articles
Emails
Blogs and social media
Data sets
Committee reports
Working papers
Company reports
Catalogues
Speeches
Reports on websites

Figure 2: Types of grey literature

If we examine only one GL type, namely 'data set' (marked above in red), this type alone typically includes a tremendous amount of data and information coming from the Internet of Things (IoT), Machine to Machine communication (M2M), self-driven cars, robots, sensors, security systems, surveillance cameras, and many other systems using AI. Estimates for the number of connected devices creating specific data trace vary by billions.[6] Such a huge number of devices, generating tons of data, in multiple formats, mostly unstructured and application specific, will represent a considerable challenge for GL researchers, practitioners,

[5] http://www.greynet.org/greysourceindex/documenttypes.html

[6] Gartner says that there will be some 20 billion connected devices communicating to each other by 2020. Allied Business Intelligence says more than 30 billion, Nelson Research says 100 billion, Intel says 200 billion, and International Data Co. says 212 billion.

and managers. Such highly contextual and software dependent data and information would be hard to collect and process, and even harder to make sense of and preserve for future use.

Closely related to single data sets is the question of connected multiple data sets, often referred to as linked data. Linked data represents the main ingredient of Semantic Web that's understandable not only to humans but also to computers. A good example of a large linked data set is DBpedia, which, in fact, makes the content of Wikipedia available in the Resource Description Framework (RDF), while including links to other datasets on the Web, such as GeoNames. By providing those extra links, the application offers much better access to knowledge and a more satisfying user experience. However, identifying and finding a role for classical GL management becomes a challenge.

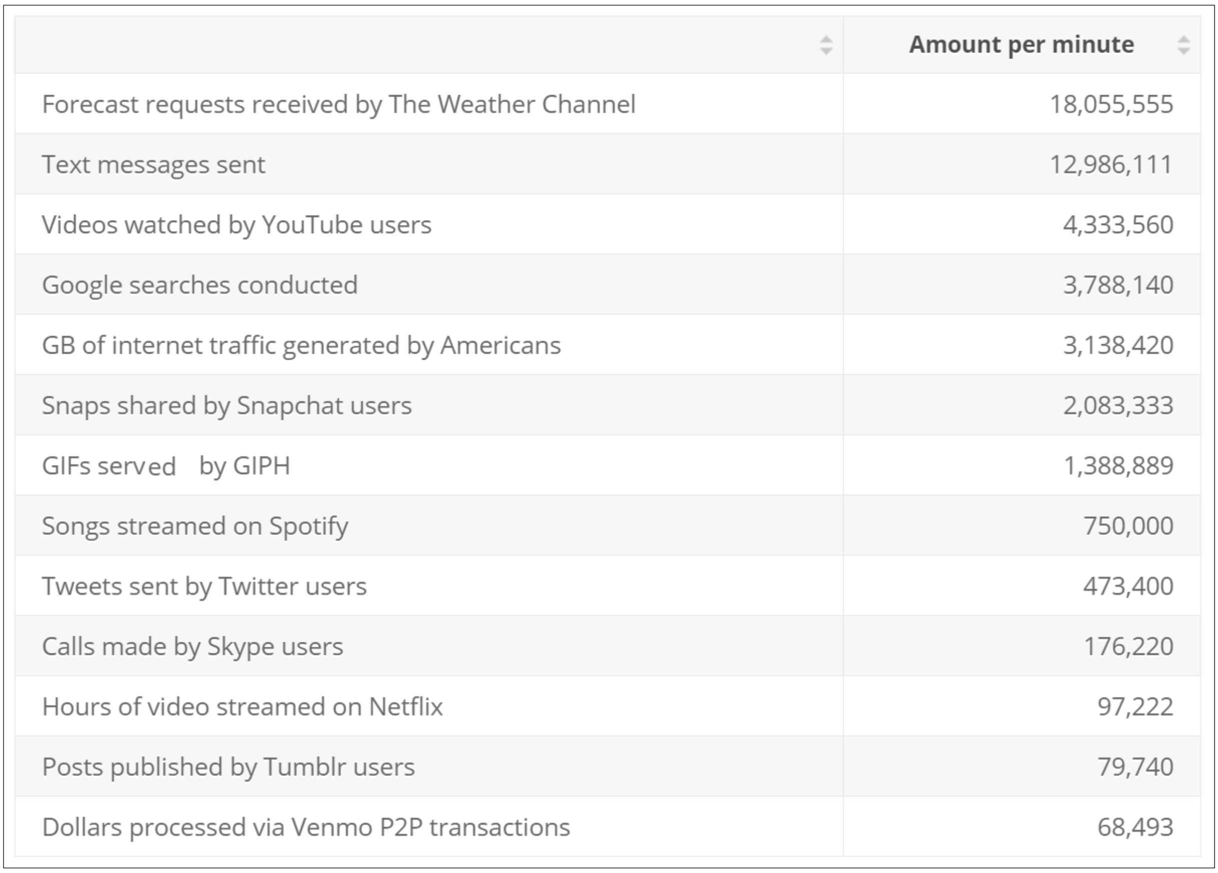
The volume of grey literature is the next area already undergoing a visible change which will be further impacted by the use of AI. According to available statistics, 2.5 exabytes of data are produced every day, which is equivalent to 250,000 Libraries of Congress. In comparison, the human brain has an estimated storage capacity of 1000 terabytes or one petabyte. 90% of all the data in the world has been generated over the last two years. There are 130 million published books around the world, with over 800,000 new titles added annually. At the same time, the digital world is moving towards increased use of mobile phones creating even more data. Currently, over half the world uses a smartphone. According to Cisco's (2017) prediction, the number of devices connected to IP networks will be three times as high as the global population in 2021. At the moment, the number of worldwide users of the four most popular messaging apps[7] have reached 4 billion.

If we mention just dissertations as one of more important GL types, Google Scholar hosts almost 4.3 million dissertations from all around the world, while ProQuest adds annually more than 130,000 new dissertations and theses to its largest dissertation database, ProQuest Dissertations & Theses (PQDT) Global[8].

Figure 3 gives an interesting statistical overview of some of the parameters regarding the volume of Internet traffic.

[7] WhatsApp, Facebook Messanger, WeChat, QQ Mobile. Source: **https://goo.gl/UYArhS**

[8] **https://www.proquest.com/products-services/dissertations/ProQuest-Dissertations-FAQ.html**



	Amount per minute
Forecast requests received by The Weather Channel	18,055,555
Text messages sent	12,986,111
Videos watched by YouTube users	4,333,560
Google searches conducted	3,788,140
GB of internet traffic generated by Americans	3,138,420
Snaps shared by Snapchat users	2,083,333
GIFs served by GIPH	1,388,889
Songs streamed on Spotify	750,000
Tweets sent by Twitter users	473,400
Calls made by Skype users	176,220
Hours of video streamed on Netflix	97,222
Posts published by Tumblr users	79,740
Dollars processed via Venmo P2P transactions	68,493

Figure 3: Media usage in an internet minute as of June 2018 (statista.com)

From the aspect of GL management, even more alarming than the above statistics is the fact that 56% of all internet traffic is from automated sources such as hacking tools, scrapers and spammers, impersonators, and bots. There are 269 billion emails sent and received each day, out of which 60% is spam. Still, the world is hungry for information which is nicely supported by the figure issued by Google that it processes daily over 6.6 billion queries, out of which 15% have never been searched for on Google before.[9]

If we combine the variety of GL formats with the volume, as partially described above, and the increased emergence of AI, the challenge that GL professionals are facing, and will continue to face in the future, is enormous. The questions of storage, sustainability, processing, usability, and many others are overwhelming. With current operational capacity, general interest, and available resources, the fear that most of GL will disappear or become unusable over time is well founded.

The increased volume of GL also merits the question of ways to measure its impact and popularity of the scientists and researchers. A common way was „counting the number of articles citing other articles, resulting in journal impact factors, normalized citation rates, and the h-index. Even those rare studies including conference papers are limited to published

[9] 100+ Internet Stats and Facts for 2018. Available from: **https://goo.gl/iUcnxc**

proceedings. Grey literature remains out of scope“ Schöpfel (2017). Major effort and additional resources are needed to demonstrate the value and extent of the GL impact.

The veracity of grey literature looks at its validity and trustworthiness. It is defined as "conformity to facts; accuracy; habitual truthfulness"[10]. In particular, it deals with GL accuracy, authenticity, information source, and security. Spam email, fake news, computer bots, botnets, web spiders, crawlers, viruses, plain misinformation and disinformation, they all represent multiple dangers that web users, including GL users, face today. Uncovering deception and estimating the veracity of information and data is difficult, in particular when prior background knowledge about content, context, or source is weak or not available.

The main assumption about establishing the veracity of some information is its originating source. If it is a well-known source with a long tradition of trustworthiness, information is usually regarded as reliable, although there are many cases of inadvertently created false information placed on the web or included in some documentation. Such cases lead us to conclude that, as users, we always need to be on the lookout for possible errors or false information. Multiple checkpoints, such as source, independent confirmation, a best practice used in preparing the information, and even intention, all need to be taken into consideration when establishing information veracity.

The use of artificial intelligence in almost any sphere will increase problems with defining the actual information veracity. We can look at two facets of the information created while relying on AI in the process of its creation. The first is the question of documented procedures, steps followed, inference paths, and decision justifications. Quite often, the whole process becomes a ‘black box’ where all we have is the input and the output, without any trace of the logic or reasoning used. Such machine-learning models are already having an impact on people’s lives. A system called COMPAS offers to predict an offender’s likelihood of reoffending and is used by some judges to determine whether an inmate is granted parole. Some suspect bias against minorities (Knight, 2017). Such AI machines do not offer any documented justifications and could display a strong potential bias, especially when there is a probability that the training data used was biased.

The velocity of grey literature refers to the speed of information creation, processing, analysis, distribution, and use. Figure 3 shows the amount of data and information created, but in addition to the amount, we also need to look at the speed at which this humongous amount of data is being created.

It is estimated that more data has been created in the past two years than in the entire history of the human race. The speed of creation results in zettabytes of stored information which, unfortunately, is barely being processed. Technical, physical, financial, and other challenges limit the possibilities for analyzing such a huge amount of data. Research shows that 99.5% of

[10] Oxford Dictionary. Available from: **https://goo.gl/AkG95E**

all data is not currently being analyzed and used (Bansal, 2014). This represents a big financial, business, and information loss for everyone involved.

This huge amount of information and data enables artificial intelligence and machine learning to turn data analysis from retrospective practice into a proactive approach to strategic decision making. AI can greatly increase the frequency, flexibility, and immediacy of data analysis across a range of industries and applications.The International Data Corporation (IDC) estimates that the amount of global data subject to data analysis will grow by a factor of 50 to 5.2 ZB in 2025; the amount of analyzed data that is “touched” by AI systems will grow by a factor of 100 to 1.4 ZB in 2025. (Reinsel et al., 2017).

As information grows in variety, volume, veracity, and velocity, business needs would focus on the information that has the most important value. Not all data is equally important to businesses or consumers, providing an opportunity for GL and other information managers to offer tools, expertise, and visible results to identify that specific value of information from the ocean of available data. The organizations that succeed during this transformation will be those that can successfully identify and take advantage of the critical subset of data that will have a meaningful, positive impact on user experience, solving complex problems, and creating new economies of scale (Ibid).

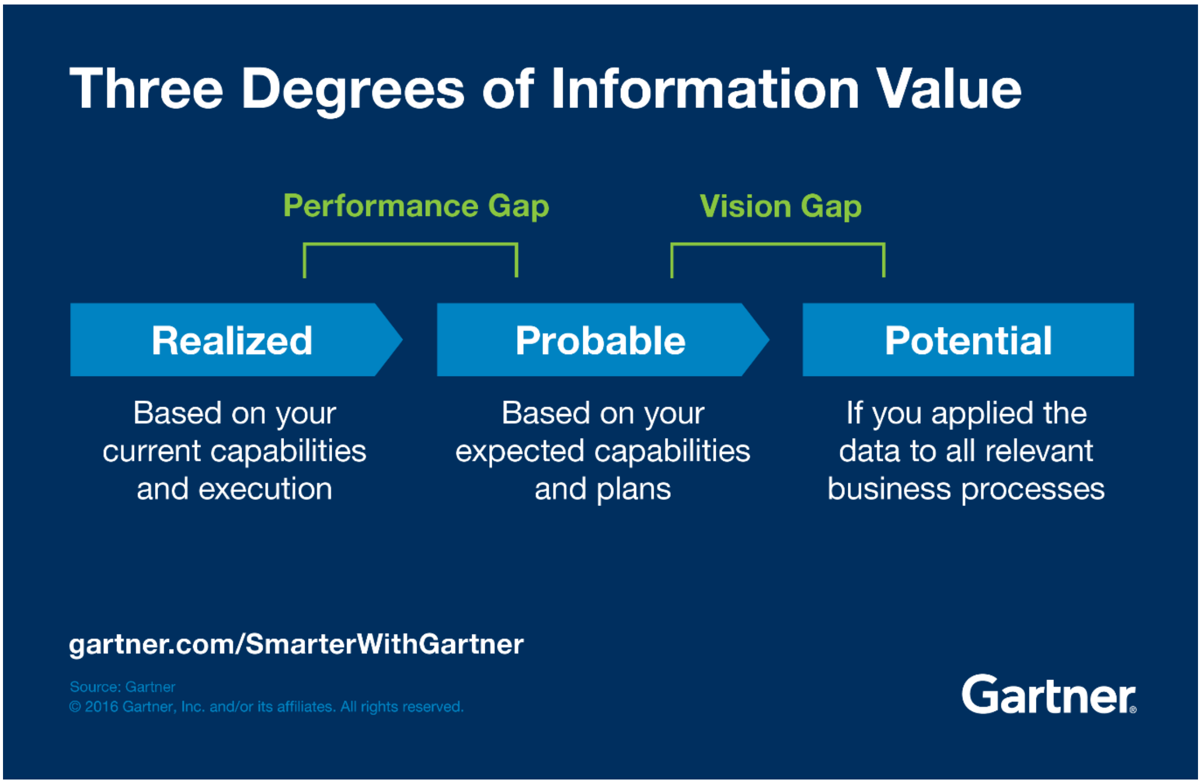
The value of grey literature, and the value of information in general, rarely finds its place on a balance sheet. Almost everyone agrees that information is an asset that costs millions[11], but hardly anyone can tell where the asset sits, its quantity, or even where it came from. It is difficult to measure, although many claim to own the asset while trying to avoid any accountability for it.

The glossary of the Queensland Government Chief Information Office offers a valuable and widely applicable definition of an ‘information asset’. It defines it as „An identifiable collection of data stored in any manner and recognized as having value for the purpose of enabling an agency to perform its business functions thereby satisfying a recognized agency requirement. Data or information that is referenced by an agency, but which is not intended to become a source of reference for multiple business functions is not considered to be an information asset of the agency. This is merely information.“ (Information Asset, 2017).

Many organizations and industries recognize information as a strategic business asset and Gartner predicts that by 2020 10% of organizations will have a highly profitable business unit specifically for making and commercializing their information assets (Pettey, 2017). They also claim that information assets have great potential, beyond the utility for which they were originally produced. Unlike most of your enterprise’s other assets, information isn’t depleted after it’s consumed. In order to utilize its value, they propose (see Figure 4) to review

[11] An asset is a resource with economic value that an individual, corporation or country owns or controls with the expectation that it will provide a future benefit. Assets are reported on a company's balance sheet and are bought or created to increase a firm's value or benefit the firm's operations. Investopedia **https://www.investopedia.com/terms/a/asset.asp#ixzz5QKbl1CnA**

performance and vision gaps that exist between the three levels of information value – realized, probable, and potential.



Three Degrees of Information Value

Performance Gap

Vision Gap

Realized

Based on your current capabilities and execution

Probable

Based on your expected capabilities and plans

Potential

If you applied the data to all relevant business processes

gartner.com/SmarterWithGartner

Source: Gartner

© 2016 Gartner, Inc. and/or its affiliates. All rights reserved.

Gartner.

Figure 4: Three degrees of information value

The performance gap is the difference between the realized value of an information asset and its probable value, while the information vision gap represents the difference between probable and potential information value (Levy, 2016).

Impact of artificial intelligence on grey literature management

In the era of intensive artificial intelligence use and machine learning, three main areas of grey literature management will be directly impacted. They include GL processing, sustainability, and usability.

GL processing and the related management tools used are directly impacted by the increased volume, variety, velocity, veracity, and value of the grey literature created. A single document and ad-hoc approach to management will be neither appropriate nor sufficient. What is required is the transparent inclusion of GL-type processing[12] during document creation, rather than post-processing. An additional requirement is to have a GL management system in place all the time, eliminating the need for any ad-hoc solutions or deviations from an already set-up plan.

[12] E.g. meta-data creation, retention scheduling, distribution channels, copyright, confidentiality, etc.

GL sustainability, the next challenge regarding the impact of artificial intelligence on grey literature management, includes three broad challenges – environmental and technical; economic and financial; and social or organizational. Preservation of documents, technical knowledge transfer over long periods, information continuity, technical operability, and usability are just some of the important aspects related to environmental sustainability. Economic and financial sustainability focuses attention on the availability of long-term adequate funding, public vs. commercial interests, and the future value of collected GL as it relates to the value of information. Social and organizational considerations emphasize the existence of multiple stakeholders, information ownership and governance, international cooperation, and also safety and security.

GL usability of large amounts of information generated by the use of AI creates an additional category of problems, such as the existence of adequate IT tools, the availability of qualified human resources, and the protection of intellectual property (IP) and privacy. Let's look at some of the challenges here. IT technology and tools are constantly changing, contributing to new software functionality, concepts, and expectations in quantum leaps, and making previous technology obsolete in almost no time. Related to this issue is the creation of dynamic vs. static information and documents, and their visualization (e.g. 2D, 3D, VR, AR). Many predictions about the impact of AI relate to job loss,[13] which translates to staffing requirements, but there is also the issue of the required technical skills, education, and training[14]. Intellectual property involves issues such as over-protectionism; open access and open science; and the role of current IP in helping world development, health, and innovation. Related to this are issues of privacy, including the protection of commercial information, and the protection of the sensitive public and personal information.

The above mentioned three areas of the expected impact of AI on grey literature management mainly deal with challenges. However, there are also some great opportunities that the use of AI on GL can offer to its management process. The first and probably most important opportunity is the reliable automation of repetitive tasks, with great accuracy, and without fatigue[15]. AI can improve current services by adding intelligence, semantic understanding, and powerful analytics to existing GL management processes. A great amount of the created and easily available data can be used to further improve learning algorithms, increasing AI accuracy, and it can create new knowledge and extract new value from existing GL resources. This coupling of big data and AI can bring a new type of AI often referred to as 'data intelligence'.[16]

[13] A two-year study from McKinsey Global Institute suggests that by 2030, intelligent agents and robots could eliminate as much as 30 percent of the world's human labor. (McKinsey, 2017). Available from: **https://goo.gl/RHr53a**

[14] By one popular estimate, 65% of children entering primary school today will ultimately end up working in completely new job types that don't yet exist. (World Economic Forum, 2016). Available from: **https://goo.gl/JKwsbn**

[15] SAS Insights: Artificial Intelligence: What it is and why it matters. Available from: **https://goo.gl/zgpJHz**

[16] The UNESCO Courier, 2018-3. Available from: **https://goo.gl/jeRxkk**

Conclusion

In the last few decades, developments in information technology have had an immense impact on the way we manage information in general, and on the way we create, disseminate and use grey literature. This paper examined the potential impact of AI on grey literature management and elaborated on its main facets, such as value, volume, variety, velocity, and veracity. Based on the growing volume of data, information, and knowledge generated and further increased by the use of AI, we can conclude that GL will not disappear in the future, that its volume will probably experience exponential growth, and that the number of GL types, its velocity, and value will increase.

The impact of AI on GL management will be especially felt in the way GL is processed, kept sustainable, and used in the long run. GL-type processing needs to be included in the f document creation process, rather than post-processing; and the GL management system should be in place all the time, eliminating the need for any ad-hoc interventions. Environmental and technical; economic and financial; as well as social or organizational constraints need to be taken into consideration if long-term GL sustainability is to be provided. Usability of GL depends on the existence of adequate IT tools, the availability of qualified human resources, the protection of intellectual property (IP) and the protection of personal privacy.

Artificial intelligence will impact every aspect of our work environment therfore, in order to secure the future and maintain the value of grey literature, intensive training, wide cooperation, and rigorous preparation need to be organized by all stakeholders and key-players.

Studies show that only a very small percent of businesses extract full value from the information they hold[17]. Use of artificial intelligence in GL management might enable organizations to focus on what matters most – business results, efficiency gains, quality of products and services.

References

BANSAL, Manju, 2014. Big Data: Creating the Power to Move Heaven and Earth. *MIT Technology Review* [online]. [Accessed 19 October 2018]. Available from: **https://goo.gl/Uv1Aw9**

CISCO, 2017. *Cisco Visual Networking Index: Forecast and Methodology, 2016–2021* [online]. Cisco [Accessed 19 October 2018]. Available from: **https://goo.gl/zn9SyV**

DICKSON, Ben, 2017. What is Narrow, General and Super Artificial Intelligence? In: *TechTalks* [online]. 2017-05-12 [Accessed 19 October 2018]. Available from: **https://goo.gl/YN3JXa**

[17] Information management and strategy – an executive guide. Available from: **https://goo.gl/Tu5GNi**

ERTEL, Wolfgang, 2011. *Introduction to artificial intelligence*. Translated by Nathanael BLACK, illustrated by Florian MAST. London: Springer. Undergraduate topics in computer science. ISBN 978-0-85729-299-5.

FARACE, Dominic John and Joachim SCHÖPFEL, 2010. *Grey literature in library and information studies*. New York: De Gruyter Saur. ISBN 978-3-598-11793-0.

KNIGHT, Will, 2017. Forget Killer Robots-Bias Is the Real AI Danger. *MIT Technology Review* [online]. [Accessed 19 October 2018]. Available from: **https://www.technologyreview.com/s/608986/forget-killer-robotsbias-is-the-real-ai-danger/**

LEVI, Heather Pemberton, 2016. Three Degrees of Information Value. In: *Gartner* [online]. 2016-07-07 [Accessed 19 October 2018]. Available from: **https://goo.gl/KYLScQ**

MARR, Bernard, 2016. What Is The Difference Between Artificial Intelligence And Machine Learning? In: *Forbes* [online]. Forbes media, 2016-12-06 [Accessed 19 October 2018]. Available from: **https://goo.gl/iJfNtw**

MCKINSEY GLOBAL INSTITUTE, 2017. *Jobs lost, jobs gained: What the future of work will mean for jobs, skills, and wages* [online]. McKinsey & Company [Accessed 19 October 2018]. Available from: **https://goo.gl/RHr53a**

MINSKY, Marvin, 1969, p. 20. Quoted in: BLAY, Whitby. *Reflections on Artificial Intelligence*. Exeter: Intellect Books, 1996. 157 p.

PETTEY, Christy, 2017. Treating Information as an Asset. In: *Gartner* [online]. Gartner, Inc., [Accessed 19 October 2018]. Available from: **https://goo.gl/kAdpyA**

Information Asset, 2017. In: *Queensland Government Chief Information Office Glossary* [online]. QGCIO, 2017-11-30 [Accessed 19 October 2018]. Available from: **https://goo.gl/mPpySC**

REINSEL, David, John GANTZ and John RYDNING, 2017. *Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big* [online]. International Data Corporation (IDC) [Accessed 19 October 2018]. IDC White Paper. Available from: **https://goo.gl/qdc3fP**

RICH, Elaine, 2010. *Artificial Intelligence*. New Delhi: Tata McGraw Hill, 2010. ISBN 978-0070678163.

SAVIĆ, Dobrica, 2017a. Impact of Disruptive Technologies on Grey Literature Management. *ITlib* [online]. Bratislava: Centrum vedecko-technickych informacii, **2017**(4), 42 - 45 [Accessed 19 October 2018]. ISSN 1336-0779. Available from: **http://itlib.cvtisr.sk/buxus/docs/42_impact%20of%20disruptive.pdf**

SAVIĆ, Dobrica, 2017b. Rethinking the Role of Grey Literature in the Fourth Industrial Revolution. In: *10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017 [Accessed 19 October 2018]. ISSN 2336-

5021. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. Also published by TGJ (The Grey Journal) Special Winter Issue, Volume 14, 2018.

SCHÖPFEL, Joachim and Hélène PROST, 2017. Altmetrics and Grey Literature: Perspectives and Challenges. Altmetrics and Grey Literature: Perspectives and Challenges. In: *18th International Conference on Grey Literature* [online]. New York [Accessed 19 October 2018]. Available from: **https://hal.univ-lille3.fr/hal-01405443/document**

SCHÖPFEL, Joachim, 2011. Towards a Prague Definition of Grey Literature. In *Twelfth International Conference on Grey Literature: Transparency in Grey Literature* [online]. Prague, TextRelease [Accessed 19 October 2018]. pp.11-26. Available also from: **https://goo.gl/Jr2Fg1**

WORLD ECONOMIC FORUM, 2016. *The Future of Jobs: Employment, Skills and Workforce Strategy for the Fourth Industrial Revolution* [online]. World Economic Forum [Accessed 19 October 2018]. Available from: **https://goo.gl/JKwsbn**

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: **https://nusl.techlib.cz/en/conference/conference-proceedings**

THE REFLECTION OF LITERARY ACTIVITIES IN DIGITAL SPACE

Pavla Hartmanová

hartmanova@ucl.cas.cz

Institute of Czech Literature, Czech Academy od Sciences

Paulina Czwordon-Lis

paulina.czwordon-lis@ibl.waw.pl

The Institute of Literary Research of the Polish Academy of Sciences, Poland

Abstract

The Czech Literary Bibliography comprises a set of bibliographical records which reflect cultural journalism and specialist texts on Czech literature. The aim of the contribution is introduction to a new project of the Institute of Czech Literature: The Czech Literary Internet. The project has extended our sources to excerpt platforms, web pages and electronic magazines whose content is not easily searchable through classic search engines. It turns out that this resource illustrates the professional debate on literary events and development and, in particular, brings new information on culture in regions and popular literature.

Keywords

Czech literary bibliography, Czech literary internet, databases, Webarchiv, archiving, search engine, Czech literary life, Polish literary bibliography, literary blogs

Introduction

Česká literární bibliografie (Czech Literary Bibliography – CLB) is a specialized analytical bibliography acquired at the Institute of Czech Literature of the Czech Academy of Sciences. Timewise, its database covers the period from the final third of the 18th century to the present and serves the needs of basic research of literature and literary life in the Czech lands. The bibliography holds almost 600,000 articles processed in standardized MARC21 format and around 1.6 million excerpts in the form of a digitised card index for the years 1770 to 1945.

The Polish Literary Bibliography (PBL), created by Pracownia Bibliografii Bieżącej (Department of Current Bibliography) operates at the Poznan department of the Institute of Literary Research of the Polish Academy of Sciences (Instytut Badań Literackich Polskiej Akademii Nauk). It has, for 70 years (since 1948), collected data from the sphere of Polish and foreign literature and literary theory and about Polish theatre and film. Processed data from 1944 to 1988 are made available through printed yearbooks (with an estimated total of 1.8 million records), data from 1989 to 2003 in the form of an electronic database which currently holds around 700,000 records. A card index which covers the 19th century and the 1st half of the 20th century is stored at the Warsaw branch of IBL and holds around 830,000 cards.

Both databases are, thanks to their broad scope, frequently used by the expert and lay public from the ranks of Bohemists, Polonists and researchers from related humanistic disciplines.

When processing the bibliography of contemporary production, both institutes come across an increasing number of e-sources. The "changing social and cultural status"[1] of electronic sources dealing with literature led both institutes to take the decision that the current literary bibliography cannot ignore such sources in their excerption. CLB therefore launched the "Czech Literary Internet"[2] project in 2017, the aim of which is to map out bibliographically literary life on the Czech Internet from the 1990s to the present day. PBL focuses on the same issue, among others, as part of the IBL.eu project, which also involves research into the issue of processing literary blogs. It is clear that the specific environment of the Internet holds texts similar to printed sources and materials that present those that process national literary science bibliographies with a methodological challenge.

In the paper that follows we endeavour to summarise the conceptual starting point and the initial experiences of both sister institutes in the bibliographic processing of Internet material. In addition to general information about both projects, we concentrate mainly on the general methodological questions associated with processing electronic sources for the needs of literary science bibliographic databases. We will therefore consider in more detail the issues connected to the collection of excerpted material and its archiving and will outline the fundamental problems associated with processing Polish literary blogs.

[1] KAŹMIERCZAK, Marek. Użytkownik, nadawca i odbiorca w Web 2.0. Uwagi o różnych sposobach odnoszenia się do literatury w serwisie Twitter. Teksty Drugie, 2012, 6, 217. ISSN 0867-0633.

[2] The study was established as part of implementation of the project *Czech Literary Bibliography – Czech Literary Internet: data, analyses, research*, CZ.02.1.01/0.0/0.0/16_013/0001743, which is cofinanced by the European Union through European Structural and Investment Funds within the Operational Programme Research, Development and Education.

The Czech Literary Internet project

The Czech Literary Internet project will be handled at CLB from 2017 to 2021 and is supported by EU operational programmes. The aim of the project is to carry out comprehensive research of the Czech literary Internet. In addition to actual analytical processing of e-articles from Internet servers and electronic magazines, it also focuses on the development of software tools for the analyses of excerpted data and scientific research into this material. We are able to summarise the first, constituent results following the first year of work on the project, in particular the specifics of the bibliographic processing of e-sources.

During the first year of Internet excerption, the excerpt base was extended by 44 e-sources, which are processed retrospectively from the beginning of their existence to the present according to the CLB excerption criteria, i.e. we capture both primary texts (fiction) and secondary texts (reflections) in relation to Czech literature and literary culture. It was possible to increase the database to include 10,000 new records from the Internet environment by 31st August 2018. It became clear from preliminary analyses that the emerging database contains reference to articles which the user is often unable to find by merely using an Internet search engine and that they are ordered far to the back in lists of results of background research on Internet search engines due to their low page rank value. This value is primarily taken from the number of references from other pages to the relevant page. Periodicals published in PDF format are entirely invisible to search engines due to the impossibility of indexing their content.

The specifics of records of articles from the Internet

A separate base, given the internal label of INT (internet.ucl.cas.cz), was created for work with records of articles from the Internet as part of improving user services and more comfortable work with data, at the same time records of articles from the Internet are available from the main CLB database (biblio.ucl.cas.cz). Records from websites and electronic magazines can be recognised primarily by the supplement [online] stated after the name of the excerpted source (in MARC21 subfield 773t), for example Ikaros [online], and also by the systematically allocated URL links. Each record of an e-article contains a link to its online version, and if possible to the archival version in Webarchiv.

All excerpted titles, printed and electronic, are also kept on record in the base of excerpted sources (excas.ucl.cas.cz). Here the user finds information about whether the website is processed in full, or merely in part (part, section, etc.). Here we also have on record information about the numbers of records for one excerpted year, and thus volume, to allow the user to have a better idea of the extent to which the server in question deals with literature.

Criteria of selection of e-sources

The criteria for selecting appropriate servers for inclusion in the excerption base in many ways concur with the criteria set out for printed periodicals. We also took into consideration, in particular, the thematic focus of servers, the quality of content and a specific criterion for Internet sources, i.e. the archival possibilities of the selected servers. We consider the perspective of findability of texts through a classic search engine to be secondary. We have not yet included blogs in the exception selection for a number of reasons. There are few blogs written by established Czech literary critics and those that do often contain texts already having been published in a printed periodical, and consequently captured in the article database. Blogs written by authors that publish only on the Internet, on the other hand, predominantly feature reviews of books written by foreign authors, which does not meet the criterion for inclusion in the CLB article database.

Archiving

The transience of electronic content relates on the one hand with the variability of links (impersistent links, transfers of articles to new sections) and, on the other, the lifetime of an actual e-source. The risk that a server and its entire content will disappear without replacement or back-up archiving is certainly not negligible. For these reasons archival possibilities have become an important criterion for the inclusion of individual servers among the excerpted platforms. Given the unforeseeable development of the Internet environment, we consider it necessary for each e-article to contain a link to its original location within the network and to the full text, backed-up at a trustworthy repository.

At first, we considered establishing our own repository in order to ensure the long-term archiving of online content. However, such a solution would be demanding on IT support and would slow down the work of the excerptors themselves. We would also have to deal with the issue of copyright to be able to make texts available to a third party, which would place considerably higher demands on the administrative assurance of the project.

For these reasons we eventually decided to use Webarchiv, the digital library of Czech electronic online sources, which is managed by Národní knihovna ČR (National Library of the Czech Republic) and which has been systematically involved in the archiving of the Czech web for a long time. Individual websites are archived by regular harvests and made available on a "wayback machine" platform, which makes it possible to view the concerned page in various historical versions. Webarchiv makes digital content available in accordance with copyright law and the contracts which it signs with the operators of individual websites, or based on the Creative Commons licence under which the pages are displayed. Thanks to Webarchiv, we always add a link to the oldest archived version of the document to the record so that the user has access to the text in its original form.

Of the 10,000 records we have in the base, we do not have an archival version connected to 11% of records. This result is, however, distorted by servers that do not yet have a contract in place with Webarchiv. Without them we arrive at only 6%. This portion is mainly made up of the latest records which have not yet been processed by Webarchiv.

Types of processed servers

We broadened the existing excerption base to include electronic magazines in PDF format, which mainly take the form of paginated, regularly segmented documents, and to include integration sources (web servers) that do not have regular pagination. Their segmentation is, in contrast to electronic magazines, far more variable and content updates are irregular. The bibliographic data of e-articles from integration sources for these reasons usually only contain the date of publication and the ISSN of the relevant server.

We therefore newly included the following in excerption:

- Websites that supplement/broaden the content of printed excerpted magazines;
- Literary websites;
- Journalism websites;
- University magazines;
- Library magazines;
- Titles already having been excerpted that have moved from printed to electronic version.

Subject-matter

The thematic selection of websites reflects the criteria set for printed periodicals in the base. Even in the digital environment we are interested primarily in a reflection of literature and Czech culture from the perspective of other humanistic sciences. We therefore mainly monitor websites that focus on literature and culture, as well as specialised production published online. We are currently beginning to process large news servers without printed version (for example, Neviditelný pes, Aktuálně.cz, etc.) and in a further stage we plan to work with the online versions of national newspapers or public media websites (Czech Radio, Czech Television).

Quality of content

The web environment can be termed extremely liberal: to exaggerate a little, anyone can publish anything there, which is reflected in the varying quality of individual texts. This is influenced by the fact that there is often no editorial team and that articles are written by literary enthusiasts with varying degrees of experience in writing texts and with the genre that they are trying to achieve. We decided to confront this volatility of content using the criteria of the quality and information repleteness of texts. Whereas in printed periodicals, with the exception of the publishers' annotations and evidently commercial communications, we try to reflect any mention of a literary event, with materials from literary servers each excerptor must evaluate the information and content value of individual texts him/herself and define his/her own criteria. It is necessary to determine whether a particular text has any meaningful value at all and is worthwhile capturing or whether, on the contrary, it is merely a commercial communication or a text completely taken from another source.

Benefit

We consider capturing a reflection of popular literature[3], which appears only marginally in printed periodicals, to be a clear benefit of excerption of electronic sources for the CLB base.

[3] By this we primarily mean the shift of reflection on the genres of popular literature to the Internet environment. There are specialised websites here that deal with genres such as crime novels, horror, sci-fi or fantasy. More information about the genre system of popular literature can be found, for example, in the book by MOCNÁ, Dagmar a kolektiv. *Encyklopedie literárních žánrů*. Praha: Paseka, 2004. p. 503.

The move of publication platforms from printed periodicals to the web is perhaps most striking for popular literature. This step has allowed CLB to expand to include new names of creators of Czech popular literature and the names of critics that have concentrated on such creation for some time now. We are preparing authoritative records of newly-captured authors for a base of national authorities (AUT).

More information about happenings in regions is also making its way into the base - about readings by authors and literary competitions, which national periodicals understandably neglect. We see another positive in supplementing the publication activity of authors that already appear in our base. Before launching the project, we only had their publication activity in printed sources on record. Excerption of the literary Internet therefore made it possible, for example, to monitor whether an author publishes simultaneously on the Internet or whether he/she has opted to publish his/her work within the digital environment alone, for whatever reason. It is also possible to monitor whether an author acts differently on the Internet and in a printed periodical: for example, whether he/she focuses on different literary and journalistic genres, chooses different language, notices other issues, etc.

Problems associated with the processing of e-sources: archiving

As previously indicated, an important perspective in selecting electronic platforms for the Czech Literary Internet project was the inclusion of a webpage in the regular selection collections of Webarchiv[4]. It proved, however, that the inclusion of a server in regular gathering does not guarantee the one-hundred per cent existence of an archive link for an excerpted e-article, i.e. that Webarchiv does not always automatically gather all material from the concerned website.

Most of the problems involved in obtaining an archival version until now have arisen in conjunction with some historic change on the concerned server, such as moving an e-article to a different section, a section ceasing to exist or some change to the software settings of the relevant website which disables collection robots from saving their content in Webarchiv under the current format of URL link. We deal with this situation by manually searching the archival version of the whole website because there is a high chance that the relevant article was collected in the past. Thanks to a contract with CZ.NIC, the National Library archives the full Czech Internet at least once a year. If we are unsuccessful in our search, we contact Webarchiv and agree on the inclusion of specific URL links in selective collection, which is carried out every month.

For unknown technical reasons, then, articles from, for example, the interesting literary website "Opičí revue", are not available in Webarchiv, that site having contained many stimulating literary reviews and functioning from 2010 to 2017. Webarchiv did gather it until 2017, until it closed, but an unknown error on the website means that only articles to the year 2013 are visible with the Webarchiv environment. We consequently only excerpted articles through

[4] Webarchiv distinguishes three types of harvest: blanket, selective and thematic. "Selective collection only covers selected sources, but in contrast to blanket collection the emphasis is placed on capturing sources and changes to them to the full extent." In: KVASNICA, Jaroslav, RUDIŠINOVÁ, Barbora, HAŠKOVCOVÁ, Marie, HOLOUBKOVÁ, Monika and Markéta HRDLIČKOVÁ. *Strategie budování sbírky Webarchivu: aktualizované znění* [online]. Verze 2.0. Praha: Národní knihovna, 2017 [Accessed 12 October 2018]. Available from: **https://webarchiv.cz/static/www/download/collection-policy-2017.pdf**

the archival version of the server and in doing so enhanced our database with 207 articles that are no longer findable and displayable using an Internet search engine.

We assume that is not an isolated case and it is likely that non-functioning links to other e-articles will appear in our database in a few years. We will most likely deal with this problem in the future with an automatic script that will detect and mark such non-functioning links.

The Polish Literary Internet project[5]

One of the objectives of the IBL.eu project is to capture the diversity of the Polish literary blogosphere. Implementation was initiated at an important historical milestone for this publication channel: two blog platforms were closed at the beginning of 2018 - blog.pl and bloog.pl, arousing debate over the end of the "golden days of the Polish literary blogosphere"[6]. The truth is that a whole host of writers' blogs have not been updated for a number of years, were officially closed or have disappeared from the web and the literary discussion ongoing there has moved to social networks. These provide the required "immediate exchange of ideas, general access and simplicity of use"[7]. This trend, however, does not affect review and readers' blogs, which continue to maintain their status. Opinions do appear that favour blogger-reviewers over reviewers from printed and electronic magazines. At the very least, as far as the quantity of literature which they are able to read and appraise is concerned (special attention is devoted to "parental" blogs, the number of which has risen in recent years in line with the growth of literature for children).

Literary blogs

In contrast to the Czech environment, literary blogs play a more important role on the Polish Internet. As projects that are generally backed by a single person, they show a higher degree of instability than "regular" literary websites or e-magazines, which usually have an editorial team of several members to call on. It is now clear that part of the Polish literary blogosphere has probably been lost for good, as is the case for some of the content of Czech literary servers.

The project of processing these is in its infancy and thus stands at the stage of initial research, bringing with it the need to ask fundamental methodological questions. If literary blogs become the starting point for a future project of archiving (in its entirety?) the Polish literary blogosphere, it will be required to specify its boundaries. At this stage of the project, however, a basic typology of blogs is sufficient.

We distinguish the following types of blogs for its needs:

- the blogs of writers that can be termed "author sites". These are used to publicise writing activities, new publications or readings by authors (Przemysław Dakowicz);

[5] Thank you to Anie Gnot for the translation of the Polish text into Czech.

[6] WIŚNIEWSKI, Michał R. Ludzie, którzy piszą w internecie. Dwutygodnik [online]. 01/2018 [Accessed 26 August 2018]. Available from: http://www.dwutygodnik.com/artykul/7600-ludzie-ktorzy-pisza-w-internecie.html

[7] KAŹMIERCZAK, Marek.: op. cit., p. 274. Also WIŚNIEWSKI, Michał R: op. cit.

- blogs which contain literary formations[8]: novels to be continued, aphorisms, poems, etc., including blogs which are later published in book format; in this case publication raises their literary character to a certain extent (the Zorkownia blog written by Agnieszka Kaluga); (the notepad character of blogs as such requires special attention);

- blogs by writers written in the format that can be classified as the (literary) genre of "electronic diary"[9] (Jerzy Sosnowski);

- review blogs (Bernadetta Darska) and readers' blogs (Niezatapialna armada), which vary in terms of the level of professionalism and the authorial strategies.

Evaluation of blogs

Other criteria for the selection of blogs apply to blogs established by writers and critics (or those looking for recognition)[10] and others again to amateur blogs. In the first case, a significant role in the selection process is played by the writer him/herself, because every unreproducible, unprinted piece of material is part of his/her work or illustrates the content of his/her literary texts. In the case of novice or amateur writers, the risk again rises of loss of literary creation from the web, and with it data about the author. It is easy to set up a blog and for this reason the Internet abounds with such literary endeavours. It is therefore entirely obvious that the criterion of literary quality comes into play here. We have until now, in creating PBL, not assessed the quality of literary works because we respected the decisions of the editors of individual magazines in terms of putting these into print and we therefore excerpted all texts which met the thematic criteria for the inclusion of an article in the PBL database. In terms of the bibliographic processing of blogs, this criterion will have to be modified because the quality of published texts is extremely varied.

The availability and archiving of blogs

In all cases - updated, not updated and archival blog - we would like to make a copy of such data units[11] available and to back them up. However, we do not have access to any professional data storage site equivalent to Webarchiv in the Czech Republic. The database of internal sources operated within the bounds of the SYNAT/PASSIM project in place at the Polish National Library is not an archive of electronic sources and merely contains basic metadata and links. Only the incomplete, randomly archived version of blogs can be found at the Internet Archive server (archive.org), and within the scope of analogue initiatives. Full access to blogs which are no longer active is desirable both on the grounds of their literary value (for example, the blog, no longer existing, written by writer Inga Iwasiów) and for the possibility of the reconstruction of the original version of the Polish literary Internet (for example, the following blogs: kumple.blog.pl, mydziecisieci.blog.pl). The implementation

[8] MARYL, Maciej. Życie literackie w sieci: pisarze, instytucje i odbiorcy wobec przemian technologicznych. Warszawa: Instytut Badań Literackich PAN, 2015. ISBN 978-83-61750-61-1.

[9] MARYL, Maciej, op. cit., p. 263.

[10] MARYL, Maciej, op. cit., p. 139.

[11] We see the difference between a non-updated and an archival blog to be that the content of a non-updated blog is available on the Internet, but has not changed for two years or longer. The category of archival blogs includes such pages which are no longer publicly available, no longer exist or whose content is only available through an external Internet archive (for example, from the pages of archive.org).

of a project involving the archiving of the literary Internet would require the creation of a repository for such data or the use of existing online tools (only the second solution comes under consideration at the current stage of the project).

The content of blogs having been published in book form (the Zorkownia blog written by Agnieszka Kaluga), which predominantly contains the reprints of poems, columns, articles (the blog written by poet Wojciech Wencel) or which are located on the servers of publishing houses, magazines and other similar institutions (for example, the blog written by literary critic Justyna Sobolewska at the Polityka.pl journalism site) is obviously available.

Conclusion

In spite of the fact that both of the projects described above approach the issue of the literary Internet from different positions (mass retrospective excerption versus targeted narrower probe), similar methodological obstacles can certainly be seen. These are found in the selection of appropriate e-sources, their technical processing, the quality and transience of electronic content and their archiving. *The Czech Literary Internet builds on existing methods of processing printed periodicals and on the criteria for their selection which are set within the bounds of CLB, and can therefore base itself on material already having been processed.* For these reasons the Czech project included in excerption professional and specialised e-magazines and websites which focus on literature in which the prevailing portion of reflection on Czech literature is concentrated.

The bibliographic processing of materials from the literary Internet therefore presents both national literary science bibliographic projects with new methodological challenges. The bibliographer is no longer faced with the task of simply describing the concerned document - he/she must now, as stated above, become a more active evaluator of the quality of the analysed text, mainly as a result of the absence of editorial work on certain servers, which sometimes leads to texts being published which do not satisfy the conditions for inclusion in the database of articles (commercial communications, not fitting in with the genre, poor language).

The excerptor must in this way deal with the issue of archiving the concerned document: without the existence of an archived version of the document, the work of the bibliographer might be wasted at any time in the future. In this regard, the inclusion of Internet materials in a portfolio of processed documents is a more demanding task that it might appear at first glance.

References

KAŹMIERCZAK, Marek. Użytkownik, nadawca i odbiorca w Web 2.0. Uwagi o różnych sposobach odnoszenia się do literatury w serwisie Twitter. *Teksty Drugie*, 2012, **6**, 217. ISSN 0867-0633.

KVASNICA, Jaroslav, RUDIŠINOVÁ, Barbora, HAŠKOVCOVÁ, Marie, HOLOUBKOVÁ, Monika and Markéta HRDLIČKOVÁ. *Strategie budování sbírky Webarchivu: aktualizované znění* [online]. Verze 2.0. Praha: Národní knihovna, 2017 [Accessed 12 October 2018]. Available from: **https://webarchiv.cz/static/www/download/collection-policy-2017.pdf**

MARYL, Maciej. *Życie literackie w sieci: pisarze, instytucje i odbiorcy wobec przemian technologicznych*. Warszawa: Instytut Badań Literackich PAN, 2015. ISBN 978-83-61750-61-1.

MOCNÁ, Dagmar and Josef PETERKA. *Encyklopedie literárních žánrů*. Praha: Paseka, 2004. ISBN 80-7185-669-X.

WIŚNIEWSKI, Michał R. Ludzie, którzy piszą w internecie. *Dwutygodnik* [online]. 01/2018 [Accessed 26 August 2018]. Available from: **http://www.dwutygodnik.com/artykul/7600-ludzie-ktorzy-pisza-w-internecie.html**

Databases

Česká literární bibliografie [online]. ÚČL AV ČR, 2011 [Accessed 17 October 2018]. Available from: **http://biblio.ucl.cas.cz**

Databáze excerpovaných časopisů [online]. ÚČL AV ČR, 2011 [Accessed 17 October 2018]. Available from: **http://excas.ucl.cas.cz**

Český literární internet [online]. ÚČL AV ČR, 2011 [Accessed 17 October 2018]. Available from: **http://internet.ucl.cas.cz**

Polska Bibliografia Literacka [online]. IBL PAN, 2018 [Accessed 17 October 2018]. Available from: **http://pbl.ibl.poznan.pl/dostep/**

Internet Archive [online]. The Internet Archive, 2018 [Accessed 17 October 2018]. Available from: **https://archive.org/**

Polish blogs

Brukowiec literacki kumple [online]. 2002-2015 [Accessed 17 October 2018]. Available from: **https://web.archive.org/web/20080715000000*/http://kumple.blog.pl/**

DAKOWICZ, Przemysław. *Dakowicz* [online]. 2010- [Accessed 17 October 2018]. Available from: **http://dakowicz.blogspot.com/**

DARSKA, Bernadetta. *Nowości książkowe - Blog Bernadetty Darskiej* [online]. 2015- [Accessed 17 October 2018]. Available from: **http://bernadettadarska.blogspot.com/**

IWASIÓW, Inga. *Świat książki* [online]. 2010-2016 [Accessed 17 October 2018]. Available from: **https://web.archive.org/web/20100801000000*/http://ingaiwasiow.pl/**

KALUGA, Agnieszka. *Zorkownia* [online]. 2010- [Accessed 17 October 2018]. Available from: **http://www.zorkownia.pl/**

Mydziecisieci [online]. 2001-2016 [Accessed 17 October 2018]. Available from: **https://web.archive.org/web/20070915000000*/http://mydziecisieci.blog.pl/**

Niezatapialna Armada Kolonasa Waazona [online]. 2011- [Accessed 17 October 2018]. Available from: **https://niezatapialna-armada.blogspot.com/**

SOBOLEWSKA, Justyna. *Oczytany | Blog Justyny Sobolewskiej* [online]. 2009- [Accessed 17 October 2018]. Available from: **https://sobolewska.blog.polityka.pl/**

SOSNOWSKI, Jerzy. *Jerzy Sosnowski* [online]. 2006- [Accessed 17 October 2018]. Available from: **http://jerzysosnowski.pl/**

WENCEL, Wojciech. *Wojciech Wencel* [online]. 2009- [Accessed 17 October 2018]. Available from: **http://wojciechwencel.blogspot.com/**

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: https://nusl.techlib.cz/en/conference/conference-proceedings

ARCHIVING SOCIAL RESEARCH DATA FROM THE VIEWPOINT OF CZECH SOCIAL SCIENCE DATA ARCHIVE

Martin Vávra

martin.vavra@soc.cas.cz

Institute of Sociology, Czech Academy of Sciences

Abstract

Czech Social Science Data Archive is now an established research infrastructure within the Czech Republic and also it is a part of the European research infrastructure through CESSDA organization. The presentation will address the opportunities and constraints associated with data archiving and sharing in the social sciences. Emphasis will be placed on used standards (for metadata, keywords) and tools (on-line database solution) and on how these standards and tools help to develop a pan-European system of data services in the social sciences.

Keywords

Data archiving, data archives, open access to data, standards of data archiving

Introduction

To begin with, one thing that (I hope) is quite obvious. Scientific information includes not only texts, but research data produced as the result of research processes too. Open access to scientific information therefore covers, or at least should cover, access to research data which complies with the various recommendations articulated by, for example, OECD[1] or the European Union[2]. Moreover, it can be said that we are gradually shifting from the recommendation to make research data available for reuse to the obligation to make them available, as shown, for example, by the rules of the EU Horizon 2020 programme, where it applies in principle that open access should be ensured to the data produced within that programme, naturally with the caveat that there are legitimate reasons for being able to disengage from this obligation, such as personal data protection, or commercial obligations tied to data[3]. Development in the Czech Republic would appear to be heading in the same direction, as shown by Národní strategie otevřeného přístupu k vědeckým informacím (National Strategy for Open Access to Scientific Information) for the years 2017-2020[4].

Why is open access to data important?

There are many reasons to support open access to data to as great an extent as possible, so let us summarise the most important of these (according to Corti et al. 2014 and Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020). Open access to data:

1. makes it possible to build on previous research results and in doing so support the cumulative nature and quality of knowledge;
2. simplify the control of scientific procedures (by replicating analyses);
3. encourages collaboration and reduces the likelihood of duplication of research on the same subject-matter, this supporting the effectiveness of expenditure on science;
4. makes it easier to involve a wider range of actors in the sphere of science (see the current trend of "citizen science").

The need to store data sets and make them available is accompanied by the need to build infrastructure that makes it possible to preserve such data and ensure access to them. This is primarily a matter of constructing data repositories for storing, preserving and making accessible data files and the metadata relating to this. It is important to mention here that there are different types of data repositories and that we are interested only in the ones that are to make data accessible for further analyses as their fundamental objective (Český sociálněvědní datový archiv [Czech Social Science Data Archive] being one of them), meaning that we will leave aside repositories that are only used to preserve data for the purposes of use by a specific institution or by a specific research team. By infrastructure we do not only mean

[1] In its *Principles and Guidelines for Access to Research Data from Public Funding* (OECD 2007), OECD states that access to scientific data acquired from public sources should be simple and user-friendly (preference for online access) and without unnecessary delays (OECD 2007:150).

[2] See, for example, the updated recommendations of EC from April 2018 "On access to and preservation of scientific information", which also places emphasis on data. Available from: **https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018H0790**

[3] More detailed information can be found in Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Available from: **https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf**

[4] **https://www.vyzkum.cz/FrontClanek.aspx?idsekce=851495&ad=1&attid=851502**

the technical side of the matter, i.e. hardware and software, but human resources, the principles of functioning and the standards and procedures used in everyday practice.

Czech Social Science Data Archive and its engagement in CESSDA

The Czech Social Science Data Archive[5] (ČSDA), which is part of the Institute of Sociology of the Czech Academy of Sciences, is such an infrastructure for social sciences in the Czech Republic. ČSDA was established in 1998 under the name of Sociologický datový archiv (Sociological Data Archive), with the name of the archive changing in 2011 to "Social Science" in order to express the broader focus of the acquisition policy. The core of the archive is its data collection, which as it stands (October 2018) numbers more than 800 data files, the overwhelming majority of which still come from sociology institutes, although we are now beginning to acquire data from other areas (historiography, psychology). Most of the data collection is available online[6], the only condition for downloading data being registration, which is possible using a web form[7]. ČSDA uses the Nesstar application[8] to make data available, the application allowing users to search metadata, analyse data online or download them to their computers for further analyses.

A detailed description of the procedure involved in archiving a data file, from the pre-acceptance stage, when negotiations are ongoing with the depositor on the provision of data for archiving, to the state of providing the data file to archive users, is provided in the ČSDA Preservation Policy[9]. Among the important moments of the whole process is the agreement on data deposition in the archive entered into between the depositor and the Institute of Sociology of the Czech Academy of Sciences[10]. The deposited data files and their formats[11] are specified in the agreement, as are any other conditions for handling the data which the depositor may impose. In light of the fact that the archiving procedures in place in ČSDA essentially refer to the OAIS system, we preserve data in original format (in OAIS terminology SIP - Submission Information Package) to maintain the authenticity of the original data, and in "archival" format (AIP - Archival Information Package), which counts on the migration of data formats such that they are readable and usable over the long term. In addition to the acquisition, preservation and disclosure of data, ČSDA workers collaborate in surveys conducted at the Institute of Sociology in which they ensure the management of research data.

The fundamental infrastructure at a trans-European level is the Consortium of European Social Science Data Archives (CESSDA)[12]. A distributed European infrastructure (European Research Infrastructure Consortium – ERIC) was established on the foundations

[5] **http://archiv.soc.cas.cz/**

[6] At: **http://nesstar.soc.cas.cz/webview/**

[7] Such registration means that it is not "open access" in the strictest sense. However, the fact that there is no restriction on who can register means that it is still relatively low-threshold access. Registration serves the archive as a way of binding a specific, identifiable user to adhere to the rules of access to data and also helps us in terms of reporting our activity as an infrastructure - thanks to registration we have a better overview of the numbers and characteristics of active users.

[8] **http://www.nesstar.com/**

[9] **http://archiv.soc.cas.cz/archivacni-rad**, English version at: **http://archiv.soc.cas.cz/en/preservation-policy**

[10] A specimen, which may be modified subject to negotiation with the specific depositor, is available at: **http://archiv.soc.cas.cz/sites/default/files/dohoda_o_depozici_dat.doc**

[11] In its preservation principles, ČSDA describes preferred and acceptable formats for the transfer of data to the archive. In the same way "archival" formats in which data are preserved for the purpose of their long-term storage are specified therein.

[12] **https://www.cessda.eu/**

of the original, informal CESSDA association in 2013. The Czech Republic became a member of this consortium and ČSDA, which had been a member of CESSDA, as an informal grouping, beforehand, became the national node for CESSDA. Eighteen European social science data archives were members of CESSDA in October 2018[13]. The institutions involved differ from each other to a fair extent. Some, such as the UK Data Service, are large organisations that make social science data available and data produced by public administration, for example. Others are "merely" departments at research organisations or at universities that employ only a few people (for example, ČSDA currently employs a total of around six full-time workers). At present, CESSDA (and through it the members' archives, including that of ČSDA) is involved in a number of projects at a European level. The objective of CESSDA for the foreseeable future is to establish a "one-stop shop", a web catalogue that will make it possible, among other, to search the data collections of all members' archives. This endeavour is also understandable with regard to the general attempt to create a genuinely trans-European integrated infrastructure for data access - the most important effort of this type is the European Open Science Cloud (EOSC), an initiative of the European Commission which aims to create a trustworthy and open space for the preservation and sharing of scientific data at a European level by the year 2020. CESSDA has signed up to the principles of open access to data and the term FAIR data is currently used in CESSDA documentation - findability, accessibility, interoperability and reusability[14]. This actually involves the updating (for more details see Wilkinson et al. 2016) and specification of the principles of open access to data that were mentioned in the introduction.

Standards of archiving

To make it possible to fulfil the principles of access to data outlined above, it is necessary to create an infrastructure in the form of "hardware" and to establish and push through standards of work and methods of evaluating their upholding in practice.

As far as the general format of the functioning of an archive is concerned, a number of the archives associated under CESSDA refer in their preservation policies and other documents to the model of the Open Archival Information System (OAIS) as a conceptual framework. As previously mentioned, the ČSDA Preservation Policy is based on OAIS and makes reference to its fundamental principles[15]. OAIS makes it possible to structure the work of archives according to basic functions, processes and positions that are responsible for the execution of functions and processes.

Entirely fundamental to the usability of data in archives are quality and comprehensible metadata, which make it possible for researchers to understand downloaded data sets and assess them from a methodological perspective. The DDI schema of description of data (the abbreviation is taken from the initiative that stands behind for this metadata standard - Data Documentation Initiative[16]) serves the archives of CESSDA, including ČSDA, for this purpose. DDI is a schema (or rather schemata - it has several variations) for data description. The standardisation of metadata is, inter alia, a prerequisite for the creation of the integrated CESSDA portal for data searching mentioned - if metadata were not to have a uniform,

[13] More detailed information about members can be found at **https://www.cessda.eu/Consortium/Membership**

[14] **https://www.go-fair.org/fair-principles/**

[15] **http://archiv.soc.cas.cz/sites/default/files/csda_archivacni_rad.pdf**

[16] **https://www.ddialliance.org/**

machine-processable format, such efforts would be unthinkable given the volume of data sets in the CESSDA members' archives.

Thanks to the CESSDA data portal, the European Language Social Science Thesaurus[17] should be able to fully fulfil its purpose. This is a "thesaurus of key words" in social sciences, now available in 13 different languages, including Czech. As a result of this, ELSST is hierarchically arranged from top terms to specific expressions and its language versions are reciprocally transferable, meaning that it will be possible, for example, to enter a key word in the search engine of CESSDA data catalogue in Czech and the result should be all data files described using this key work in all languages of ELSST.

The final standard that I shall mention here is CoreTrustSeal[18] (originally Data Seal of Approval), which is a system of certifying digital archives. Awarding a CoreTrustSeal should mean that an archive has in place such processes and standards that ensure that the stored data will remain securely stored even over the long term. An archive that wishes to obtain a CoreTrustSeal must, for example, have clear rules in place for secure storage and back-up of data or for the updating and migration of data formats. ČSDA successfully obtained a CoreTrustSeal in 2017[19].

Challenges

There are still a number of problem areas and challenges in the archiving of social science data. It is important that we raise the willingness of researchers to share data - without this, meaning without a culture of data sharing, archiving is troublesome (especially in situations in which the principle of open access to data is not in any way enforceable).

Data should also be prepared for archiving from the very beginning of its life cycle - this means that researchers should have compiled data management plans that are ideally part of research projects, and reference would be made to them in contracts with providers of public money for research. Procedure would then follow these plans. CESSDA is also aware of the fact that archives should be more active at this stage too, i.e. in familiarising researchers with how data management should look, and for this reason compiled the Expert Tour on Data Management[20], which should allow data producers to better plan work with data and their future archiving.

It can be said that "big data" is another major challenge for archives. There has been discussion ongoing in sociology, at least since the article published by Savage and Burrows (2007), on the need to use big data in analyses. The problem is that when we look into data archives, we find practically no big data there (to be blunt, there are none at all in ČSDA). There are various reasons for this, a typical one being that in most cases there is a combination of problems with copyrights, personal data protection and technical problems - meaning how to tailor big data to existing technology and standards at archives.

[17] **https://elsst.ukdataservice.ac.uk/**

[18] **https://www.coretrustseal.org/**

[19] A document providing information on source documents and the results of certification is available from: **https://www.coretrustseal.org/wp-content/uploads/2018/01/Czech-Social-Science-Data-Archive.pdf**

[20] **https://www.cessda.eu/Research-Infrastructure/Training/Expert-Tour-Guide-on-Data-Management**

Conclusion

The infrastructure for archiving and sharing research data in social sciences already exists: however, it obviously has its limits for potential onward development (this stands at the Czech and the EU level). CESSDA is an entirely fundamental organisation at the European level for access to social data. ČSDA is such an organisation within the Czech research environment, and is also a member of CESSDA. Thanks to the development of internet technologies and of archiving standards, this infrastructure is now relatively simple for the user to use. What is important is that the idea of a single place (in the form of a website) at which the user is able to look for data in all CESSDA members' archives is no longer merely an idea, but is at an advanced stage of development.

In order to direct future development, however, it will be necessary to determine what form of open access to scientific data we would like, how we will support it and how we will motivate scientists and their institutions to collaborate on it, such determination obviously only being possible in terms of the science policy of the state, but if possible in line with debate among the parties involved.

References

CORTI, Louise, Veerle van den EYNDEN, Libby BISHOP and Matthew WOOLLARD, 2014. *Managing and sharing research data: a guide to good practice* [online]. Los Angeles: SAGE [Accessed 19 October 2018]. ISBN 978-1-4462-6726-4. **Available from: https://data-archive.ac.uk/media/2894/managingsharing.pdf**

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT, 2007. *OECD principles and guidelines for access to research data from public funding* [online]. Paris, France: OECD [Accessed 19 October 2018]. ISBN 9789264034020. Available from: **https://doi.org/10.1787/9789264034020-en-fr**

SAVAGE, Mike and Roger BURROWS, 2007. The Coming Crisis of Empirical Sociology. *Sociology*. **41**(5), p. 885–899. DOI: 10.1177/0038038507080443.

WILKINSON, Mark D., Michel DUMONTIER, IJsbrand Jan AALBERSBERG, et al., 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* [online]. **3** [Accessed 19 October 2018]. DOI: 10.1038/sdata.2016.18. ISSN 2052-4463. Available from: **http://www.nature.com/articles/sdata201618**

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: https://nusl.techlib.cz/en/conference/conference-proceedings

WHAT ABOUT OTDs? ARE THEY GREY?

Joachim Schöpfel

joachim.schopfel@univ-lille.fr

University of Lille

Snježana Ćirković

snjezana.cirkovic@gmail.com

Austria

Hélène Prost

helene.prost007@gmail.com

CNRS, France

Abstract

The term of grey literature is sometimes applied for older material and special collections, especially in the field of digitization projects of scientific heritage. The following paper will analyse this term of “grey scientific heritage” and, based on empirical and conceptual elements, contribute to a better understanding of grey literature. Special attention will be paid on older theses and dissertations (OTDs), as a main part of scientific heritage especially from universities.

Keywords

Theses, dissertations, heritage collections, scientific heritage, grey literature

Introduction

In systematic literature reviews, meta-analyses and library guidelines, grey literature is often described as unpublished material disseminated outside commercial channels, not peer-reviewed and with limited information referencing. The Encyclopaedia of Library and Information Sciences defines grey literature as material "produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers, i.e., where publishing is not the primary activity of the producing body" (Schöpfel & Farace, 2010). The Prague definition of grey literature raises awareness of its documentary nature, intellectual property rights and quality issues (Schöpfel, 2011). More recently, the 2014 Pisa Declaration on Policy Development for Grey Literature Resources[1] mentions inter alia "research and technical reports, briefings and reviews, evaluations, working papers, conference papers, theses and multimedia content" and recognizes them as an "essential resource in scholarly communication, research and policy making, (and) a key source of evidence, argument, innovation and understanding in many disciplines, (and) an important and valuable part of research and information", which generally means that this is current material with new and unpublished results. However, the term grey literature is also applied to older material and special collections, especially in the field of scientific heritage digitization projects. The following paper will analyse the term "grey scientific heritage" and, based on empirical and conceptual elements, contribute towards a better understanding of grey literature. Special attention will be paid to older theses and dissertations (OTDs) as an essential part of scientific heritage, especially from universities. We recently published a debate on whether and why electronic theses and dissertations (ETDs) should still be considered grey literature in the digital age (Schöpfel & Rasuli, 2018). Here, we will try to assess the grey characteristics of older items from the Gutenberg era. The methodological approach is twofold:

First, relevant papers on scientific heritage from the last two decades will be reviewed, with special attention paid to the definition of grey literature and to the inclusion of theses and dissertations.

Second, the paper will assess more than one hundred digitization projects from a recent French public digitization program to valorise otherwise hidden scientific heritage collections through digital libraries and open repositories. Do these projects contain grey literature? What kind of grey literature? What role is played by OTDs?

This paper will discuss these papers and projects in the light of the usual definitions of grey literature.

Grey literature and heritage - an overview

A search via Google Scholar, OpenGrey and the GreyGuide[2] reveals that few papers make explicit links between grey literature and scientific or cultural heritage collections; a text-mining study of a large corpus of papers on grey literature confirmed that the bigram "cultural heritage"

[1] **http://greyguide.isti.cnr.it/pisa-declaration/**

[2] GreyGuide Repository and Guide to Good Practices and Resources in Grey Literature **http://greyguide.isti.cnr.it/**

is rarely used (Bartolini et al., 2017). Most papers on grey literature deal with recent items, not older material.

Scientific vs cultural heritage

The meaning of scientific heritage is also a matter of discussion. The term itself is “diverse, complex, multi-layered (and) difficult to define” (Lourenco & Wilson, 2013, p.745). It is part of cultural heritage, i.e. a shared collective legacy, a corpus of material signs handed on by the past in every culture and which constitutes a source of identity and cohesion for communities, everything “we want to keep, share with others and pass on to the next generation” (idem)[3]. Heritage of research is one part of these tangible or intangible assets, everything researchers have produced and what is of interest for future research. In other words, “what the scientific community as a whole perceives as representing its identity, worth being passed on to the next generation of scientists and to the general public as well” (idem, p.746). Yet some ambiguities remain, for instance about the legacy character of research libraries, and if scientific heritage must be a scientific contribution or whether it can also include anything produced by scientists, for instance private materials, diaries, letters etc.

The papers on grey literature that deal with heritage collections are from very different domains, including for example Holocaust literature, urban planning, Polish underground literature, Newton’s journal, computer science, Antarctic research, Iceland research publications and the Serbian cultural enlightenment. Some of these papers clearly focus on cultural heritage without any scientific legacy character, such as two studies on historical documents produced by public authorities (de Biagi & Puccinelli, 2017) and on Yizkor books (Jones & Siegel, 2006). Other examples are papers on private collections in the Prado Museum (Docampo, 2010) and on the Australian Baptist heritage collection (Burn, 2006). These items and holdings may be of interest for scientists but are not produced by scientists.

Other studies include both types of heritage, scientific and not, like a paper on unpublished material (i.e. manuscripts, letters, photographs and sketches) by Sir Julius von Haast, a New Zealand scientist in the 19th Century (Nolden, 2017), an analysis of Polish unpublished and prohibited “underground” literature, including translations of scientific and technical items (Nahotko, 2008), or the presentation of the Virgin Islands Heritage Collection with a core collection of digitized material with funeral booklets, historical photographs and newspaper articles, but also research reports and occasional papers produced by local research units (Marsicek & Weiss, 2002).

Types and definitions of grey scientific heritage

Some papers explicitly address grey literature in terms of scientific heritage, especially scientific and technical reports, working papers, proceedings and surveys (cf. Japzon & Anderson, 2005; Stock et al., 2006; Juliusdottir, 2014). Less frequent grey items in these papers include technical drawings (Jackson, 2005; Biagioni & Giannini, 2010), newsletters and workshop/training materials (Ramos-Lun & Vogel, 2006), handwritten notes (Cirkovic, 2016) and materials from conferences that are not readily available (Gheen & Olmsted, 2010). Two studies mention older theses and dissertations (Costello, 2007; Biagioni & Giannini,

[3] Cf. also **https://en.unesco.org/** By the way, the report of the Horizon 2020 expert group on cultural heritage commissioned by the European Commission and published in 2015 (*Getting cultural heritage to work for Europe*) simply doesn’t define the term.

2010). All these items were produced by scientists, are scientific and technical information and written for the scientific community.

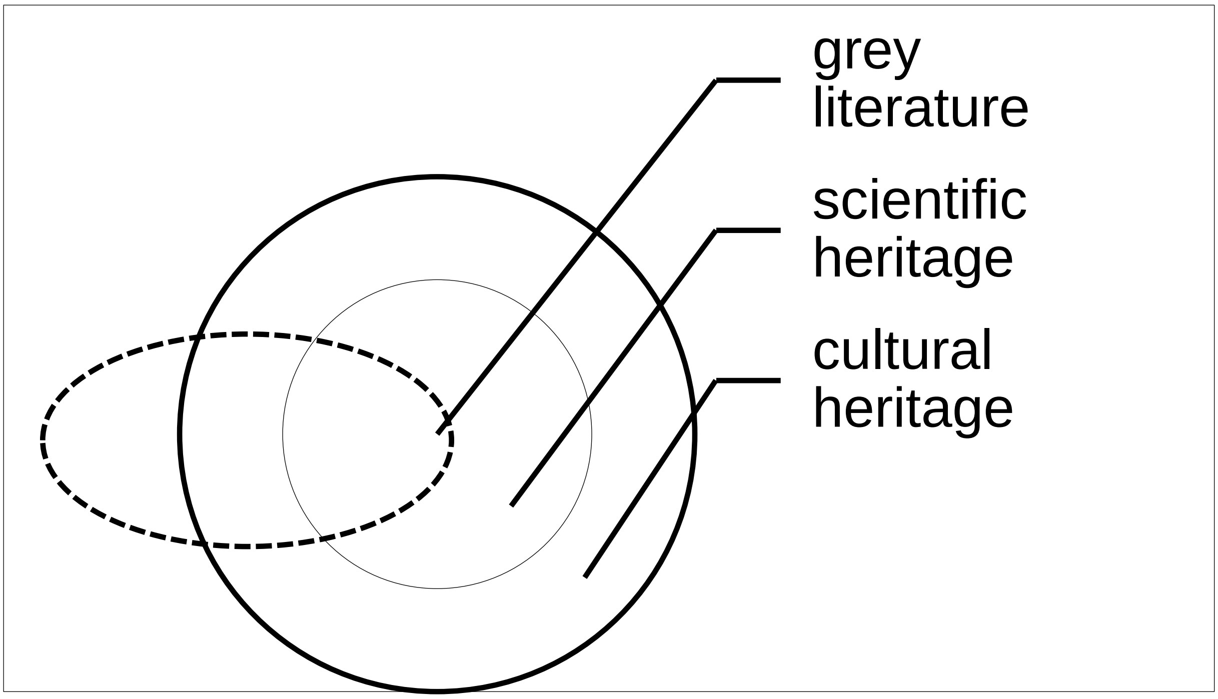


Figure 1: Cultural heritage, scientific heritage, grey heritage, grey literature

How do these papers define grey literature? Some papers just provide a typology of documents considered as grey. Ramos-Lun & Vogel (2006) cite the GreyNet Luxemburg definition of grey literature and distinguish it from “ephemera”, i.e. “materials (physical or electronic) that regardless of appearance, quality or quantity, and that at some point were considered disposable and of little value or no value, through time, had become valuable in such a way that it had broadened their appeal and made them desirable to be collected and preserved by individuals, collectors and information institutions”. In fact, they describe “ephemera” as a special kind of grey literature.

Rucinski (2015) comments that the ephemeral and variable nature of grey publication types, editions and formats makes them hard to describe and define. The risk of being lost if no investment is made is one of the major characteristics attributed to grey heritage items; they are considered endangered due to their small production quantities and various preservation challenges (Jackson, 2005). Another elements commonly used to distinguish grey items from other scientific heritage is the fact that they were not (or not widely) published and may even be confidential, secret or prohibited (“underground”), non-traditional materials, or writings kept away from public review (von Hofe, 2005; Ramos-Lun & Vogel, 2006; Nahotko, 2008). Juliusdottir (2014) states that these items are often “neither bibliographically accessible in catalogues open to the public nor available through traditional market publishing distribution channels”, while Jackson (2005) adds that they are usually difficult to discover or obtain, often difficult to find in libraries, in online databases and on the web, (and) non-accessible.

The distinction between current items and older material (50 years or more) is not always simple. Some studies are “borderline” and do not make it clear if the main purpose is the discovery of scientific heritage or access to recent resources. For instance, one of the first papers on grey literature remains ambiguous about the wartime German research reports collected by the US Army and the UK National Lending Service (now British Library Document Supply Centre) from 1945 on (Chillag, 1994) - was this historical material, or did the US Army gather this material and ship it westward because of its value for cutting edge technology?

Preventing the risk of loss

Why is it important to invest in older grey literature? The papers provide different reasons, e.g. their historical value as "hidden treasures" (Stock et al., 2006; Biagioni & Giannini, 2010), their interest for institutional history and commemoration (Anderson et al., 2007), like the celebration of the 50th anniversary of the first Italian computer (Biagioni & Giannini, 2010). In other words, for these authors the value is the record of progress, not the information itself.

The value of older materials can be increased by making them more readily available. "The greatest challenge remaining for our library is to make our grey literature and ephemera collection available to our users" (Ramos-Lun & Vogel, 2006). A review of older material can include the retrospective enhancement of descriptive and name authority records, thus resulting in "improved documentation of the collections, thereby maximising the discoverability of historical evidence and utilisation of the informational value of a personal and scientific archive" held in library collections (Nolden, 2017). This means a systematic search and collection of resources from a variety of agencies and organizations (Costello, 2007) or digitization and dissemination on the web of collections that were previously fairly inaccessible (Gheen & Olmsted, 2010).

Another related purpose is digital preservation which, alongside dissemination via web servers, prevents grey literature "from moving further toward the black" (Ramos-Lun & Vogel, 2006). Here, various authors are in favour of centralized archives, especially of institutional repositories (Jackson, 2005; Stock et al., 2009; Lynch, 2017) but other solutions may exist.

Standards

Some grey literature, primarily older items, is poorly described in catalogues and databases. Some initiatives insist on the importance of standards to improve the findability and interoperability of heritage collections. Yet, except for a general call for standards, there is no consensus about which kind of standard – generic or not, etc. – should be applied to grey literature.

Jackson (2005), for instance, recommends a current and generic cataloguing standard such as AACR2 for the conversion of print resources to accelerate the inclusion of urban planning resources in online databases and catalogues. Kansa et al. (2010) promote a field-specific approach to primary data, i.e. "a common and highly abstracted framework for expressing archaeological observations, their descriptive properties and their contextual relationships".

Two metadata standards are generally accepted, the Dublin Core and XML. Anderson et al. (2007) developed an extension of the DC, called Goddard Core Metadata Element Set for the metadata of resources produced by the NASA Landsat Legacy Project.

In a quite different environment, namely the Cuban Heritage collection at the University of Miami, Baur et al. (2016) apply the Library of Congress Encoded Archival Description format

(EAD-XML) for the conversion of legacy search aids in typewritten, MS Word, PDF, HTML and even poorly executed EAD-XML formats. Standardized and valid EAD-XML mark-up, according to the authors, is crucial to provide deep access to and interoperability of their archival description and metadata.

Legal aspects

Some papers address legal issues, but this does not seem to be a major concern for this kind of project. Of course, intellectual property rights must be assessed thoroughly before digitizing heritage collections (Blackwell & Blackwell, 2013). Yet the general idea is that older primary source and public domain materials should be freely available, without user agreements and terms of service as a precondition for content access: "(…) our cultural heritage is vulnerable, and risks becoming encompassed within a modern enclosure movement if action is not taken" (Clark & Chawner, 2014). Licensing and laws should be in favour of projects designed to foster the preservation and accessibility of hidden and rare material which may otherwise be "locked up under copyright" (Lynch, 2017).

The case of a national digitization program

As mentioned above, scientific heritage can be described as tangible or intangible assets, everything researchers have produced and what is of interest for future research. The last aspect, in particular, was a central element of a digitization program launched in 2013 and 2014 by the French Ministry of Higher Education and Research to valorise scientific heritage collections[4]. The funding criteria were above all the collections' interest for research communities, together with the technical quality, accessibility (open access), interoperability and the added value of the service environment. The document types were not a specific condition of the program, and the participants – mainly academic libraries and documentation centres – were invited to submit proposals including all kinds of library materials such as journals, books, unpublished papers, posters, photographs, maps and so on. Also, the program was not limited to scientific documents – literary, political or business items could be part of the proposals, as well as press products or personal archives, if their value for research could be established.

123 proposals were submitted and evaluated by an expert group of librarians and scientists. Based on the expert evaluation, the Ministry selected 33 projects with a total grant of €1,296,000 for one or two years. The objective was to valorise otherwise hidden scientific heritage collections through digital libraries and open repositories in an environment of open science and new library services. 38 proposals included grey literature (31%), most of them issued by universities. Nearly all (34) involved consortia with two or more partners; seven

[4] The digitization program was part (segment 5) of the *Bibliothèque Scientifique Numérique* (BSN) or Digital Scientific Library framework (2009-2018). The corresponding author was a member of the steering committee in charge of project selection. The call, criteria and results of the program are available at **http://www.bibliothequescientifiquenumerique.fr/bsn-5-numerisation/**

consortia incorporated international cooperation with institutions from Italy, Canada, Belgium and Germany.

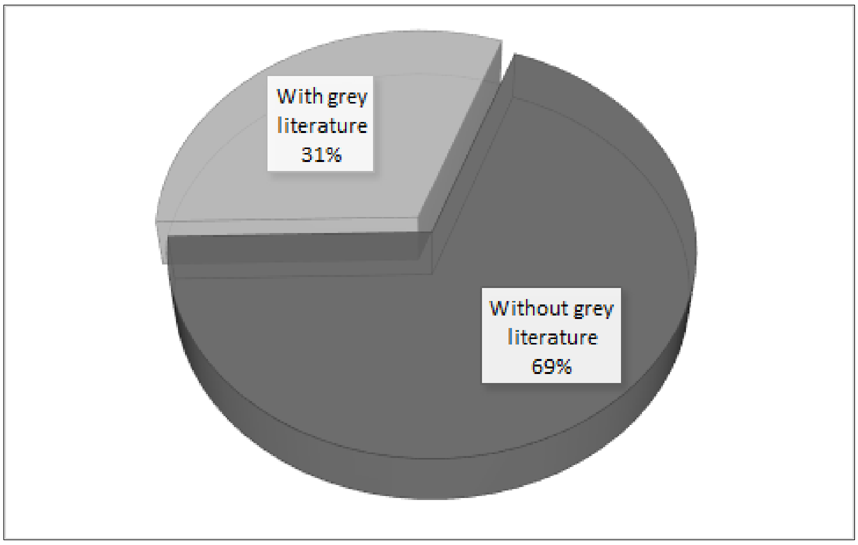


Figure 2: The share of grey literature projects in the French digitization program (N=123)

Two thirds (25) of these 38 proposals are 100% (pure) grey literature projects, with nothing else than grey items like reports, field work, memoirs, minutes or simply “unpublished work”. Six projects included dissertations, such as 444 theses in medicine from the University of Lyon (19th and early 20th Centuries), about 360 theses in law from the University of Toulouse (19th Century) and about 1,000 older theses from Bordeaux. The other 13 projects included other documents than grey literature, often non-textual items like maps, drawings and photos. It is difficult to give reliable figures on the overall size of these projects in terms of volumes or pages. A best estimate of these “hidden treasures” is at least 22,500 volumes with 2.9 million pages. The theses and dissertations represented about 8,218 volumes and 1.15 million pages. Even if only a smaller part (15%) of the whole number of projects with grey literature included theses and dissertations, their real share in the grey heritage collections which academic libraries and research laboratories submitted for funding can be assessed as 35%-40%.

Older theses and dissertations

In the French digitization program, older theses and dissertations (ODTs) formed a significant part of the academic collections – outside of book and journal holdings – for which investment, digitization and valorisation through platforms, digital libraries etc. was sought. Yet what exactly does “older” theses and dissertations mean? In the six proposals, one part of the theses and dissertations was written during the 19th and early 20th Centuries, while others are from the last 50 years up to 2005. In other words, one part of these heritage collections is already in the public domain while other items are still protected by intellectual property rights.

A comparison with the figures from a union catalogue (Trove), a search engine (BASE), a portal (DART-Europe) and two repositories (TEL, NTK) reveals some interesting insights, even if these figures should be considered with caution because these tools are not really designed for older items, except the Australian Trove catalogue (Table 1).

Online service	Up to 1950	Up to 1900
Trove	6.4%	0.5%
BASE	1.5%	0.7%
DART-Europe	1.0%	0.4%
TEL	0.5%	0.0%
NTK GL Repository	0.2%	0.0%

Table 1: Share of older theses and dissertations (in %)

These figures are anything but comprehensive or representative[5], yet illustrate two aspects: on the one hand a non-negligible part of theses in academic catalogues are ODTs, and on the other a small part of them have already been made findable and accessible in the new digital and open science environment via metadata digitization and production. In absolute figures, this small part represents more than 10,000 theses (best estimate).

Table 1 distinguishes between two categories, for two reasons: theses more than 100 years old are probably already in the public domain; theses dating from before 1950 were written by authors who have probably finished their academic (or other professional) careers. Both categories can be considered scientific heritage, and the first category facilitates valorisation through digitization programs.

Age is not the only criteria for scientific heritage. Another, even more important criteria, is the material's quality and value for the research community. Academic theses and dissertations have always been considered as the result of at least three years' original, independent and critical thinking by young scientists, subsequently validated by a commission of senior scholars representing an academic institution and the research community. Depending on the discipline, older theses may still be of interest for today's research. In any case, they represent a unique testimony on the history of science and academic life.

Grey heritage

When does grey literature become scientific heritage? The overview on published studies and papers does not provide a clear indication. Age plays a role, as well as legal status (public domain) and the real usage and interest for research in a given field. In some disciplines like medical and life sciences, scientific documents "expire" and become obsolete more quickly

[5] For instance, we could have added the French academic union catalogue SUDOC or the catalogue of the National Library of the Czech Republic; but would the results have been very different?

than in others, like mathematics or history. In institutions which switched to ETDs many years ago, media might be another criterion, with all native print theses being considered part of the research heritage; yet this criterion makes no sense where print theses are still accepted.

Grey literature is not easy to define, and the fuzzy term of heritage adds to the confusion. Nevertheless, the analysis of recent papers and other evidence from digitization programs and discovery tools enable three aspects to be clarified:

Grey literature is usually defined as being difficult to get because of limited quantities, non-commercial dissemination etc. Yet the problem of scientific heritage is not acquisition because the items are already in library holdings, research collections etc., but are hidden, invisible and not included in bibliographic records. So the first challenge is not the identification of interesting items somewhere outside of the library or the discovery of external information sources and channels – the real challenge is the discovery of valuable material inside the library or research institutions, and the decision to invest into “digging up the treasure” and make it visible which, in the open science environment, means digitization and online publishing.

Grey literature is above all a challenge for acquisition and collection building. Yet the secondary challenge with scientific heritage is not collection building but how to improve the findability and accessibility of formerly hidden items in a new service environment on the internet. Findability means that all items must be described with rich metadata and, if possible, linked to the semantic web, applying the usual standard formats. Accessibility means that these items should be freely available on the web via digital libraries, open repositories or similar platforms, without unnecessary restrictions and without artificial enclosure, and accessible via standard protocols such as OAI-PMH.

Grey literature needs curation and conservation efforts. The third challenge for scientific heritage is similar, even if part of the hidden treasures are already well preserved - perhaps too well. In the era of open science, the preservation of scientific heritage means digital preservation in a secured environment but, at the same time, reusability with new digital tools and services, including content mining and linking to data (cf. France 2018).

Scientific heritage like ODTs, as valuable and useful textual sources of information and insofar as it needs discovery, curation and preservation, can thus be considered grey literature. The intermediary role and importance of academic and research libraries remains intact, and their skills, investment and engagement are needed to discover these hidden treasures and make them findable, accessible and reusable. However, the direction of intervention and perimeter of action are shifted from outside to inside, as these treasures are already in the holdings, waiting for discovery on bookshelves or in microfilm containers like Sleeping Beauties.

Conclusion

Grey literature is not easy to define, and the terms cultural and scientific heritage add even more fuzziness to the concept. Our findings can be summarized in three sentences:

- ETDs are scientific grey literature.
- OTDs are grey scientific heritage.
- Some ODTs are less grey than others.

There is growing awareness of the economic and social benefits of cultural heritage6. Scientific heritage is part of this cultural heritage – it is produced by scientists and is of interest for the research community. For over 20 years, theses and dissertations, “the most useful kinds of invisible scholarship and the most invisible kinds of useful scholarship” (Suber, 2012) are part of ETD programs, and an increasing number of universities and countries such as Australia, New Zealand and France, have gone 100% digital for theses. It is time to take care of the older theses and dissertations present in many academic and research libraries. The problem with ODTs is not dissemination and acquisition but findability, accessibility and reusability. They must be digitized, described and republished in a modern service environment. We recently asked the question of whether ETDs are still grey literature or not, and concluded that “if by 2020 ETDs should be completely integrated in the emerging open science infrastructures, as open as possible (and just as closed as necessary), easily retrievable and accessible, and largely reusable by content mining tools, greyness would no longer be a problem” (Schöpfel & Rasuli, 2018).

The same conclusion applies to ODTs. Obviously, as they are part of scientific output, intellectual work and valuable for scientific and historical research, they are grey literature because of their limited availability, lack of description and risk of loss. Just like all grey literature, ODTs need care and curation by information professionals, especially in the academic and research libraries in the frontline, to increase their findability, accessibility and potential reuse *(Figure 3).*

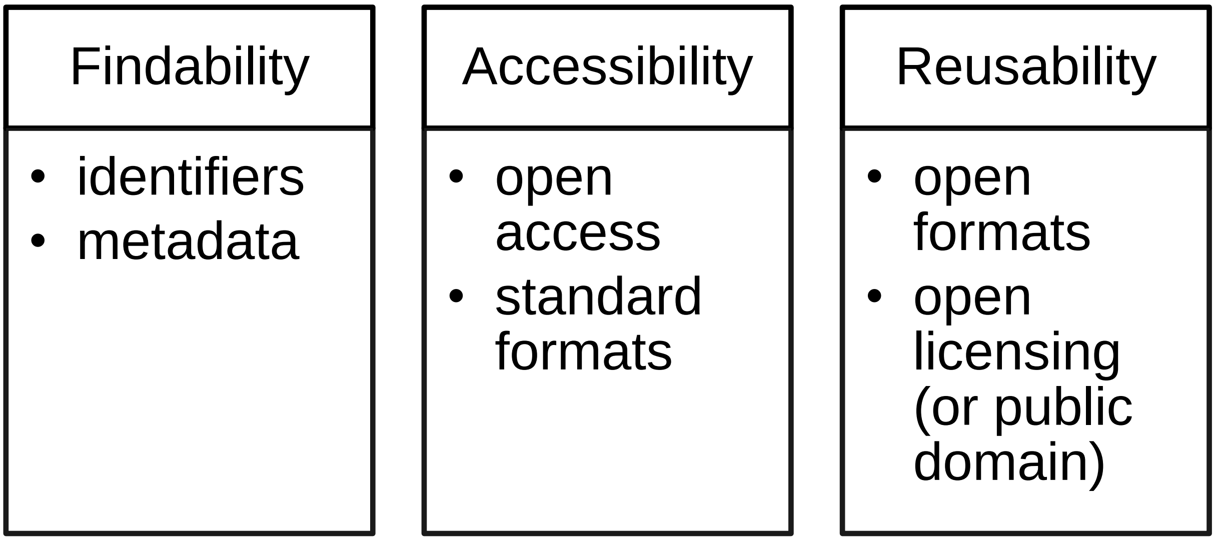


Figure 3: Challenges of grey scientific heritage

The challenge of grey scientific heritage is conservation and dissemination through open access for the scientific community. However, in the digital age and the emerging open scientific age, it must be insisted that readers include not only scientists but also interested

[6] See the 2015 report of the European Commission, cf. footnote 3.

citizens, and that they are not only humans but also machines, with a corresponding impact on decisions on how to publish and how to describe the documents.

As the number of older theses and dissertations is by definition limited, we may be hopeful that one day all or most of these documents will be searchable and available on open repositories or via academic portals as an essential contribution to global scientific heritage.

References

ANDERSON, Nikkia, Gail HODGE and Andrea JAPZON, 2007. Harnessing NASA Goddard's Grey Literature: The Power of a Repository Framework. In: *Eighth International Conference on Grey Literature - Harnessing the Power of Grey: Proceedings.* Amsterdam: TextRelease, 2007. p. 21-24. ISBN 90-77484-08-6.

BARTOLINI, Roberto et al., 2017. A terminological “journey” in the Grey Literature domain. In: *Eighteenth International Conference on Grey Literature - Leveraging Diversity in Grey Literature: Proceedings.* Amsterdam: TextRelease, 2017. p. 117-130. ISBN 978-90-77484-30-2.

BAUR, Natalie, Lyn MACCORKLE and Sevika SINGH, 2016. The Home Stretch: Developing Automated Solutions for Legacy Container List Data at the Cuban Heritage Collection, University of Miami Libraries. *Archival Practice* [online]. Vol. 3. ISSN 2378-4032. Available from: **http://libjournal.uncg.edu/ap/article/view/1158/876**

DE BIAGI, Luisa and Roberto PUCCINELLI, 2017. Grey crossroads' in cultural heritage preservation and resource management. In: *Eighteenth International Conference on Grey Literature - Leveraging Diversity in Grey Literature: Proceedings.* Amsterdam: TextRelease, 2017. p. 162-168. ISBN 978-90-77484-30-2.

BIAGIONI, Stefania and Silvia GIANNINI, 2010. A Hidden Treasure on Computer Science Pre-History in Pisa: The CSCE Collection. *The Grey Journal: An International Journal on Grey Literature.* Vol. 11, no. 1, p. 104-109. ISSN 1574-1796.

BLACKWELL, Amy H. and Christopher W. BLACKWELL, 2013. Hijacking Shared Heritage: Cultural Artifacts and Intellectual Property Rights. *Chicago-Kent Journal of Intellectual Property.* Vol. 137, iss. 1. ISSN 1559-9493.

BURN, Kerrie L., 2006. The Australian Baptist Heritage Collection: implications for the management of geographically distributed special collections. In: *2nd Research Applications in Information and Library Studies Seminar.* Wagga Wagga, NSW: Centre for Information Studies, Charles Sturt University.

CHILLAG, John P., 1994. From Weimar to Maastricht and beyond: half a century with grey literature. In: *First International Conference on Grey Literature: Weinberg Report 2000: Amsterdam (The Netherlands), 13 - 15 December 1993*. Amsterdam: GreyNet.

ĆIRKOVIĆ, Snježana, 2016. Grey Literature Sources in Historical Perspective: Content Analysis of Handwritten Notes. In: *Seventeenth International Conference on Grey Literature -*

A New Wave of Textual and Non - Textual Grey Literature: Proceedings. Amsterdam: TextRelease. p. 131-136. ISBN 978-90-77484-27-2.

CLARK, Alex and Brenda CHAWNER, 2014. Enclosing the public domain: The restriction of public domain books in a digital environment. *First Monday.* Vol. 19, no. 6. DOI: 10.5210/fm.v19i6.4975.

DOCAMPO, Javier, 2010. Creating a heritage collection: the entry of three private libraries into the Prado Museum Library. *Art Libraries Journal*. Vol. 35, no. 2, p. 19-24.

COSTELLO, Gina R., 2007. Louisiana Coastal Wetlands and Louisiana Coastal Grey Literature: Vanishing Treasures. In: *Eighth International Conference on Grey Literature - Harnessing the Power of Grey, Proceedings*. Amsterdam: TextRelease. p. 59-65. ISBN 90-77484-08-6.

FRANCE, Fenella. 2018. Integrated Scientific Research and Data Management for Cultural Heritage. In: *Göttingen - CODATA RDM Symposium 2018: March 19, 2018*. Göttingen: eResearch Alliance.

GHEEN, Tina and Sue OLMSTED, 2010. Digitizing Grey Literature from the Antarctic Bibliography Collection. *The Grey Journal: An International Journal on Grey Literature*. Vol. 11, no. 1, p. 68-72. ISSN 1574-1796.

HOFE, Hal von, 2005. Towards a Genealogy of Grey Literature via Newton's Journals. In: *Sixth International Conference on Grey Literature. Work on Grey in Progress: Proceedings*. Amsterdam: TextRelease. p. 119-122. ISBN 90-77484-04-3.

JACKSON, Rose M., 2005. Grey Literature and Urban Planning: History and Accessibility. In: *Sixth International Conference on Grey Literature: Work on Grey in Progress: Proceedings*. Amsterdam: TextRelease. p. 169-176. ISBN 90-77484-04-3.

JAPZON, Andrea and Nikkia ANDERSON, 2005. Wallops Island Balloon Technology: Can't see the Repository for the Documents. *The Grey Journal: An International Journal on Grey Literature.* Vol. 6, no. 1, p. 24-29. ISSN 1574-1796.

JONES, Faith and Gretta A. SIEGEL, 2006. Yizkor Books as Holocaust Grey Literature. *The Grey Journal: An International Journal on Grey Literature*. Vol. 7, no. 1, p. 152-159. ISSN 1574-1796.

JULIUSDOTTIR, Stefania. 2014. For better or for worse: Knowledge output 1944-2001 and effects of the Legal Deposit Act no. 20/2002 and e-publishing on access to GL in Iceland. In: *Fifteenth International Conference on Grey Literature: The Grey Audit, A Field Assessment in Grey Literature: Proceedings.* Amsterdam: TextRelease. p. 119-128. ISBN 978-90 90-77484-22-7.

KANSA, Eric, KANSA, Sarah, BURTON, Margie and Cindy STANKOWSKI, 2010. Googling the Grey: Open Data, Web Services, and Semantics. *Archaeologies.* Vol. 6, no. 2, p. 301-326, DOI: 10.1007/s11759-010-9146-4.

LOURENCO, Marta C. and Lydia WILSON, 2013. Scientific heritage: Reflections on its nature and new approaches to preservation, study and access. *Studies in History and Philosophy of Science Part A. V*ol. 44, no. 4, p. 744-753.

LYNCH, Clifford, 2017. Updating the Agenda for Academic Libraries and Scholarly Communications. *College & Research Libraries.* Vol. 78, no. 2, p. 126-130. DOI: 10.5860/crl.78.2.126.

MARSICEK, Catherine and Susan WEISS, 2002. Virgin Islands Heritage Collection: A Global Resource Sharing Model. In: *28th Annual Conference of IAMSLIC, Mazatlan, Sinaloa, Mexico, 6-11 October 2002*.

NAHOTKO, Marek, 2008. Some Types of Grey Literature: A Polish Context. In: *Ninth International Conference on Grey Literature - Grey Foundations in Information Landscape: Proceedings*. Amsterdam: TextRelease. p. 89-98. ISBN 978-90-77484-10-4.

NOLDEN, Sascha, 2017. Sir Julius von Haast: exploring an archival documentary heritage collection in the Alexander Turnbull Library. *Journal of the Royal Society of New Zealand.* Vol. 47, no. 1, p. 125-131.

RAMOS-LUM, Marisol and Steve VOGEL, 2006. Entering Grey Waters: Challenges and Solutions of Providing Access to Non-traditional literature in an Aquarium's library. In: *Seventh International Conference on Grey Literature: Open Access to Grey Resources: Proceedings*, Amsterdam: TextRelease. p. 147-151. ISBN 90-77484-06-X.

RUCINSKI, Taryn L, 2015. The Elephant in the Room: Toward a Definition of Grey Legal Literature. *Law Library Journal*. Vol. 107, no. 4, p. 543-559.

SCHÖPFEL, Joachim, 2011. 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: Proceedings*. Amsterdam: TextRelease. p. 11-26. ISBN 978-90-77484-16-6

SCHÖPFEL, Joachim and Dominic FARACE, 2010. Grey literature. In: Bates, Marcia C. and Mary N. Maack (Eds.). *Encyclopedia of Library and Information Sciences*. Third Edition. London: CRC Press. p. 2029-2039.

SCHÖPFEL, Joachim and Behrooz RASULI, 2018. Are electronic theses and dissertations (still) grey literature in the digital age? A FAIR debate. *The Electronic Library*. Vol. 36, no. 2, p. 208-219.

STOCK, Christiane, ROCKLIN, Emmanuelle and Aurélie CORDIER, 2006. LARA - Open access to scientific and technical reports. *Publishing Research Quarterly.* Vol. 22, no. 1. (6 March 2006), p. 42-51. DOI: 10.1007/s12109-006-0007-3.

SUBER, Peter, 2012. *Open access*. Cambridge, Mass: MIT Press. MIT Press essential knowledge series. ISBN 9780262517638.

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: **https://nusl.techlib.cz/en/conference/conference-proceedings**

PROVIDING PUBLIC ACCESS TO THESES AND COPYRIGHT LAW

Matěj, Myška

matej.myska@law.muni.cz

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

Abstract

The paper presents a basic overview of the issue of providing public access to electronic theses and dissertations from copyright's law perspective in the Czech Republic. The first part introduces the purpose of this institute as well as the development of legal regulation. The second part identifies and discusses the emerging problems of the current regulation, especially the ones related to providing online access to theses. The third part presents the relevant existing Czech case law. The last part proposes possible evaluates the applicability of the existing case law to the current regulation and recommends how achieve the desired balance of the two legal obligations (access providing and copyright protection).

Keywords

Copyright law, exceptions and limitations, three-step test, electronic theses and dissertations

The publication of this paper is supported by the Czech Scientific Foundation – project no. GA17-22474S "Adapting Exceptions and Limitations to Copyright, Neighbouring Rights and Sui Generis Database Rights to Digital Network Environment"

Introduction

Providing public access to theses and dissertations seems an ideal way to achieve the broadest dissemination of the achieved results, ensure transparency in education and control the handling of public funds. The development of the relevant infrastructure is also helping to achieve these aims more effortlessly[1]. What still remains a challenge is the "rights thicket" of protection regimes pertinent to theses.

Firstly, a thesis should, by definition, be an original copyrighted work of the author. The thesis is consequently protected by copyright and the author has the moral right to decide whether or not to make her work public (Section 11 of the Copyright Act ("CA"))[2] and the economic right to use her work (Section 12 CA), including the communication of such work to the public (Section 18 CA). In order not to infringe these rights, the provision of public access must be based on an adequate legal title. This might include either an explicit (contractual) license with the author of the thesis, or a statutory license.

Secondly, based on the content of the thesis, the provision of public access might also potentially infringe rights and interests of third parties such as copyright, trade secrets, privacy and personal data. Even in this case, a proper legal title to provide public access to these protected assets is needed (i.e. consent granted by the respective concerned person or legal permission).

This brief paper mainly focuses on the first of the abovementioned issues[3]. The legislative framework for discussion is primarily Czech legislation[4] with the necessary overlaps into international and European regulatory frameworks. This brief paper does not, however, discuss the history, basic aspects and fundamental notions of the issue at hand, as this has already been done elsewhere. Specifically, no attention is paid to the question of defining grey literature (Schöpfel, 2011), copyright protection for grey literature (Schöpfel and Lipinski, 2011), theses as grey literature (Schöpfel and Rasuli, 2018) or theses as copyrighted works and their treatment as grey literature (Polčák and Šavelka, 2009; Polčák, 2010).

Purpose and regulation

Providing public access to theses is an excellent practical example of the balancing of the various interests and rights of the individual and of the public. On the one hand, the author of the copyrighted work enjoys the copyright protection that is also guaranteed by

[1] For an overview of initiatives in this field see e.g. The Networked Digital Library of Theses and Dissertations (**http://www.ndltd.org/about**) and (Suleman, Atkins, Gonçalves, France, Fox, Chachra, Crowder, Young 2001a; 2001b). In Czech see (Mach 2015).

[2] Act No 121/2000 Sb., on copyright, on rights related to copyright, and on amending certain other Acts (Copyright Act), as amended ("CA").

[3] The second issue (i.e. the rights and interests of third parties) is a very complex one and cannot be dealt with in the necessary detail within the scope of this brief paper as this would only lead to misleading oversimplification. At national level, the issue of the protection of personal data in research has been dealt with in (Koščík, Polčák, Myška, Harašta 2017, pp. 59–76); and trade secrets protection in (Horáček, Čada, Hajn 2017, pp. 298–306). Both trade secrets and privacy protection are also elucidated and elaborated upon in the respective commentaries to the Czech Civil Code (e.g. Lavický, Dávid, Dobrovolná, Handlar, Havlan, Horecký, Hurdík, Hrdlička, Koukal, Ronovská, Ruban 2014; Melzer, Tégl 2014)

[4] Primarily Act No 111/1998 Sb., on higher education institutions and on amendments and supplements to some other Acts (Higher Education Institutions Act") or "HEIA") and the CA.

the Charter of Fundamental Rights and Freedoms.[5] On the other hand, there is the political right to information. The public interest in the transparency[6] of functioning of higher education institutions is substantiated mainly because the public higher educations are financed from public funds. Telec (2006) also mentions the legitimate public interest in "improving the state of science, technology and art". This interest is demonstrated in the introductory provision of the HEIA (Section 1 HEIA), which refers to higher education institutions as "the leading centres of education, independent knowledge and creative activity" that "play a key role in the scholarly, scientific, cultural, social and economic development of society". Telec (2006) also argues that public access to theses helps in the discovery of malpractice during the elaboration of the theses, such as plagiarism.

However, until 2006, theses could be used by the respective higher education institution only based on the statutory license for the school work (Section 35 CA). Based on this provision, the thesis could be used for the non-commercial "internal needs" of the higher education institution. For any other uses not covered by further statutory exceptions and limitations, the institution needed a contractual license from the author. Pursuant to Section 60 CA, however, the institution had the right to conclude such a license agreement under the usual terms unless the author did not demonstrate valid reasons for not doing so. In 2006[7] the library exception (Section 37 CA) was amended in such a way that a higher education institution may lend Bachelor's, Master's, Doctoral, advanced Master's ("rigorosum") and habilitation theses on its own premises for the purposes of research and private study, provided the author did not exclude such use.[8]

Since 2006[9] the legislative approach to providing public access to theses has fundamentally changed. In brief, the basic modality changed from "permitted use under certain circumstances" to "an obligation to provide public access". Pursuant to Section 47b(1) HEIA higher education institutions are obliged to provide public access to a defended Bachelor's, Master's, Doctoral, and advanced Master's[10] thesis in a publicly accessible database. The text of the provision does not however expressly stipulate the specific means of achieving this goal and leaves it to the higher education institution to decide this in its internal regulations. It is thus upon the higher education institution itself to adequately balance the abovementioned rights. A university might therefore also set up an online repository (electronic database)

[5] Art. 34 of the Constitutional Act No 2/1993 Sb., on the declaration of the Charter of Fundamental Rights and Freedoms as a part of the constitutional order of the Czech Republic, as amended, declares that: "*The rights to the fruits of one's creative intellectual activity shall be protected by law*".

[6] Transparency and open data are mentioned as the leading reasons for providing public access in the Explanatory Memorandum to the Act that most recently revised the regulation in the Czech Republic (Explanatory Memorandum to Act No 137/2016 Sb. on amending Act No 111/1998 Sb., on higher education institutions and on amendments and supplements to some other Acts (Higher Education Institutions Act) – Parliamentary press 464/0, p. 136–137).

[7] Act No 216/2006 Sb., amending Act 121/2000 Sb. Act, on copyright, on rights related to copyright, and on amending certain other Acts (Copyright Act), as amended, and certain other Acts.

[8] However, with the public obligation introduced in the HEIA discussed below, this provision lost its main purpose and is only applicable to theses defended before 1 January 2006 (Telec, Tůma 2007, p. 391).

[9] Section 47b HEIA was introduced by Act No 552/2005 Sb., on amending Act No 111/1998 Sb., on higher education institutions and on amendments and supplements to some other Acts (Higher Education Institutions Act), as amended, and certain other Acts.

[10] Furthermore, readers' reports must also be provided. The Act amending Act No 563/2004 Sb., on educational workers and on amendments to certain other Acts, as amended, Act No 227/2009 Sb., amending certain laws in connection with the adoption of the Act on Basic Registers, as amended, and Act No 111/1998 Sb., on higher education institutions and on the amendment of certain other Acts (Higher Education Institutions Act), as amended, added the obligation to also provide access to the document describing the course of the defence process.

of theses, as was done e.g. by Charles University[11] and Masaryk University.[12] As regards not-yet-defended theses, the HEIA (Section 47b(2)) foresees the obligation of the institution to make these publicly available at least five days before the defence on-site (i.e. on the premises of the institution) and allows anyone to make copies thereof. The HEIA also contains the presumed consent of the candidate with the provision of public access to the thesis (Section 47b(3) HEIA) effective at the moment the thesis is handed in.

In 2017[13] Section 47b HEIA was significantly amended. Firstly, doctoral theses do not have to be made public if they have been made available to the public in another way (e.g. published as a scientific book). The reason for this exclusion is the actual fulfilment of the purpose of the provision in the case of such publication, i.e. public access to the result. Pursuant to Section 74(5) HEIA habilitation theses are also made public in the same manner if they were not made available to the public in another way (e.g. published as a scientific book). Secondly, the amendment introduced the possibility to delay the provision of public access to a thesis for three years and thus avoid the possible negative consequences of such an action.[14] The reasons for such delay are not mentioned explicitly, but a footnote referencing the illustrative list of Acts regulating potentially infringed rights and interests is provided. Consequently, the protection of copyright (presumably the author's)[15] and the protection of classified information[16] and trade secrets[17] are regarded as credible reasons to delay publication. In order to avoid the misuse of this delay,[18] in such a case one copy of the thesis must immediately be sent to the Ministry of Education, Youth and Sports for archiving, and the reason for the delay must be presented in the same manner in which the thesis would be made available.

Due to its uncertain formulation, relatively broad scope and less-than-ideal legislative technique,[19] the introduction of the obligation to provide public access to theses immediately came under the scrutiny of the Ministry of Culture,[20] practitioners and jurisprudence (Telec, 2006; Křesťanová and Holcová 2008; Polčák, 2010). The next part thus discusses the related problems as regards author rights.

[11] See https://dspace.cuni.cz/?locale-attribute=en

[12] See https://is.muni.cz/thesis/?lang=en

[13] Act No 137/2016 Sb. on amending the Act No 111/1998 Sb., on Higher Education Institutions and on Amendments and Supplements to some other Acts (Higher Education Act).

[14] The previous version of sec. 47(b) HEIA did not entail such mitigating provisions and was prone to, if interpreted rigorously, generate negative consequences such as disclosure of trade secret (Polčák 2010, pp. 74–75).

[15] CA.

[16] Act No 412/2005 Sb., on the Protection of Classified Information and on Security Competence, as amended. Dostál (2018) already convincingly presented that this referral has basically no meaning and practical application in this area, mainly due to the time limitation of the delay with provision of public access to maximum of three years.

[17] Sections 504, 2976, and 2985 of the Act No 89/2012, Civil Code, as amended.

[18] Explanatory memorandum to the Act No 137/2016 Sb. on amending the Act No 111/1998 Sb., on Higher Education Institutions and on Amendments and Supplements to some other Acts (Higher Education Institutions Act) – Parliamentary press 464/0, p. 137.

[19] I.e. that the HEIA does not use the same terms as the CA.

[20] Stanovisko Samostatného oddělení autorského práva Ministerstva kultury k právnímu názoru odboru legislativního a právního Ministerstva školství, mládeže a tělovýchovy k aplikaci § 47b zákona o vysokých školách č. 111/1998 [Opinion of the Independent Department of Copyright of the Ministry of Culture on the Legal Opinion of the Legislative and Legal Department of the Ministry of Education, Youth and Sports on the Application of Section 47b of the Higher Education Institutions Act No 111/1998] Available from: http://ipk.nkp.cz/legislativa/01_LegPod/autorske-pravo/Stanovisko111_98.htm

Emerging issues

In order to frame the discussion, it must firstly be noted that providing public access to theses, might, without the proper legal title, infringe the author's rights provided in the CA. Specifically, both the moral right to decide whether to make the work public (Section 11 CA) as well as the economic rights of reproduction (Section 13 CA), lending (Section 16 CA) and communication to the public through making it available (Section 18 para 2 CA) in the case of provision of online access. As already stated in the introduction, disposition with the thesis is, based upon copyright legislation, primarily in the hands of the author. Without a contractual license, a thesis may only be used based upon a statutory exception or limitation of exclusive rights. As a framing reference, it must also be noted that the question of limitation of exclusive right are regulated by the CA, which must however be compliant with the respective international agreements[21] to which the Czech Republic is a party, and EU legislation.[22] Specifically, the right of reproduction and right of communication to the public are harmonized rights, and Member States can only limit them in cases foreseen in Art. 5 InfoD. Moreover, such exceptions and limitations must also pass the so-called "three-step test" set in Art. 5(5) InfoD, i.e. they "shall only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightsholder." This three-step test was also implemented in Section 29 CA and serves as a general limitation on exceptions and limitations, i.e. a material precondition for their application (Telec and Tůma, 2007, p. 340).

Section 47b HEIA contains more exceptions and limitations of author rights. Firstly, in para. 3, it stipulates the legal fiction of consent with making the work public. Consequently, the moral rights of the author are not infringed by providing public access to the thesis. This conclusion is undisputed both by jurisprudence (Telec and Tůma, 2007, p. 382) as well as the Ministry of Culture.[23] The main argument for this conclusion is that the author is informed from the beginning of her studies about this consequence. The second paragraph of Section 47b HEIA limits the right of reproduction. Again, this issue is not disputed in jurisprudence (Telec and Tůma, 2007, p. 382; Křesťanová and Holcová, 2008, pp. 44–45) or by the Ministry of Culture[24] and is to be regarded only as duplicity of the "free use" (i.e. private copying) exception in the CA regulated through Sections 30 and 30a CA.

On the other hand, the last remaining limiting paragraph of Section 47b HEIA is regarded as highly controversial and problematic. The main reason is its unclear legal nature that has been subject to debate in Czech copyright jurisprudence. Two main opinion streams are identifiable.

The first, represented by Křesťanová and Holcová (2008, p. 45), criticizes its legislative quality and claims that it might be unconstitutional. Namely, this exception is not based on any of the available exceptions and limitations in the InfoD.[25] Even if treated as a sui generis copyright exception, the section would not pass the three-step test as it is not specific enough due to it

[21] In particular, the Berne Convention for the Protection of Literary and Artistic Works of 1886, as amended by the Paris Revision of 1971, the World Intellectual Property Organization World Copyright Treaty of 1996.

[22] 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 ("InfoD").

[23] Stanovisko Samostatného oddělení autorského práva Ministerstva kultury k právnímu názoru odboru legislativního a právního Ministerstva školství, mládeže a tělovýchovy k aplikaci § 47b zákona o vysokých školách č. 111/1998 Sb. Available from: **http://ipk.nkp.cz/legislativa/01_LegPod/autorske-pravo/Stanovisko111_98.htm**

[24] Ibid.

[25] 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.

permitting a university to set its own rules regarding the "database of theses" (Křesťanová and Holcová, 2008, p. 45). This opinion was also expressed by the Ministry of Culture, which classified this exception as non-compliant with European legislation[26]. As a result, these authors recommended the conclusion of licensing agreements with the students submitting their theses (Křesťanová and Holcová, 2008, p. 47).[27]

On the other hand, Telec and Tůma regard this paragraph as a specific "quasi-license", i.e. a sui generis limitation of rights with only a specific beneficiary (i.e. the higher education institution) and that is based only on a specific legal relationship between the student and the respective institution. As a result, the higher education institution is not only obliged, but also entitled, to use the thesis for such purposes – namely on the basis of the legal fiction of a legal license.[28] Consequently, such a limitation is not subject to Art. 5 InfoD. However, as with any limitation of exclusive right, it must be interpreted restrictively and in compliance with the three-step test. Polčák (2010, p. 74) opined that the obligation to provide access to theses "using a publicly accessible database gives implicit permission (a license) in and of itself for the university".

Case law

The provision was extensively examined by Czech courts in the dispute between T. H. and Masaryk University. This public university generally provides online access to full texts of submitted theses and readers' records. The question was, however, discussed under specific factual circumstances. Namely, the claimant submitted two versions of his dissertation thesis (2011, 2012) for defence, and at the same time concluded an exclusive (sic!) licensing agreement with a publishing house (2010) that consequently published the first version of the dissertation as a book in 2011. Furthermore, the internal regulations of Masaryk University contained the option not to provide public access to a respective thesis, due inter alia to the protection of legitimate third-party interests, of which the author did not take advantage.

The Regional Court in Brno[29] applied Section 47b HEIA dismissed the cease-and-desist claim and ruled that the Masaryk University had proceeded secundum et intra legem.

The appellate Higher Court[30] overruled this decision and stated that Masaryk University had infringed the appellant's copyright. The reasoning was, however, rather unclear as the appellate court simply stated that the university had not made the theses accessible only to the inner academic community as foreseen by Section 35(3) CA. Thus, it completely omitted and did not apply Section 47b HEIA. Masaryk University filed a recourse to the Supreme Court which dealt with the main question of whether the legal obligations, i.e. to provide public access

[26] Stanovisko Samostatného oddělení autorského práva Ministerstva kultury k právnímu názoru odboru legislativního a právního Ministerstva školství, mládeže a tělovýchovy k aplikaci § 47b zákona o vysokých školách č. 111/1998. Available from: **http://ipk.nkp.cz/legislativa/01_LegPod/autorske-pravo/Stanovisko111_98.htm**

[27] The same recommendation was reached by the Expert Committee on Providing Electronic Access to Theses of the Association of Libraries of the Czech Universities (**http://www.evskp.cz/dokumenty.php?tsekce=2&sek=&ukol=1)**.

[28] As elucidated aptly by (Křesťanová, Holcová 2008, p. 47). Under this view, the legislator tried to balance the rights and interest of the institution and the author beforehand.

[29] Regional Court in Brno of 29 April 2014, file No 23 C 61/2013-117.

[30] Judgement of the Higher Court in Olomouc of 26 February 2015, file No 7 Co 5/2014-142.

and to protect the rights of the author, had been properly balanced, especially as regards the form and manner of such provision.

The Supreme Court[31] preferred the position of Telec and Tůma (2007, p. 381) that Section 47b HEIA is a "quasi-license" limitation of the author's rights[32] that must be interpreted restrictively and in compliance with the three-step test (Section 29 CA). The Supreme Court also elucidated the relationship between the CA and HIEA, namely that the CA is lex generalis and HIEA lex specialis. The Supreme Court did not however dwell on the international and/or European aspects of this issue. Interestingly, Section 47b HEIA with its reference to the internal regulation of the higher education institution, was found to be sufficient to pass the first step of the three-step test, i.e. that the limitation of copyright is applicable only in "certain special cases"[33]. The provision of online access did not seem in conflict with the second step of the test, as making theses available is a common part of the graduation process that does not conflict with the normal exploitation of the work[34]. Also, Masaryk University (defendant) had set up an internal procedure regarding how to block access to the thesis that might have been used before it was handed in. However, the author did not take advantage of this opportunity. Consequently, the third step was sufficiently passed, as the legitimate interests of the author were not unreasonably prejudiced.

On the second occasion, the Higher Court,[35] bound by the legal opinion of the Supreme Court, actually applied the three-step test in casu. The first step was satisfied, as the rights of the author were limited on a legal basis (specifically Section 47b HEIA) and the internal regulations of Masaryk University also reflected this legal basis. There was also no problem with the second step as "the submission of the final thesis is a one of the prerequisites for completing the studies"[36]. The internal regulations of Masaryk University had foreseen this manner of publication and also reflected Section 47b(3) HEIA in that submission of the thesis also implies consent to provide public access to it. The Higher Court further inferred - from the fact that the thesis is accessible in the university information system - that its use is for study purposes. Such use can be deemed normal. Even the third step had been complied with in the current case. The legitimate interest could not have been unreasonably prejudiced as he had the opportunity to proceed according to the internal regulations and request that the thesis not be made publicly accessible but had not done so.

After this decision, T. H. filed a constitutional complaint directly against this judgment of the Higher Court as it allegedly infringed his right to the protection of the rights 'to the fruits of one's creative intellectual activity'. The complainant also filed for the declaration of Section 47b HEIA as unconstitutional. However, as he "failed to exhaust all the procedures afforded

[31] Judgement of the Supreme Court of 29 October 2015, file No 30 Cdo 2864/2015.

[32] Limitations of copyright might also be included in Acts other than the CA, which was reflected in the most recent amendment to Section 29 CA (Act No 102/2017, amending Act No 121/2000, on copyright, on rights related to copyright and amending certain other Acts (Copyright Act), as amended).

[33] Notwithstanding the opinion of Křesťanová and Holcová (2008, p. 45) that such a limitation of author's rights is too broad and no certain enough.

[34] Including the in casu situation, where the first version of the thesis was published as a book – the author notably had the possibility to block/delay access to it.

[35] Judgment of the Higher Court in Olomouc of 9 February 2016, file No 7 Co 5/2014-186.

[36] Judgment of the Higher Court in Olomouc of 9 February 2016, file No 7 Co 5/2014-186.

him by law for the protection of his rights"[37] the complaint was rejected as inadmissible[38]. As a result, the Constitutional Court did not carry out the assessment so sought by Czech jurisprudence (see e.g. Křesťanová and Holcová, 2008). Thus it is still possible that this provision might be declared unconstitutional, especially as the Supreme Court only evaluated the previous version of the HEIA that did not contain a maximum delay of three years. Furthermore, the provision might also be challenged as to its conformity with the InfoD in the form of a reference for a preliminary ruling.

Evaluation and recommendation

The fundamental conclusion of the decision is that making theses available online does not conflict with the rights of the author, and consequently no license agreement with the author is currently needed[39]. However, the decision of the Supreme Court discussed above dealt with the previous version of Section 47b HEIA that did not include the possibility to delay the provision of access for a maximum of three years. If a similar suit were filed again, the courts would have to re-evaluate the compatibility of the provision and also the specific chosen form and manner of providing of public access with the three-step test. In other words, whether the balance between the legal obligation of the higher education institution and the protected rights of the author will still be achieved. As regards the provision itself, the fundamental problem might potentially lie with the second step of the three-step test, i.e. conflict with the normal of a work. However, it could be argued that the normal exploitation of theses is precisely to make them available to the public for public scrutiny. Also, even though public access might only be delayed for three years, this term should be regarded as sufficient to commercially exploit the theses. Furthermore, it could be also argued[40] that the provision stipulates a "one-time" delay that could, however (under the strictest conditions), also be renewed. At the level of internal regulations, the use of technological protection measures could also achieve the desired balance. Providing public access (even online) with the application of technological measures that would only enable mere access (i.e. not the making of reproductions in the form of prints and digital downloads) would surely be regarded as compliance with the second and third steps. However, non-display uses (Borghi and Karapapa, 2011), e.g. for data mining, should still be allowed.

De lege ferenda, the "publication clause" concerning a dissertation thesis (47b(1) HEIA in fine), i.e. no obligation to provide public access if it has already been provided, e.g. through the publication of a scientific book, should be extended to all the concerned theses. This would again make the provision more acceptable in the context of the three-step test.

In order to address all the identified problems, a more systematic reform of access provision would be needed. The Slovakian Central Repository of Theses[41] might serve as good

[37] Section 75 of the Act Constitutional Court Act 182/1993 Sb. (English translation of the Act cited from: **https://www.usoud.cz/fileadmin/user_upload/ustavni_soud_www/Pravni_uprava/AJ/ZUS_EN_verze_2018.pdf**). The correct procedural step would have been the filing of an appeal to the Supreme Court.

[38] Order of the Constitutional Court of 15 February 2017, file No II.ÚS 1317/16.

[39] These findings directly contravene the conclusions of Křesťanová and Holcová (2008, pp. 46–47) and confirm the conclusions of Polčák (2010, p. 74).

[40] I would like to thank my colleague Michal Koščík for this idea.

[41] Central Register of Final and Qualifying Works [Centrálny register záverečných a kvalifikačných prác]. For general information, see its home page at http://cms.crzp.sk/. The register is primarily regulated in Section the sec. 63(7) to (13) of Act No 131/2002 Z. z., on higher education and on amendments to some other Acts ("HEIA-SVK"). The details are set out in the respective Decree (Decree of the Ministry of Education, Science, Research and Sport of the Slovak Republic No 233/111 Z.z., implementing some

inspiration. The system is state-run, and higher education institutions have an obligation to submit theses into this repository, which also provides an originality check (Section 63(7) HEIA-SVK). Public access is only provided on the basis of a contractual license with the author, which they are, however, obliged to conclude (Section 69(9) HEIA-SVK). The provision of access might be delayed for at most three years (Section 69(10) HEIA-SVK). The manner of provision might also include “protected” access limited using technological protection measures blocking the copying and printing of the theses (Section 69(10) HEIA-SVK). In order to properly balance the rights, published theses are not subject to the access obligation, as in the Czech Republic.

Conclusion

The Czech legislation concerning the providing of public access to theses is prima facie not suitable to provide clear answers and potentially not compliant with the InfoD. The approach to this regulation was also dichotomic. Part of the national jurisprudence as well as the expert bodies recommended that higher education institutions conclude license agreements with their students. On the other hand, another part of the doctrine suggested that this is completely legal, as the obligation to do so also implies the necessary authorisation. Finally, the Supreme Court sanctioned that a solution based on this legislation consisting of making theses available online without restriction was compliant provided there was also the opportunity to prevent the provision of public access to the theses. The last HEIA amendment introduced yet another level of legal uncertainty, as the provision of access might be delayed but only for three years. Currently there is no case law reassessing the compliance of this specific condition.

Higher education institutions are therefore in an unenviable situation. They must create a system to provide public access to theses that would carefully balance all the involved interests and at the same time fulfil legal obligations. Following the decision of the Supreme Court, these details should be adequately stipulated in the internal regulations of the higher education institution. The second step, i.e. the conflict with the normal exploitation of the work, seems to be just as crucial, however still passable even if the delay is limited to at most three years, if the blocking mechanisms are set up.

A more radical approach would be the complete overhaul of the system and introduction of a complex state-backed system for providing public access to theses such as the one in the Slovak Republic.

References

ADAMOVÁ, Zuzana and Branislav HAZUCHA, 2018. *Autorský zákon: komentár.* 1. vydanie. Bratislava: C. H. Beck. Beckova edícia Komentované zákony. ISBN 978-80-89603-58-9.

BORGHI, Maurizio and Stavroula KARAPAPA, 2011. ID 2358912: *Non-Display Uses of Copyright Works: Google Books and Beyond* [online]. SSRN Scholarly Paper. Rochester,

provisions of Act 131/2008 Z. z., on higher education and on the amendment and supplementation of some other Act). This decree contains, inter alia, a licensing agreement template. The legal analysis of its functioning is provided in (Adamová, Hazucha 2018, pp. 721–723).

NY: Social Science Research Network [Accessed 15 November 2018]. Available from: **http://papers.ssrn.com/abstract=2358912**

DOSTÁL, Jakub, 2018. Vysokoškolská kvalifikační práce coby utajovaná informace. *Právní rozhledy* [online]. 2018. Vol. 26, no. 17, p. 601- [Accessed 15 November 2018]. Available from: **http://beck-online.cz**

HORÁČEK, Roman, Karel ČADA and Petr HAJN, 2017. *Práva k průmyslovému vlastnictví.* 3., doplněné a přepracované vydání. Praha: C. H. Beck. ISBN 978-80-7400-655-5.

KOŠČÍK, Michal, Radim POLČÁK, Matěj MYŠKA and Jakub HARAŠTA, 2017. *Výzkumná data a výzkumné databáze: právní rámec zpracování a sdílení vědeckých poznatků.* Praha: Wolters Kluwer. ISBN 978-80-7552-952-7.

KŘESŤANOVÁ, Veronika and Irena HOLCOVÁ, 2008. Ustanovení § 47b zákona o vysokých školách jako autorskoprávní omezení? In: KŘÍŽ, Jan, Tomáš DOBŘICHOVSKÝ, Irena HOLCOVÁ, Veronika KŘESŤANOVÁ, Hana LENGHARTOVÁ, Janine SMITKIEWICZ, Artur-Axel WANDTKE and Petra MALÁ ŽIKOVSKÁ. *Aktuální otázky práva autorského.* Praha: Karolinum, p. 40-47. ISBN 978-80-246-1528-8.

LAVICKÝ, Petr, Radovan DÁVID, Eva DOBROVOLNÁ, Jiří HANDLAR, Petr HAVLAN, Jan HORECKÝ, Jan HURDÍK, Miloslav HRDLIČKA, Pavel KOUKAL, Kateřina RONOVSKÁ and Radek RUBAN, 2014. *Občanský zákoník: komentář. I. Obecná část (§ 1-654).* 1. vyd. Praha: C. H. Beck. Velké komentáře. ISBN 978-80-7400-529-9.

MACH, Jan, 2015. *Správa, vyhledávání a zpřístupňování elektronických vysokoškolských kvalifikačních prací* [online]. PhD thesis. Praha: Univerzita Karlova v Praze, Filozofická fakulta [Accessed 15 November 2018]. Available from: **https://dspace.cuni.cz/handle/20.500.11956/64653**

MELZER, Filip and Petr TÉGL, 2014. *Občanský zákoník: velký komentář. Svazek III: § 419-654 a související společná a přechodná ustanovení.* Vyd. 1. Praha: Leges. ISBN 978-80-7502-003-1.

POLČÁK, Radim and Jaromír ŠAVELKA, 2009. *Digitální zpracování tzv. šedé literatury pro Národní úložiště šedé literatury* [online]. [Accessed 15 November 2018]. Available from: **http://www.nusl.cz/ntk/nusl-111528**

POLČÁK, Radim, 2010. Legal Aspects of Grey Literature. In: *Grey Literature Repositories* [online]. Zlín: VeRBuM, p. 67-89 [Accessed 15 November 2018]. ISBN 978-80-904273-6-5. Available from: **http://invenio.nusl.cz/record/97129/files/idr-285_1.pdf**

SCHÖPFEL, Joachim and Tomas A. LIPINSKI, 2011. *Legal Aspects of Grey Literature. The Grey Journal* [online]. Vol. 8, no. 3, p. 137-153 [Accessed 15 November 2018]. Available from: **https://archivesic.ccsd.cnrs.fr/sic_00905090/document**

SCHÖPFEL, Joachim and Behrooz RASULI, 2018. Are electronic theses and dissertations (still) grey literature in the digital age? A FAIR debate. *The Electronic Library* [online]. Vol. 36, no. 2, p. 208–219 [Accessed 15 November 2018]. DOI: 10.1108/EL-02-2017-0039. Available from: **https://www.emeraldinsight.com/doi/full/10.1108/EL-02-2017-0039**

SCHÖPFEL, Joachim, 2011. Towards a Prague Definition of Grey Literature. *Grey Journal (TGJ).* Vol. 7, no. 1, p. 5–18.

SULEMAN, Hussein, Anthony ATKINS, Marcos A. GONÇALVES, Robert K. FRANCE, Edward A. FOX, Vinod CHACHRA, Murray CROWDER and Jeff YOUNG, 2001a. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 2: Services and Research*. D-Lib Magazine* [online]. Vol. 7, no. 9 [Accessed 15 November 2018]. DOI: 10.1045/september2001-suleman-pt2. Available from: **http://www.dlib.org/dlib/september01/suleman/09suleman-pt2.html**

SULEMAN, Hussein, Anthony ATKINS, Marcos A. GONÇALVES, Robert K. FRANCE, Edward A. FOX, Vinod CHACHRA, Murray CROWDER and Jeff YOUNG, 2001b. Networked Digital Library of Theses and Dissertations: Bridging the Gaps for Global Access - Part 1: Mission and Progress. *D-Lib Magazine* [online]. Vol. 7, no. 9. [Accessed 15 November 2018]. DOI: 10.1045/september2001-suleman-pt1. Available from: **http://www.dlib.org/dlib/september01/suleman/09suleman-pt1.html**

TELEC, Ivo and Pavel TŮMA, 2007. *Autorský zákon: komentář.* Praha: C. H. Beck. ISBN 978-80-7179-608-4.

TELEC, Ivo, 2006. Zveřejňování závěrečných prací vysokými školami. *Právní rádce* [online]. Vol. 14, no. 10, p. 69-70 [Accessed 15 November 2018]. Available from: **https://pravniradce.ihned.cz/c1-19622050-zverejnovani-zaverecnych-praci-vysokymi-skolami**

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: https://nusl.techlib.cz/en/conference/conference-proceedings

THE UNIVERSITY TEACHERS' ATTITUDES TOWARDS GREY LITERATURE: A SURVEY LED AT THE UNIVERSITY CONSTANTINE 2

Azzedine Bouderbane

azzedine.bouderbane@univ-constantine2.dz

University Constantine 2

Nadjia Gamouh

nadjia310@hotmail.com

University Constantine 2

Teboura Benkaid Kesba

dirbu2007@yahoo.fr

University of Oum-el-Bouaghi

Abstract

Scientific research requires the collection of a variety of information gathered from diverse resources. Grey literature is one of the information resources that may enrich researchers' scientific works, but, in some situations, some users may not have a positive attitude towards grey literature for various reasons. We led a qualitative study at the University Constantine 2 where we arranged an interview with a representative sample of university teachers. The descriptive approach was adopted. Significant results will be presented to the audience of the conference.

Keywords

Scientific research, university teachers' attitude, grey literature, institutional digital repository, qualitative study, survey, University Constantine 2

Introduction

Researchers endeavour to identify their information needs. After exploiting the information, they collect with tremendous efforts, they seek to publish their intellectual output. Whenever they succeed in this task, they give more visibility to their research. Thus, they benefit scientifically. This process underlines the importance of gathering a variety of information from different resources. Grey literature is among the information resources researchers require to produce quality scientific works. However, in some situations and for various reasons, researchers do not have the opportunity to benefit from these fruitful resources.

Methodology

In this chapter, we state the problem and explain how we arranged the survey.

Stating the problem

We live in an era of information explosion. The output of literature is undergoing extraordinary expansion thanks to the development of the publishing process and technology. University teachers are astonished by the multiplicity of modern and evolving research tools now available. For scientific reasons, university teachers and academic institutions endeavour to produce published and unpublished works. Scientific published literature is visible and available in libraries, bookshops, academic sites, databases and so on, whereas unpublished literature seems hidden and unavailable. University teachers in Constantine had a negative attitude towards grey literature. This factor created a difficult situation for teachers who needed information in general, and grey literature in particular.

The survey

This stated problem encouraged us to run a qualitative survey to identify the basic reasons for such attitudes towards these useful resources. We adopted a descriptive approach that helped us, on the one hand, describe the phenomenon and, on the other hand, analyse the data collected from the interview which we used to gather elements to provide explanations and predicted answers to our questions. The interview process lasted ten (10) days from 15 June to 25 June 2018. It was arranged at the Central University Library. Fifty (50) teachers from University Constantine 2 constituted our sample. Twenty-five (25) university teachers were females, while the other twenty-five (25) were males. All the selected university teachers worked in the social and human sciences. We did not experience any difficulty getting in touch with our respondents because they were colleagues, we often meet at university pedagogical meetings, scientific conferences, and when frequenting the Central University Library. Through our study, we sought to achieve he following objectives: a) to determine how frequently university teachers use this literature; b) to check whether these teachers encountered obstacles when they needed to use these specific collections; c) to evaluate the university teachers' attitudes towards grey literature; and d) to highlight the fundamental role of universities in promoting the use of grey literature.

Scientific research

Scientific research is the means used when seeking the truth concerning objects and phenomena and when identifying the relationships and links that exist between these matters. This enables us to find answers to queries and solutions to problematic situations. Scientific research is an activity carried out at universities and research laboratories and centres. It can assist society to develop and renew itself. It can also assist society in overcoming a variety of obstacles. Scientific research is an efficient instrument that enriches faculty members' knowledge, and develops their scientific experience. Students benefit when their teachers transmit this knowledge.

Nobody can deny that scientific research is fed by information and knowledge. Quality research output requires information collected from a variety of scientific documents. A great many scientific, academic and technical studies and reports are produced yet, often remain unpublished - these are useful and valuable scientific documents that can be added value for the researchers' scientific output.

Information search and the knowledge culture

Information is transmitted in many formats; it can be printed, electronic, digitized and/or audio-visual. "Information resources are understood as the totality of information gained and accumulated during the scientific and practical activities of people for use in production, management and everyday life" (Odintsov, 2012). Information search means a set of actions, methods and procedures that attempt to extract desired information from a set of documents (Dinet, Rouet and Passerault, 1998). Individuals need to be skilled enough to be able to access information rapidly and efficiently. Consequently, they should be familiar and competent with information search techniques. Nowadays, technological development plays key role in storing, processing and retrieving information available in a variety of forms and in multiple modern formats. Information users should, therefore, be competent in electronic information searches. This skill has now become fundamental, so that information users can establish search strategies and use information in a rational manner (Candallot, 2005). Nowadays, searching information in data-bases, on the internet or in a document has become a common activity accessible to all: it is no longer only undertaken by information specialists (Maury, 2011). Information professionals should, however, "Know that even though search engines may be important tools, it is also important to help users develop a critical approach to assessing the information resources that are readily available" (Dinet, Rouet and Passerault, 1998). Competency with information search techniques permits the progressive acquisition of knowledge and culture. Furthermore, it is no longer sufficient to merely know how to use information; it is also important to produce the knowledge that is a prerequisite for any progress and for any success. For this reason, society should valorise knowledge, encourage citizens to keep learning throughout their lives, and oblige its institutions to dispense and diffuse knowledge. M. F. Blanquet (1999) was right when she said: "The future belongs to those who handle knowledge".

Grey literature

Grey literature is a broad concept that is in constant evolution. The best global experts, such as Joachim Schöpfel (2010), have mentioned the difficulty in giving a precise definition to grey literature, adding: "The definition of grey literature is intimately conditioned by the fact that it is an object of collection and acquisition. A document becomes grey not only because it is a work of the mind and not sold by a vendor but insofar as someone – an institution, a library, an information service, and a professional – shows interest in acquiring it." The fourth international conference on grey literature defined it as: "That which is produced at all levels of government, by academics, business and industry, in print and electronic formats, but which is not controlled by commercial publishers (Fourth International Conference on Grey Literature, 1998). Other writers state that grey literature represents "a reservoir of rich information which is however not well structured (Expernova, 2015). Marzi cited IGLWG (Interagency Gray Literature Working Group) which defined GL in 1995 as "Open source material that is usually available through specialized channels and may not enter normal channels or systems of publication, distribution, bibliographical control, or acquisition by booksellers or subscription agents (Marzi, Pardelli and Sassi, 2010).

Data collection and analysis

When asking the respondents to give examples of grey literature documents, 90 % mentioned: manuscripts, theses and technical reports. 10 % of our sample only cited: theses. These examples expressed the teachers' interests and needs in terms of information resources. To highlight the distinction between grey literature and other documents, 50 % of our sample explained that it was a matter of published and unpublished documents, while the other teachers were not able to identify any distinction. This statement proved that for many teachers, their conception of these specific resources was still ambiguous. Regarding the location of grey literature documents, 70 % spoke only of university libraries, 20 % mentioned university libraries and digital university repositories, while 10% spoke about university libraries, digital university repositories, information centres in the industrial sector, and the Internet. Though the respondents were not perfectly informed about the multiple locations of grey literature, 30 % of them mentioned digital university depositories, an instrument which might encourage them to archive their scientific output, and gain a clear idea about the concepts of grey literature and open access. An institutional repository can be viewed as "...a set of services that a university offers to members of its community for the management and dissemination of digital materials created by the institution and its community members" (Institutional repository, 2018). A question related to possible difficulties when attempting to gain access to grey literature, and when using it, showed that 60 % of the respondents faced technical difficulties. This problem was definitely related to the age of the university teachers and the related difficulty to gain competency in today's technological culture. We should also mention that 80 % of our sample consisted of university teachers aged between 50 and 60. Concerning electronic information searches, 65 % of the respondents underlined its complexity. This proved that they had not mastered this discipline. 80 % of our sample spoke about administrative obstacles, explaining that academic institutions rigidly conserved and protected unpublished documents using the permanent slogan of "copyright regulations". Regarding the frequency of use of these documents, 80 % of respondents mentioned that they "rarely" used grey literature in their information searches because of the tremendous obstacles they faced when attempting to use these resources. Just 15 % of them affirmed that they "never" used grey literature documents in their research because of the harsh and complex environment surrounding these specific resources. When asked about assistance provided by their university in the use of these documents, 90 % of respondents attested that the contribution of their university in this matter was very modest. "As proof", they explained, "our university launched a project in 2016 to establish a digital university depository that would be useful and helpful for teachers on the scientific side, but we are still eagerly looking forward to its birth". **The following list shows the digital university repositories available in the country:**

- **Tlemcen University**
  http://dspace.univ-tlemcen.dz/
  Copyright
- **Biskra University**
  http://dspace.univ-biskra.dz:8080/jspui/
  Copyright
- **Ouargla University**
  https://dspace.univ-ouargla.dz/jspui/
  Copyright
- **Chlef University**
  http://dspace.univ-chlef.dz:8080/jspui/
  Copyright
- **Bouira University**
  http://dspace.univ-bouira.dz:8080/jspui/
  Copyright

We asked them another question to see whether they could benefit from other digital university depositories available in the country. 75 % did not know anything about these repositories in Algeria. 25 % of respondents who were informed about these informational tools attested that the online university repositories were under copyright protection, and that no repository was licensed under Creative Commons. The survey showed that male university teachers who were informed about grey literature were more numerous than female university teachers (5 %), which could be explained by women's double responsibility as housewives and as university teachers. They cannot spend enough time at the university to frequent university libraries and communicate with librarians about library collections in general, and grey literature in particular. The respondents added that full-text access was not possible, and that some of these online repositories provided access to bibliographical data, whereas others also provided abstracts of theses and articles, and in addition they affirmed that the Internet connection was weak, making access to these repositories quite impossible. Universities should work hard in this field to develop modern depositories. With the new digital university repositories, “advanced users are also given the opportunity to perform advanced queries” (Caffaro and Kaplun, 2010).

Main survey results, comments and suggestions

The global attitude of Constantine 2 University teachers, (respondents to our interview), was negative for various reasons: a) A lot of them find it difficult to define the concept of grey literature or to identify the various types of these specific documents; b) They confront multiple obstacles (administrative, technical, technological and personal) when attempting to access grey literature documents; c) A lot of them think that the geographical space where grey literature resources are located is very narrow, whereas it is actually wide for those who are well informed about these documents; d) The majority of university teachers are not competent in electronic information searches.

University teachers should know that grey literature can significantly contribute to the promotion of scientific research. Unpublished resources that seem invisible to teachers for various reasons have the potential to make teachers’ scientific output more valuable, fruitful, pertinent and of higher quality. University teachers have just begun to be aware about the importance of the digital university depository: this is positive, yet the number of teachers, who understand the usefulness of this instrument, use it, feed it with their scientific output, and benefit from it should be expanded. “Now-a-days, libraries cannot meet the needs of their users solely through their own collections. To meet the various needs of users and interdisciplinary approaches, libraries need to use information sources (…) in a network environment.” (Chowdappa and Ramasesh, 2010)

The role of academic institutions in improving university teachers’ sensitivity to the importance of grey literature is predominant. Arranging meetings and organizing conferences and, programming workshops… are some of the activities that can improve teachers’ awareness of grey literature, as well as their interest in accessing it to use it. Each institution should establish training programs to enhance teachers’ understanding of grey literature and the importance of archiving their scientific output in their institutional repository when the latter is available. Training is the dynamo for academic progress. Using modern tools that promote intensive communication, will help them develop the open access spirit and increase their willingness to share, transmit and exchange information. They will discover a great amount of grey literature resources in their digital university repository and be able to exploit it. “Grey literature has significantly contributed to the open access movement and, as such, has bolstered the public’s trust in science” (Farace, 2010). This will make teachers’ scientific output richer and more visible. The impact of research at the university will be improved and, consequently, the university will also be more visible, and will demonstrate its performance through international academic rankings.

Conclusion

Progress generates unprecedented upheavals in all fields. The acquisition of information and knowledge by citizens is giving new dimensions to learning and development (Bouderbane, 2013). Information resources are available in a variety of shapes and formats. People who know how to use and exploit these resources will feel strong, useful and productive in this complex society. It would be a pity to conserve information resources that cannot be used and exploited by users. University teachers – people with sensitive roles in educating generations – should possess the faculty of adaptation that they need to easily integrate this changing society. Grey literature is among the resources that university teachers need, above all in their scientific research. Information specialists should participate in sensitizing users to the importance of grey literature. They can train them in electronic information search competency. Grey literature is a treasure that should be opened to users' eyes, especially to those who have the mission of educating generations. These specific resources should not be buried; they are alive, and, furthermore, they can represent added-value to scientific research and output.

References

BLANQUET, M. F., 1999. S'approprier l'information électronique. In: *Bulletin des bibliothèques de France (BBF)* [online]. No. 5, p. 8-16 [Accessed 12 November 2018]. ISSN 1292-8399. Available from: **http://bbf.enssib.fr/consulter/bbf-1999-05-0008-001**

BOUDERBANE, A., 2013. Information Literacy: A Key for Getting in Knowledge Society. In *Revue Madjelet El Maktabet Wa El Maaloumat*. Vol. 4, no. 2, p. 1-6.

CAFFARO, J. and S. KAPLUN, 2010. Invenio: A Modern Digital Library System for Grey Literature. In*: Twelfth International Conference on Grey Literature: Conference Proceedings: 6-7 December 2010.* Amsterdam: TextRelease. p. 91-94.

CANDALLOT, C., 2005. *Ma formation à la recherché documentaire*. Québec: Documentation et Bibliothèques. p. 231-239.

CHOWDAPPA, N. and C. P. RAMASESH, 2010. Grey Literature in Engineering Sciences and Technology and its Use Pattern in the Research Institutions in India: The Case Study of Karnataka State. In*: Twelfth International Conference on Grey Literature: Conference Proceedings: 6-7 December 2010.* Amsterdam: TextRelease. p. 101-111.

DINET, J., J.-F. ROUET and J.-M. PASSERAULT, 1998. Les "nouveaux outils" de recherche documentaire sont-ils compatibles avec les stratégies cognitives des élèves?. In: Rouet, J.-F., de la Passardière, B. *Quatrième colloque "Hypermédias et Apprentis-sages"* [online]. Poitiers, France: Association Enseignement Public & Informatique, Oct 1998. p. 149-162 [Accessed 12 November 2018]. ISSN 2-7342-0625-0. Available from: **http://www.epi.asso.fr/association/dossiers/hyper4.htm**

EXPERNOVA, 2015. Littérature grise: un gisement incontournable d'information. In: *Expernova Blog* [online]. 2015 [Accessed 1 July 2018]. Available from: **https://blog.expernova.com/litterature-grise-un-gisement-incontournable-dinformation/**

FARACE, D., 2010. Peering through the Review Process: Towards Transparency in Grey Literature. In*: Twelfth International Conference on Grey Literature: Conference Proceedings: 6-7 December 2010.* Amsterdam: TextRelease. p. 32-38.

Fourth International Conference on Grey Literature, 1998. *Asian Libraries* [online]. Vol. 7, no. 9 [Accessed 1 July 2018]. DOI: 10.1108/al.1998.17307iab.006. ISSN 1017-6748. Available from: **http://www.emeraldinsight.com/doi/10.1108/al.1998.17307iab.006**

Institutional repository, 2018. In: *Wikipedia: the free encyklopedia* [online]. Wikipedia Foundation, Inc., last modified on 24. 10. 2018 [Accessed 1 August 2018]. Available from: **https://en.wikipedia.org/wiki/Institutional_repository**

MAURY, Yolande, 2011. Boubée, Nicole and André Tricot. Qu'est-ce que rechercher de l'information? 2010. [online report] In: *Spirale: Revue de recherches en éducation*. Vol. 2011, no. 48, p. 191-192 [Accessed 6 November 2018]. ISSN 2118-724X. Available from: **https://www.persee.fr/doc/spira_0994-3722_2011_num_48_1_1787_t11_0191_0000_2**

MARZI, C., G. PARDELLI and M. SASSI, 2010. *A terminology based re-definition of grey literature.* In*: Twelfth International Conference on Grey Literature: Conference Proceedings: 6-7 December 2010.* Amsterdam: TextRelease. p. 27-31.

ODINTSOV, B., 2012. Some quantitative relationships between traditional and information resources. In: *Inf. Resur. Ross.* No. 6, p. 11–14.

SCHÖPFEL, J., 2010. *Towards a Prague Definition of Grey Literature*. In*: Twelfth International Conference on Grey Literature: Conference Proceedings: 6-7 December 2010.* Amsterdam: TextRelease, 2010. p. 11-26. Available also from: **https://archivesic.ccsd.cnrs.fr/sic_00581570/document**

Appendix

Interview grid

1. Could you give examples of documents that are specified as grey literature?
2. What is the main distinction between grey literature and other resources?
3. Are you able to determine the location of grey literature?
4. Do you encounter difficulties when you need to use grey literature? If so, what are the main difficulties?
5. How often do you use grey literature?
6. Does your university help you gain access to grey literature? If so, how?
7. What do you know about your digital university repository?
8. As a university teacher, are you able to use grey literature via the digital university repositories available in your country?
9. How would you generally evaluate your use of grey literature?
10. How can your university promote teachers' use of grey literature?

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: https://nusl.techlib.cz/en/conference/conference-proceedings

SUI GENERIS DATABASE RIGHT AND OFFICIAL WORK EXCEPTION

Míšek, Jakub

jakub.misek@law.muni.cz

Institute of Law and Technology, Masaryk University, Czech Republic

Abstract

Sec. 94 of Copyright Act was amended by “Open Data Act” in the way that official work exception applies mutatis mutandis, to the maker of the database. The new regulation makes it easier to work with public sector databases, including grey publicly funded literature repositories. However, it presents interpretation difficulties. This paper points to them and suggests solutions. The first part analyses the official work exception and its legal construction. The second part briefly deals with sui generis database rights and their exceptions. The third part connects first two topics. It analyses the possibility of application of the official work exception to sui generis database rights and deals with the transitional provision, which contains the amending “Open Data Act”.

Keywords

Sui generis database rights, official work, grey literature

The publication of this paper is supported by Czech Scientific Foundation – project GA17-22474S Adapting Exceptions and Limitations to Copyright, Neighbouring Rights and Sui Generis Database Rights to Digital Network Environment.

Introduction

Institutions interested in the collection, cataloguing and publication of grey literature must properly address intellectual property rights which may relate to the content. Apart from copyright, they must not forget the sui generis database right. If the grey literature database is protected by such right, it must also be licensed during the publication process. Otherwise, any future ambitions to reuse grey literature works in the database might be thwarted even though they are not directly protected by copyright.

This article considers as grey literature documents within the meaning of the Prague definition of grey literature.[1] In the Czech Republic, a vast number of such documents are excluded from copyright protection thanks to the official work exception. Section 2 of this article briefly addresses this exception and its application to grey literature.

Once collected, grey literature works are placed in repositories and made accessible to the public. These repositories are databases within the meaning of European database directive No 96/9/EC.[2] Even though we usually do not think of database content directly as grey literature[3], intellectual property rights which relate to such databases are important because they play a supporting, instrumental role for the database content. Part 3 of this paper presents basic sui generis database right protection concepts in connection with grey literature repository practice.

The so-called “Open Data Act” (No 298/2016 Sb.[4]) introduced a new rule stating that the official work exception applies mutatis mutandis to the maker of the database. This change is important for institutions operating repositories created during the exercise of legal duties. In these cases, such institutions would not need to concern themselves with any licensing of database rights. Therefore, the change ensures easy the accessibility and reusability of provided content. Unfortunately, however the amendment also introduced a transitional provision which makes its interpretation more complicated. The presented article addresses this concern in Section 4 and seeks to provide an interpretation aid for practical use.

Official work

The official work exception is stipulated in Section 3(a) of the Copyright Act[5], and states that official works are excluded from copyright protection.[6] Its purpose is to ensure the free

[1] See Joachim Schöpfel, ‘Towards a Prague Definition of Grey Literature’, in *Grey Tech Approaches to High Tech Issues.* (Presented at the Twelfth International Conference on Grey Literature: Transparency in Grey Literature, Prague, 2011), pp. 11–26. This definition was chosen because of its recency, clarity and exhaustiveness.

[2] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

[3] Schöpfel argues that grey literature is acquired and put into collections. That indicates that the content of a database (of a collection) can be grey literature, but the database of such works itself will most likely not be. See ibidem p. 17. Nevertheless, if present, database rights must be addressed during the publication of grey literature.

[4] This Act is primarily harmonisation legislation enacted in connection with European Regulation No 910/2014 on electronic identification and trust services for electronic transactions in the internal market. The regulation of open data is present only as a relatively independent part of the Act, and it amended Act No 106/1999 on the freedom of information and Act No 121/2000, the Copyright Act.

[5] See Act No 121/2000 Sb., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts.

[6] The full text of the mentioned section states: “*Copyright protection shall not apply to an official work, such as a legal regulation, decision, public charter, publicly accessible register and collection of its documents, and also any official draft of an official work and other preparatory official documentation including the official translation of such work, Chamber of Deputies and Senate*

availability of public and publicly important documents. It is a quite traditional legal institute in Czechia. The exclusion of certain publicly important documents from copyright protection was already present in the Czechoslovak Copyright Act from 1926.[7] The application of this exception is generally broad for two reasons. Firstly, the list of excluded documents it mentions is not exhaustive. Secondly, the definition is open thanks to the statement that it covers "also any other work where there is a public interest in its exclusion from copyright protection".[8] The exception covers most of the documents created during the fulfilment of public administration duties (e.g. Acts, court decisions, opinions, recommendations etc.). However, in situations when a work is created by a third party (e.g. an attorney who writes a legal analysis), this exception generally does not apply.

A vast amount of Czech grey literature - covering all governmental documents, opinions and analyses - falls within this exception and can therefore be freely distributed and used for any purpose. On the other hand, this exception does not apply to other grey literature documents such as master and doctoral theses, and these are thus covered by copyright. Furthermore, the official work exception is applicable to copyrightable databases[9] if the abovementioned condition of public interest is met. A database can be protected by copyright when "by reason of the selection or arrangement of the content, it constitutes the author's own intellectual creation".[10] The copyright applying to such a database does not protect its content,[11] but only the structure and arrangement of that content. Good examples of copyrightable databases that fall within the application of the official work exception are the Register of Persons and the Business Register.

Sui generis database right

The sui generis database right protects the investment required for the creation of the database.[12] The maker of the database is a person or a body which "is involved both in the initial organization of the database and its financing."[13] In the case of grey literature repositories, the maker of the database is the institution which creates and operates the repository. The right is established when there is a qualitatively or quantitatively substantial investment in obtaining, verifying or presenting the content.[14] For example, if the grey literature

publications, a memorial chronicle of a municipality (municipal chronicles), a state symbol and symbol of a municipality, and any other such works where there is public interest in their exclusion from copyright protection."

[7] See Section 6 of Act No 218/1926. Sb. Telec dates this exception back to 1895. See Ivo Telec and Pavel Tůma, *Autorský zákon - Komentář* (Prague: C. H. Beck, 2007), p. 73.

[8] Some authors argue that a work can become an administrative work only once the consent of the right holder is given, because otherwise it would be a form of expropriation. See ibidem.

[9] In the light of Art. 1 of Directive No 96/9/EC, a database means "*a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means*". For an exhaustive explanation see P. Bernt Hugenholtz, 'Directive 96/9/EC', in *Concise European Copyright Law*, ed. by Thomas Dreier and P. Bernt Hugenholtz, Second edition (Alphen aan den Rijn, the Netherlands: Kluwer Law International, 2016), pp. 379–420 (pp. 379–401).

[10] See Article 3 Section 1 of Directive No 96/9/EC.

[11] See e.g. Michal Koščík and others, *Výzkumná data a výzkumné databáze. Právní rámec zpracování a sdílení vědeckých poznatků* (Prague: Wolters Kluwer ČR, 2018), p. 43.

[12] This opinion is confirmed by legal doctrine. See e.g. Hugenholtz, p. 402; Matěj Myška a Jakub Harašta, 'Omezení Autorského Práva a Zvláštních Práv Pořizovatele Databáze v Případě Datové Analýzy', *Časopis pro Právní Vědu a Praxi*, 23.4 (2016), 375–84 (p. 377); Telec a Tůma, p. 732.

[13] Hugenholtz, p. 403.

[14] See Article 7 of Directive 96/9/EC. A detailed explanation of these terms is beyond the scope of this article. For more information see e.g. Hugenholtz, pp. 402–15; and Matěj Myška and Jakub Harašta, 'Less Is More? Protecting Databases in the EU after Ryanair', *Masaryk University Journal of Law and Technology*, 10.2 (2016), 170–99. Available from: https://doi.org/10.5817/MUJLT2016-2-3

repository operator invests in creating a presentation layer which enables direct automatic computer access to the database (application interface, or “API” in short), that might constitute necessary investment.

The maker of the database has an absolute right to prevent anyone else from extracting (copying in any way) and re-using (any further use) the database content. Therefore, it can be said that the sui generis database right indirectly protects the database content. Furthermore, direct access to the database is not necessary for this protection to apply. Extracting data from a mirror copy of a database may constitute an infringement of the sui generis right of the first database.[15] The maker of the database can license and (unlike copyright) also waive his rights.

When a database protected by sui generis right is made public, anyone authorised to access it can extract and re-utilise insubstantial parts of its content. Directive 96/9/EC also provides a few exceptions from database protection, like the use of database content for illustration when teaching or performing scientific research. An insubstantial part of the content from an online repository of qualification theses can be, for example, extracted and reused by anyone.[16] In all other situations, the maker of the database must license the sui generis right to fully provide its content. Otherwise, the full scope of the content[17] cannot be used by third persons.

The sui generis database right and the official work exception

When a database is made by a public sector body in the course of fulfilling its public administration duties, such database should not be protected by the sui generis database right. This is because such database is funded from a public budget, so there is no investment and no risk in it, and thus no need for protection.[18] This idea was acknowledged by the Czech legislator through the Open Data amendment of the Copyright Act. Section 94 of the Copyright Act states that the official work exception applies mutatis mutandis to the database maker. Therefore, in situations when a database is made as a part of a public administration agenda, the sui generis database right legal protection does not apply. Unlike copyright protection, this exception should apply even in a situation when the database itself is created by a third party on behalf of a public sector body through public procurement. The initiative and funding come from the public sector body and thus the public sector body should be the maker of the database. In such cases, the public sector body can provide the whole content of the database for extraction and re-utilisation without any licence.

For example, a university is the operator of a repository of qualification theses. It is the maker of a database made to fulfil its legal duty to make the theses publicly available in accordance with Section 47b of Act No 111/1998 Sb.[19]. Thus, the database of theses should not be

[15] For more on this topic see Hugenholtz, who refers to the European Court of Justice cases British Horseracing Board and Directmedia Publishing. In Hugenholtz, p. 407.

[16] This example deals only with the *sui generis* database right. Copyright to the specific works must also be taken into consideration.

[17] This also covers a search service built upon the database. See Decision of the Court of Justice of the European Union No C-202/12 – Innoweb.

[18] Hugenholtz refers to a Dutch case in which the court used this argument to rule that the city of Amsterdam is not eligible for the sui generis database right. See Hugenholtz, p. 405.

[19] See Act No 111/1998 Sb., on higher education institutions.

protected by sui generis right, and the official work exception should apply. It must be stressed that this does not mean the theses themselves are not protected by copyright.

Act No 298/2016 Sb. which reintroduced the application of the official work exception to sui generis, contains a transitional provision which reads as follows: "This provision does not apply to databases protected by the sui generis right of the database maker obtained before this Act came into force."[20] According to the explanatory memorandum, the reason behind this provision is to limit possible negative outcomes to the existing rights of the makers of databases. There is a practical negative outcome from this provision. If there is a new substantial investment in a database (either changing, obtaining, verifying or presenting the content), the 15-year term of protection recommences. Hence it is very possible that old databases protected before 1 January 2017 will continue to be protected forever as long as substantially important changes are made to them. This results in a situation with a quite low level of legal certainty because it is almost impossible for a layperson to correctly guess in advance whether a database is protected or not.

Conclusion

Operators of grey literature repositories should address the sui generis database right in addition to copyright during the publication process. If such right exists, the operator should provide a licence which would enable the full use of the database content. In cases where the operator is a public sector body and fulfilling an administrative duty, the database made and used for such purpose is not protected by the sui generis database right because of the application of the official work exception. However, it the database was created before 1 January 2017, the protection applies. To enable maximum content use, it is recommended to waive the sui generis database right, for example by using the CC0 licence.

References

HUGENHOLTZ, P. Bernt and Thomas DREIER, ed. *'Directive 96/9/EC', in Concise European Copyright Law.* Second edition. Alphen aan den Rijn, The Netherlands: Kluwer Law International, 2016. pp. 379–420

KOŠČÍK, Michal, Radim POLČÁK, Matěj MYŠKA and Jakub HARAŠTA. *Výzkumná data a výzkumné databáze: právní rámec zpracování a sdílení vědeckých poznatků*. Praha: Wolters Kluwer, 2017. Právní monografie (Wolters Kluwer ČR). ISBN 978-80-7552-952-7.

MYŠKA, Matěj and Jakub HARAŠTA. Less Is More? Protecting Databases in the EU after Ryanair. *Masaryk University Journal of Law and Technology* [online]. 2016, **10**(2), 170-199 [Accessed 19 October 2018]. DOI: 10.5817/MUJLT2016-2-3. ISSN 1802-5951. Available from: **https://journals.muni.cz/mujlt/article/view/5180**

MYŠKA, Matěj and Jakub HARAŠTA. Omezení autorského práva a zvláštních práv pořizovatele databáze v případě datové analýzy. *Časopis pro právní vědu a praxi* [online].

[20] The Act came into force on 1 January 2017.

2015, **23**(4), 375 - 384 [Accessed 19 October 2018]. ISSN 1805-2789. Available from: **https://journals.muni.cz/cpvp/article/view/5276/4362**

SCHÖPFEL, Joachim, 2011. Towards a Prague Definition of Grey Literature. In *Twelfth International Conference on Grey Literature: Transparency in Grey Literature*. Prague, TextRelease [Accessed 19 October 2018]. pp.11-26. Available also from: **https://goo.gl/Jr2Fg1**

TELEC, Ivo and Pavel TŮMA. *Autorský zákon: komentář*. V Praze: C.H. Beck, 2007. Velké komentáře. ISBN 978-80-7179-608-4.

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: **https://nusl.techlib.cz/en/conference/conference-proceedings**

LEGAL FRAMEWORK FOR DIGITALISATION AND STORAGE OF DIGITAL WORKS BY PUBLIC ARCHIVES

Michal Koščík

michalkoscik@gmail.com

Masaryk University, Brno

Abstract

The article observes the legal framework of Czechia for making digital copies and their further use by cultural heritage institutions. The article generally describes the laws regulating copyright exemptions and then describes the functioning of individual exemptions harmonised by InfoSoc directive within the Czech legal system. Finally, the article describes the exemptions not directly presumed by InfoSoc directive that are formulated as a duties of public institutions towards public but involve copyrighted works.

Keywords

InfoSoc, copyright, digitalisation

The publication of this paper is supported by the Czech Scientific Foundation – project no. GA17-22474S “Adapting Exceptions and Limitations to Copyright, Neighbouring Rights and Sui Generis Database Rights to Digital Network Environment”.

Introduction

The aim of this article is to provide a national report on the state of copyright exceptions for the digitisation of works and the use of digital copies by public repositories such as libraries, cultural heritage institutions and educational institutions.

The legal regulation of copyright and related rights is concentrated into the Czech Copyright Act[1] ("CA"). The CA covers most of the substantive rules on copyright, related rights and database rights, including statutory exceptions and rules on the collective management of copyright rights. Basically, all the EU acquis on copyright and related rights has been implemented into the wording of the CA[2].

The concept of the CA as a single act that would cover all substantive rules related to copyright has, however, eroded over time. The rules on license agreements are currently part of the Czech Civil Code[3] and certain statutory copyright exceptions are implemented in other Acts [4]. The statutory exceptions outside the CA are usually formulated as the duty of an institution to make, preserve or enable a copy of a certain work, rather than rights that the institution can exercise at will. Typical examples of such statutory exceptions are the duty of a university to make all theses and dissertations available to the public[5], or the duty of the owner of a cultural archival relic to provide a backup "security" copy of the item[6].

Apart from the CA, the relevant provisions for digitising and preserving digital documents can be found in the abovementioned Act on Archiving and Records Management[7], the Library Act and the Free Access to Information Act[8]. The legal deposit obligations are regulated by the Act on Non-periodical Publications[9] and the Act on Periodical (Print) Publications[10]. These laws contain provisions that could be perceived as statutory exceptions or obligatory licenses and will be addressed below.

Making a digital copy part of the collection

The purpose of a library, archive or museum is to collect items, preserve them and make them available for the public[11]. Acquisition is traditionally performed through the purchase of a tangible object that is itemised and put into storage. Items in storage are subsequently indexed/catalogued and collectively form the institution's collection. The acquisition

[1] Act No 121/2000 on Copyright and Rights Related to Copyright and on the Amendment of Certain Acts.

[2] For example, the CA contains provisions harmonised by legislation on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc), the Software Directive, the Database Directive and the Orphan Works Directive.

[3] Act No 89/2012, the Civil Code, Sections 2358 to 2389.

[4] See below.

[5] See Section 47b of Act No 111/1998, the Higher Education Act.

[6] Act No 499/2004, on archiving and records management.

[7] Act No 499/2004, on archiving and records management.

[8] Act No 106/1999, the Free Access to Information Act.

[9] Act No 37/1995, on non-periodical publications.

[10] Act No 46/2000, on periodical publications.

[11] For example, Section 2 of Act No 122/2000, on the protection of museum character collections and on the amendment of certain other Acts, defines a museum as an institution that acquires, collects, preserves and indexes natural objects or human works and makes them available on equal terms to everybody. A gallery is defined as a special form of museum focused on the visual arts. The Library Act defines a library as an institution that provides a library service to everybody (Section 2), which also includes making the library collection available (Section 4).

of a tangible object is typically performed through donation, purchase or a legal obligation to hand over certain items such as a public deposit obligation.

Acquiring digital content into a collection usually relies on a different set of legal instruments. Instead of purchase or donation, the acquisition is usually performed through a license agreement or by taking advantage of a statutory exception. The use of licenses for public archive purposes has already been broadly covered elsewhere[12]. Below, we focus mainly on the statutory exceptions.

Statutory exception for preservation purposes

A physical (analogue) medium such as paper deteriorates over time or can be damaged. It is often therefore important to make backup copies to preserve the work or document it contains. Making a backup copy is explicitly permitted by a statutory exception contained in Czech law[13]. The statutory exception contained in Section 37(1)(a) of the Czech Copyright Act enables every library, archive, museum or educational institution to create a copy of any work to preserve that work for its internal archival or conservational purposes.

The formulation of the exception is very broad and does not limit its scope only to certain works as long as the copy does not serve direct or indirect commercial purposes. This exception can also be applied by analogy to records and documents protected by so-called related rights or rights related to copyright, such as broadcasting records or rights of performing artists. Curiously enough, the exception for making a copy for archiving and preserving purposes is not applicable to data contained in databases protected by the sui generis right and for computer programs[14]. The archive or library, however, may make a permanent copy of a computer program under statutory license for backup as determined in Section 66(3) of the CA, which is in fact broader in terms of possible uses than the "archiving exception".

The important question not explicitly answered by the Czech Copyright Act is whether a library or archive can create a copy of a document that was previously not in the collection of the institution. In other words, whether a library can use the exception in Section 37 of the CA to make a certain work part of its collection. For example, can a library make a copy of a literary work that has been uploaded in a pirate repository? Can a public repository make a copy of a publicly available website solely based on the archiving exception?

Historically, the exception to making a copy for archiving purposes was written to enable institutions to create backup copies of items contained in their collections. The explanatory report to the first wording of the CA states that this exception was introduced so that the backup copies can be used in the place of other copies that were "lost, damaged or destroyed".

[12] See: DAVIS, Trisha L. License agreements in lieu of copyright: Are we signing away our rights?. *Library Acquisitions: Practice & Theory* [online]. 1997, **21**(1), 19-28 [Accessed 19 October 2018]. DOI: 10.1016/S0364-6408(96)00085-3. ISSN 03646408. Available from: **http://linkinghub.elsevier.com/retrieve/pii/S0364640896000853**; ŠAVELKA, Jaromír and Michal KOŠČÍK. Jaké podmínky musí splňovat autorské dílo, aby mohlo být vloženo do veřejně přístupného repozitáře? In: *Seminar to access of grey literature 2010: 3rd year of the seminar* [online]. Praha: Národní technická knihovna, 2010. ISSN 1803-6015. Available from: **https://nusl.techlib.cz/cs/konference/sbornik-2010**; KOSCIK, Michal; SAVELKA, Jaromir. Dangers of Over-Enthusiasm in Licensing under Creative Commons. Masaryk UJL & Tech., 2013, 7: 201.; MYŠKA, Matěj. Vybrané právní aspekty otevřeného přístupu k vědeckým publikacím. Právní rozhledy, 2014, 22.18: 611-619.

[13] Making a copy of any work in the collection of a library or archive is explicitly permitted by Section 37 of the Czech Copyright Act.

[14] Due to the explicit exception contained in Section 66(7) of the CA.

However, the wording of the exception for archiving purposes does not explicitly state that the exception can only be enjoyed by the legitimate holder or user (like in the exceptions under Sections 36 or 66 of the Czech Copyright Act). The grammatical formulation of the exception is rather broad. The wording has to be interpreted in line with the general CA clause, in particular in accordance with the provisions on a "three step test" in Section 29. The archiving exception can also be enjoyed by a user without ownership rights to the medium from which the copy is made as long as this does not conflict with the normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author. The legislation regulates and restricts how a particular copy may be utilized further, and therefore the mere act of creating a copy for archiving or preservation purposes does not have the potential to interfere with the commercial interests of the rightsholder and is also easily foreseeable by any author who publishes his/her work. It is also important to note that the license to make a copy for preservation purposes is granted even for works where the author has not decided to make such works public[15]. It is clear that Czech legislation attributes significant importance to the preservation of cultural heritage, even in cases where the author or rightsholder does not consent to preservation of the work. It can therefore be concluded that a library or archive can make a preservation copy of a cultural heritage item even if the source has questionable rights regarding making the work available to the archive or library in the first place. It is relatively easy for any library or archive to digitise any of its books or items regardless of the nature of the work or the existence of copyright protection for the work.

Making a digital copy accessible for users

Making a digital copy of a work to protect it from destruction or loss is only the first part of a library's or archive's work. It is understandable that an archive needs to share its documents with the public. Making a digital copy available to users is considered a use of a copyrighted work and either a license or statutory exception is needed. The CA provides for three major copyright exceptions for libraries, public archives, schools and other educational institutions to make the content of their collections accessible to the general public.

The first of these exceptions is the right to lend a "lawfully" made reproduction of a work that has been damaged or lost, providing that the work no longer has any commercial nature[16]. The CA does not restrict such copies to analogue form, and this provision might be applicable to lending digital copies, especially after it was made clear by the Court of Justice that the concept of lending can extend to digital copies[17].

The second exception allows public archives and libraries[18] to display digital copies via dedicated terminals located on their premises, such a work being made available in this way exclusively for the purposes of research or private study by members. The digital copy needs to be a copy of an item that is part of the institution's collection. We described above that it is relatively easy for a public archive or library to make a digital copy of any item and make it a part of its collection. Hence the three-step test has to be applied to determine whether the

[15] See Section 29(2) of the CA.

[16] Section 37(1)(b) of the CA.

[17] C-174/15 - Vereiniging Openbare Biblioteken ECLI:EU:C:2016:856.

[18] The exception extends to library, archive, museum, gallery, school, university and other non-profit school-related and educational establishments.

actual on-site display does not conflict with the normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author[19]. As the case law of the CJEU[20] shows, printouts from terminals can be considered the use of a work that would need either license or compensation via the collective management. Czech legislation enables the provision of printed copies of such reproductions via an intermediary (upon request), providing that the intermediary pays the remuneration to the collective rights manager. The aspects of collective rights management of digital works will not be discussed further, as this topic has recently been discussed by Straková[21]. Another practical problem faced by the archives is that the use of digital works can be prohibited by contractual arrangements that take precedence over the statutory exception. This puts major parts of contemporary content that are usually distributed digitally outside the scope of this exception. Unlike the exception for "preservation purposes", the exception for displaying digital copies via terminals does not relate to works that have not yet been made available to the public[22].

The third exception relates to orphan works. Again, this exception will not be discussed in detail as it has recently been addressed on the same forum by Myška[23].

Statutory limits to exceptions

The limits of the abovementioned exceptions are contained in the general clauses of the CA (Section 29) which, in addition to the three-step test, contain specific protection for unpublished works. The second paragraph of Section 29 CA stipulates that the exceptions can be generally enjoyed only in relation to works that have already been published. This restriction was a significant obstacle to repositories of grey literature because large parts of the works archived as grey literature were never created with the intention of being published[24] (such as minutes of meetings and sessions) or were auxiliary records (such as technical drawings). The restriction from Section 29 CA did not serve much purpose as regards protection of rights-holders, since the mere creation of a copy for archiving purposes did not automatically give a repository the right to publish it. This restriction was partially lifted in 2017 when an amendment to the CA[25] enabled libraries and museums to collect even unpublished works. The practical impacts of these restrictions remain low because repositories still have to rely on the authors' consent[26] if they want to publish unpublished works in their collections.

Another statutory restriction to the digitisation of works was discussed by Telec in his work on the digitisation of films[27], where he concluded that the archiving exception does not allow for the remastering of original works and that the original work also has to be archived with its flaws and imperfections.

[19] Section 29(1) of the CA.

[20] C-117/13, Technische Universität Darmstadt, ECLI:EU:C:2014:2196.

[21] See: STRAKOVÁ, Lucie. Changes in the Area of Extended Collective Management in Relation to Memory and Educational Institutions in the Light of the Czech Amended Copyright Act. An International Journal on Grey Literature, TextRelease, 2018, ron. 14, Special Winter Issue, p. 61-65. ISSN 1574-1796.

[22] See below.

[23] MYŠKA, Matěj. Orphan and Out-Of-Commerce Works After the Amendment of the Czech Copyright Act. The Grey Journal, Amsterdam: TextRelease, 2018, ron. 14, Special Winter Issue, p.55-60. ISSN 1574-180X.

[24] Instead, they were meant to be used as documents for internal use.

[25] Act No **102/2017**, amending the CA.

[26] Or eventually the death of the author (if known).

[27] TELEC, Ivo. Digitalizace filmů. Právní rozhledy, Prague: Nakladatelství C. H. Beck, s. r. o., 2015, vol. 23, 15/16, p. 526-528.

Exceptions for digitisation outside the CA

As mentioned in the introduction, most statutory exceptions outside the CA are not construed as exceptions but rather as duties to provide some kind of public service. Of these exceptions, we can list the obligation of tertiary educational establishments to make all theses and dissertations available to the public, including reviews[28], the obligation to make digital conversions in the Act on Archiving[29] and the obligation of every public institution to provide information under the Freedom of Information Act ("FIA").

The general principle of the FIA is that any public entity must, on request, provide basically any information it has generated or has in its possession. Therefore, the archives of public entities can be a valuable source of content for museums and libraries. On the other hand, the content generated by libraries and museums can also fall under the scope of the FIA. One of the limits to this obligation is a third party intellectual property right to the requested document or information[30]. Certain intuitions, such as universities and orchestras, are also excepted from providing copyrighted works they created, however libraries and museums do not fall into this privileged category.

Interestingly, the FIA gives the public entity who provides the information or document an explicit license to digitise the requested document and even to display such document online[31]. This, however, cannot be abused to circumvent or broaden the limits to copyright exceptions imposed by the general clause contained in Section 29 CA if the document is protected by third persons' rights.

The obligation of a public depository that is imposed on every publisher of periodic or non-periodic publications is a specific form of statutory exception. Publications delivered to selected libraries through the public deposit obligation can be digitised for preservation purposes. The benefits of handing over legal deposits purely in electronic form have already been described by Polčák[32]. However, the concept of electronic legal deposit remains unknown to Czech legislation, even if the question of whether purely digital publications meet the definition of non-periodic publications remains open[33], but largely academical, because libraries and archives are entitled to make digital copies for preservation purposes based on the abovementioned exception.

Copyright rights to digitised copies

Czech legislation does not recognize any specific intellectual property rights to the digitisation of analogue copies. Putting a document into a scanner does not make it a unique product of the author's creative activity, which is a necessary criterion for copyright protection. It is, however, possible that collecting, selecting, scanning and presenting large volumes of digitised documents would constitute a sui generis database right for a repository operator. It is also possible that certain results of elaborate digitisation and three-dimensional modelling would

[28] Act No 111/1998, on tertiary education, Section 47b.

[29] See Act No 499/2004, on archiving and records management, Section 13(5); Section 15 and Section 69a.

[30] See Section 11(2) of Act No 106/1999, the Freedom of Information Act.

[31] See Section 4a FIA

[32] POLČÁK, Radim. Práva k datům spravovaným veřejnými knihovnami ve světle změn informačního zákona. Knihovna, Prague: National Library of the CR, 2016, vol. 27, n. 1, p. 61-73. ISSN 1801-3252. Cf. MATUŠÍK, Zdeněk. K některým autorskoprávním otázkám činnosti knihoven v současnosti. Knihovna plus.

[33] Act No 37/1995, on non-periodical publications, explicitly excludes computer programs and audio-visual works.

acquire copyright protection. Polčák warns that "the possibility of copyright protection for digital images might then provide for the emergence of subsequent copyright protection of old cultural heritage that itself is not protected by copyright"[34]. Czech legislation solves this problem by explicitly stating that libraries, museums and galleries are obliged to provide their own intellectual property upon request under the FIA[35].

Conclusion

The aim of this article was to describe the current state of copyright exception for the digitisation and making digital copies available via cultural heritage and educational institutions in Czechia. Czechia has implemented virtually all the applicable exceptions contained in the InfoSoc directive, however did not go much further beyond the basic exceptions. The regulation on the statutory deposit of electronic digital copies or the creation of web-based archives is still unsatisfactory to non-existent. The current state of copyright exceptions limits cultural heritage institutions to using these exceptions for contemporary content subject to licensing terms and conditions. We have, however, identified that many statutory licenses are in fact disguised as obligations to perform a public service, especially in the field of archiving and tertiary education.

References

DAVIS, Trisha L. License agreements in lieu of copyright: Are we signing away our rights?. *Library Acquisitions: Practice & Theory* [online]. 1997, **21**(1), 19-28 [Accessed 19 October 2018]. DOI: 10.1016/S0364-6408(96)00085-3. ISSN 03646408. Available from: **http://linkinghub.elsevier.com/retrieve/pii/S0364640896000853**

MATUŠÍK, Zdeněk. K některým autorskoprávním otázkám činnosti knihoven v současnosti. *Knihovna plus* [online]. 2010, n. 1 [Accessed 19 October 2018]. ISSN 1801-5948. Available from: **http://knihovna.nkp.cz/knihovnaplus101/matus.htm**

MYŠKA, Matěj. Vybrané právní aspekty otevřeného přístupu k vědeckým publikacím. *Právní rozhledy*. Nakladatelství C. H. Beck, 2014, **22**(18), p. 611-619. ISSN 1210-6410.

MYŠKA, Matěj. Orphan and Out-Of-Commerce Works After the Amendment of the Czech Copyright Act. *The Grey Journal.* Amsterdam: TextRelease, 2018, **14**(Special Winter Issue), p. 55-60. ISSN 1574-180X.

POLČÁK, Radim. Práva k datům spravovaným veřejnými knihovnami ve světle změn informačního zákona. *Knihovna.* Praha: Národní knihovna ČR, 2016, **27**(1), p. 61-73. ISSN 1801-3252.

[34] POLČÁK, Radim. Digitisation, Cultural Institutions and Intellectual Property. Masaryk University Journal of Law and Technology, Brno: Masaryk University, 2015, vol. 9, n. 2, p. 121-141. ISSN 1802-5943. DOI:10.5817/MUJLT2015-2-7.
[35] See Section 11(5)(d) of the Freedom of information Act.

POLČÁK, Radim. Digitisation, Cultural Institutions and Intellectual Property. *Masaryk University Journal of Law and Technology.* Brno: Masaryk University, 2015, **9**(2), p. 121-141. DOI: 10.5817/MUJLT2015-2-7. ISSN 1802-5943.

TELEC, Ivo. Digitalizace filmů. *Právní rozhledy.* Praha: Nakladatelství C. H. Beck., 2015, **23**(15/16), p. 526-528.

STRAKOVÁ, Lucie. Changes In The Area Of Extended Collective Management In Relation To Memory And Educational Institutions In The Light Of The Czech Amended Copyright Act. *An International Journal on Grey Literature.* Amsterdam: TextRelease, 2018, **14**(Special Winter Issue), p. 61-65. ISSN 1574-1796.

ŠAVELKA, Jaromír and Michal KOŠČÍK. Jaké podmínky musí splňovat autorské dílo, aby mohlo být vloženo do veřejně přístupného repozitáře? In: *Seminar to access of grey literature 2010: 3rd year of the seminar* [online]. Praha: Národní technická knihovna, 2010. ISSN 1803-6015. Available from: **https://nusl.techlib.cz/cs/konference/sbornik-2010**

Judicial documents and statutory law

Court of Justice of the European Union, C-117/13 Judgment of the Court (Fourth Chamber), Technische Universität Darmstadt v Eugen Ulmer KG; 11 September 2014, ECLI:EU:C:2014:2196

Court of Justice of the European Union, C-174/15 Judgment of the Court (Third Chamber), Vereniging Openbare Bibliotheken v Stichting Leenrecht; 10 November 2016, ECLI:EU:C:2016:856

Act No. 111/1998 Coll. Higher education Act

Act No. 106/1999 Coll. Freedom of Information Act

Act No. 121/2000 Coll. on Copyright and Rights Related to Copyright and on Amendment to Certain Acts

Act No. 122/2000 Coll. on the Protection of Collections of Museum Character and the Amendment of Certain Other Laws

Act No. 89/2012 Coll., Civil Code,

Act No. 499/2004 Coll. on Archiving and records management

Act No. 37/1995 Coll. on Non-Periodical Publications

Act No. 46/2000 Coll. on Periodical publications

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: **https://nusl.techlib.cz/en/conference/conference-proceedings**

LOCKSS DISTRIBUTED DIGITAL PRESERVATION NETWORKS

Anthony Leroy

anthony.leroy@ulb.ac.be

Université libre de Bruxelles, Belgium

Abstract

As university libraries, preserving digital objects for future generations is one of our key missions.

This paper briefly discusses the essential features required for an ideal digital preservation solution to mitigate the many risks that endanger our digital assets.

The LOCKSS open source technology can help libraries build a robust distributed digital preservation network to ensure the very long-term availability of our scientific heritage. Many examples of existing implementations illustrate the wide variety of preservation networks currently based on the LOCKSS software.

Keywords

Digital preservation, LOCKSS, distributed preservation network

Introduction

In our digital era, the production of information grows at an unbridled rate. Access to the information has also become easier and faster than ever.

However, digital information also became much more fragile and vulnerable. A study from the University of British Columbia estimates that more than 80% of the research data at the origin of publications in zoology dating from the nineties is definitely lost (Vines, 2014).

The preservation of digital assets is a complex matter. Analog media such as paper or microfilms can be preserved for hundreds of years just by ensuring appropriate environmental storage conditions. By contrast, preserving digital objects in order to make sure that they will be reusable in the very long term requires elaborated strategies and rigorous processes to protect them against a large variety of threats. Most of these threats are not easily quantifiable and many catastrophic events have very low probability to occur but in the long term, some will inevitably happen.

Evaluating the risks

It is generally thought that the main risks concern hardware breakdowns, obsolescence or natural disasters and mitigation measures tend to focus mainly on those aspects.

However, in practice, it is observed that data losses find mostly their origin in human errors, external or internal computer attacks, financial or organizational problems (Rosenthal, 2005).

To mitigate these risks, the commonsense solution consists in making multiple copies. As a matter of fact, having more copies of the digital objects is the criteria which has the largest impact on the preservation solution reliability (Rosenthal, 2011)**.** Storing the copies on more reliable or more heterogeneous hardware will help but it will actually have a lower impact on overall reliability than having more copies.

Just how many copies are required depends on the threat model which includes the financial risk of managing too many expensive copies. Also, the law of diminishing returns states that over a certain number of copies, the cost of an extra copy is not paying off compared to the relative reliability improvement.

There is thus no single answer to this question but having many copies on consumer-grade media is much more reliable than having few copies on advanced expensive hardware.

Hence, the statement that « Lots of Copies Keep Stuff Safe » or LOCKSS (Maniatis, 2005).

Georeplication and Diversity for Better Preservation

Of course, having many copies is not enough. To protect them from small-scale disasters, it is necessary to disseminate them throughout the world, in places considered safe from natural and man-made hazards.

The management of each individual archive copy should also be left to parties which are autonomous and independent on the financial, administrative and organizational levels. It is also necessary to check the integrity of the data regularly and, if required, to migrate the files to long-term sustainability format.

No outsourced solution can guarantee data preservation according to these criteria. For reasons of economic profitability, commercial companies resort to the mutualization of technical and human resources and adopt the most profitable technology. The only guarantee offered to the customer of such third-party services is the existence of a contract stating that the supplier provides assurance to deploy economically “reasonable” efforts to preserve customer data and, possibly, in the event of an unfortunate loss, to grant a compensation.

The Role of University Libraries

The preservation of academic and scientific heritage is the responsibility of university libraries (Skinner, 2010). Digital objects of scientific interest that need to be shared and preserved are commonly stored in institutional repositories: theses, scientific publications and research data but also the grey literature for which the university is generally the only holder: internal reports, working papers, laboratory notebooks, white papers and research data.

Collaboration for an efficient preservation solution

A university library alone is unlikely to have at their disposal the human and technological resources needed to build an efficient preservation solution.

Collaboration between institutions to build a distributed preservation network is thus inevitable.

The organizational structure of a distributed preservation network can take multiple forms. The network can emerge from a pre-existing organization seeking for a preservation solution for their own locally created resources or it can be specifically created by a community for the preservation of objects of global interest.

Distributed preservation networks need to rely on a common technological infrastructure to support, at low cost and with limited human intervention, the coordination of the multiple archive copies which are stored on independent preservation nodes.

One of the most important services that a preservation infrastructure should provide is a mechanism to automatically and regularly check the integrity of the preserved copies.

The LOCKSS technology precisely offers one of the most sophisticated integrity monitoring services.

LOCKSS: a state-of-the-art technology

The LOCKSS technology is implemented as an open-source software originally developed for the global LOCKSS network to provide a solution for the post cancellation access or perpetual access to subscription e-journals and e-books.

It is an awarded technology: in 2014, it received the first ever perfect score when audited for the TRAC certification of the CLOCKSS archive (Rosenthal, 2014).

What makes LOCKSS software unique is precisely its robust integrity check and repair protocol enabling a secure audit mechanism between independently administered preservation nodes to test the integrity of the distributed copies. In the event of a corruption, the altered copy can be automatically repaired based on the valid copies (Maniatis, 2005).

LOCKSS mainly addresses bit-level preservation ensuring the preserved bits remain unchanged. The advanced logical-level preservation activities, such as format normalization ensuring that the bits will still be interpreted correctly, are left at LOCKSS network users' discretion.

LOCKSS: a vibrant community

The organizations using LOCKSS networks to preserve their digital content constitute a large and vibrant international community. Currently, counting a dozen networks and hundreds of institutions, the LOCKSS community is constantly growing.

While the Global LOCKSS network is probably the most famous LOCKSS preservation network, there are many other networks exploiting the same technology to preserve a large variety of content with various scopes, goals, governance and membership models (Reich, 2009).

Some examples are provided hereunder, classified by types of preserved content.

LOCKSS Networks Preserving General Interest Scientific Content

- The Global LOCKSS Network (GLN) is the first and largest LOCKSS network, counting more than 200 participating institutions all over the world. It preserves the content from over ten thousand e-journals from a wide variety of commercial and learned society publishers (Publishers & Titles, 2018).
- The Controlled LOCKSS (CLOCKSS) network is a dark archive jointly governed by academic publishers and university libraries to ensure the long-term survival of scholarly publications (CLOCKSS, 2018).
- *The Public Knowledge Project Preservation Network (PKP PN) is a dark archive preserving OJS journals. The PKP PN currently preserve more than 700 OJS journals and primarily targets the content not preserved elsewhere* (PKP Preservation Network, 2018).

LOCKSS Networks Preserving Government Information

- The Canadian Government Information (CGI) Digital Preservation Network, preserves digital collections of Canadian government documents (Wakaruk, 2013).
- The USDocs network preserves US digital government documents.

LOCKSS Networks Preserving Scientific Journals of Local Interest to a Community

- The Council of Prairie and Pacific University Libraries (COPPUL) Network preserves e-journals from member university libraries and small journals in Western Canada (COPPUL, 2015).
- The Cariniana Network preserves over 1000 Brazilian open access journals from Sistema Eletrônico de Editoração de Revistas (SEER) (A Cariniana e a Aliança LOCKSS da Stanford University, 2018).

LOCKSS Networks Preserving Content of Local Interest to a Community, including grey literature

- The Alabama Digital Preservation Network (ADPNet) preserves digital content locally created in Alabama. It proposes a low-cost digital preservation solution for academic institutions, state agencies, and cultural heritage organizations in Alabama (Alabama Digital Preservation Network, 2018).
- The WestVault Network provides a distributed digital preservation storage network spread across 4 western Canada provinces to preserve critical digital content submitted through an ownCloud instance (WestVault, 2018).
- The MetaArchive Distributed Digital Preservation Network is an international dark archive run by the Educopia institute to preserve high value locally created digital materials for more than 60 member libraries, archives, and museums (MetaArchive Cooperative, 2018).
- The SAFE LOCKSS Network is an international distributed archive preserving the born-digital open-access collections from a variety of institutional repositories managed by the participating member (Leroy, 2015).

The SAFE Network

The SAFE Archive Federation (SAFE) LOCKSS Network provides an interesting example of an organization solely built for the purpose of preserving content from the institutional repositories of participating institutions (including a good proportion of grey literature materials).

The network is a federation of completely independent institutions sharing a common view on how their own digital collections must be preserved. It is based on a light organizational structure around a simple memorandum of understanding. The budgets of participating institutions remain fully independent. Each member of the network simply agrees to make available a portion of their preservation node storage to keep copies of their partners' content.

SAFE is an international network ensuring an efficient replication of the archives in completely independent sites on the organizational level. SAFE has preservation nodes in Belgium, Canada, Germany, Sweden and Switzerland.

Each member keeps full technical control on their preservation node as only local administrators can manage the content of their network node.

Partners are however able to check the status of their preserved content over the network thanks to a monitoring tool collecting and aggregating status information from the nodes.

The LOCKSS technology is particularly well suited to enable this type of collaboration.

The future of LOCKSS

The original LOCKSS software was designed almost two decades ago at a time when collecting digital objects from static web pages was straightforward. It then gradually evolved to fit the needs of the increasingly dynamic web content, making the software more and more complex and difficult to maintain.

In 2017, LOCKSS was awarded a Mellon Foundation Grant to modernize the LOCKSS codebase by rearchitecting the software as a collection of Web Services with fully documented REST-APIs, a project named LOCKSS Architected As a Web Service (LAAWS) (Guicherd-Callin, 2018).

The new architecture relies on the state-of-the art open-source community software for the non-core-business modules (such as web crawling or content dissemination) and will align to the web archiving standard WARC. It will also support large-scale distributed storage to cope with the ever-increasing preservation storage needs.

In particular, the core LOCKSS component that provides the state-of-the-art data integrity monitoring service will be de-siloed into an independent web service. This will undoubtedly facilitate the reuse of LOCKSS in diverse contexts where advanced integrity check is needed; which will certainly result in a significant increase of the LOCKSS user base in the coming years.

Coincidently, the LOCKSS community is consolidating by creating users and developer groups to share best practices and develop community tools across networks.

Conclusion

The LOCKSS technology empowers university libraries to fulfill their essential mission of preserving digital knowledge by collaborating with other institutions to build a robust distributed preservation network. Many examples have been provided to illustrate the wide variety of existing networks currently employing the LOCKSS software.

The future of LOCKSS is exciting: the software is being re-architected to meet state-of-the art technology standards and the user community is looking forward to welcome and support the new preservation networks that will emerge to secure the access to unique knowledge for future generations.

Acknowledgment

We would like to acknowledge the great work of the LOCKSS team at Stanford University in developing and promoting this remarkable technology.

Of course, none of this would be possible without the groundbreaking work of Dr. David Rosenthal. His thoughtful ideas and clear vision greatly inspire the whole community. His blog, blog.dshr.org, is an invaluable source of information for anyone interested in digital preservation.

References

Alabama Digital Preservation Network, 2018. *Alabama Digital Preservation Network* [online]. Alabama Digital Preservation Network [Accessed 25 September 2018]. Available from: **http://www.adpn.org/**

CLOCKSS, 2018. *CLOCKSS* [online]. [Accessed 24 September 2018]. Available from: **https://clockss.org/**

COPPUL, 2015. COPPUL Private LOCKSS Network Governance Policy. In: *Council of Prairie and Pacific University Libraries (COPPUL)* [online]. COPPUL, 2015 [Accessed 25 September 2018]. Available from: **https://coppul.ca/pln-governance**

COPPUL, 2018. WestVault. In: *Council of Prairie and Pacific University Libraries (COPPUL)* [online]. COPPUL, 2018 [Accessed 25 September 2018]. Available from: **https://coppul.ca/westvault**

GUICHERD-CALLIN, Thib, 2018. LOCKSS Software Re-Architecture. In: *34th International Conference on Massive Storage Systems and Technology* [online]. Santa Clara [Accessed 25 September 2018]. Available from: **http://storageconference.us/2018/Presentations/LOCKSS-tutorial-4.pdf**

LEROY, Anthony and Patrick HOCHSTENBACH, 2015. SAFE PLN: An International Preservation and Access Solution. In: *D-Lib Magazine: In Brief* [online]. July/August 2015 [Accessed 25 September 2018]. Available from: **http://www.dlib.org/dlib/july15/07inbrief.html**

MANIATIS, Petros, Mema ROUSSOPOULOS, T. J. GIULI, David S. H. ROSENTHAL and Mary BAKER, 2005. The LOCKSS peer-to-peer digital preservation system. *ACM Transactions on Computer Systems* [online]. **23**(1), 2-50 [Accessed 24 September 2018]. DOI: 10.1145/1047915.1047917. ISSN 0734-2071. Available from: **http://portal.acm.org/citation.cfm?doid=1047915.1047917**

METAARCHIVE COOPERATIVE, 2018. MetaArchive Resources. In: *MetaArchive* [online]. Atlanta: Educopia Institute [Accessed 25 September 2018]. Available from: **https://metaarchive.org/documentation-resources/**

PKP Preservation Network. In: *Public Knowledge Project* [online]. Simon Fraser University Library, 2014 [Accessed 24 September 2018]. Available from: **https://pkp.sfu.ca/pkp-pn/**

A Cariniana e a Aliança LOCKSS da Stanford University. In: *Portal da Rede Cariniana* [online]. 2018 [Accessed 25 September 2018]. Available from: **http://cariniana.ibict.br/index.php/noticias/377-a-cariniana-e-a-alianca-lockss-da-stanford-university**

Publishers & Titles (GLN). In: *Lots Of Copies Keep Stuff Safe* [online]. Stanford: Stanford University, 2018 [Accessed 24 September 2018]. Available from: **https://www.lockss.org/community/publishers-titles-gln/**

REICH, Victoria and David ROSENTHAL, 2009. Distributed Digital Preservation: Private LOCKSS Networks as Business, Social, and Technical Frameworks. *Library Trends*. **57**(3), 461-475 [Accessed 24 September 2018]. DOI: 10.1353/lib.0.0047. ISSN 1559-0682.

ROSENTHAL, David S. H., Thomas ROBERTSON, Tom LIPKIS, Vicky REICH and Seth MORABITO, 2005. Requirements for Digital Preservation Systems. *D-Lib Magazine* [online]. **11**(11) [Accessed 24 September 2018]. DOI: 10.1045/november2005-rosenthal. ISSN 1082-9873. Available from: **http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html**

ROSENTHAL, David, 2011. How Few Copies? In: *DHSR's blog* [online]. 2011-03-15 [Accessed 24 September 2018]. Available from: **https://blog.dshr.org/2011/03/how-few-copies.html**

ROSENTHAL, David, 2014. TRAC Certification of the CLOCKSS Archive. In: *DHSR's blog* [online]. 2014-07-24 [Accessed 24 September 2018]. Available from: **https://blog.dshr.org/2014/07/trac-certification-of-clockss-archive.html**

SKINNER, Katherine, Matt SCHULTZ and METAARCHIVE COOPERATIVE (U.S.), 2010. *A guide to distributed digital preservation*. Atlanta, Ga.: Educopia Institute. ISBN 978-0-9826653-0-5.

VINES, Timothy H., Arianne Y.K. ALBERT, Rose L. ANDREW, et al., 2014. The Availability of Research Data Declines Rapidly with Article Age. *Current Biology* [online]. **24**(1), 94-97 [Accessed 24 September 2018]. DOI: 10.1016/j.cub.2013.11.014. ISSN 09609822. Available from: **https://linkinghub.elsevier.com/retrieve/pii/S0960982213014000**

WAKARUK, Amanda, 2013. Introducing the CGI-PLN: Using the LOCKSS Program to Preserve DSP Content in a Changing Environment. In: *Government Information Day* [online]. Toronto [Accessed 24 September 2018]. Available from: **https://era.library.ualberta.ca/items/33a42ce9-a1d8-42b9-ba2c-3b8b0a776dbe**

11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: **https://nusl.techlib.cz/en/conference/conference-proceedings**

ARCLIB – LTP SOLUTION FOR LIBRARIES

Eliška Pavlásková

eliska.pavlaskova@ruk.cuni.cz

Library of the Czech Academy of Sciences

Zdeněk Vašek

zdenek.vasek@ruk.cuni.cz

Library of the Czech Academy of Sciences

Abstract

The presentation introduces project ARCLib. The project aims to create complex open source Long Term Preservation solution for libraries. ARCLib ensures long term preservation of digital data according OAIS guidelines and provides a free alternative to commercial software solutions. ARCLib is designed as a solution for all types of memory institutions – museums, galleries and archives. As part of the project two methodical guidelines were created – Methodology for logical preservation of digital data and Methodology for bit preservation.

Keywords

Long term preservation, open source, ARCLib, OAIS

This article was created within the research programme „Programme of applied research and experimental development of national and cultural identity 2016 – 2020 (NAKI II)“, for the „ARCLib Project – Designing complex Solution for Long-term Preservation in Czech Libraries“, project identification DG16P02R044.

Introduction

Work on the ARCLib project has been ongoing since 2016. The project responds to the need for memory institutions and, in particular, libraries to ensure the long term preservation of digital documents. There are several commercial and open-source solutions on the market at present which cover the issue of LTP (Long Term Preservation) to a greater or lesser extent. Of course there is no open solution within the Czech environment that would comprehensively cover the needs of libraries (and other types of memory institutions) relating to the long term storage of digital objects and, at the same time, enjoy a broader community of users. These are mainly regional and specialised libraries that need to ensure the long term preservation of such documents, which libraries have been working with for some time now and which are gaining in importance. Such institutions, however, do not have the funds to be able to bring in the bigger team required to implement a more robust open source solution and, simultaneously, frequently want to maintain the flexibility which an open source provides. It is precisely for these institutions that the results of the ARCLib project are intended.

The principal planned outcome is a newly-created open source tool for the long term preservation of digital documents prepared on the basis of OAIS and experience of other tools in use in the Czech environment. In addition to its own software tool, methodical materials will be created and subsequently made public in the form of open access. The use of a software tool and methodologies will make it possible to protect, over the long term, digital data and institutions that cannot call on a large team of specialised workers.

The first year of the solution involved the preparation of detailed technical and procedural specifications of tender dossiers for the development of the system. InQool, the company which won the tender, took on the task of actual development. A prototype now exists and individual functional requirements are being tested on this. Information about the project is available at **https://arclib.cz/**.

The objectives and development of a project solution

The objective of the project is to create a comprehensive LTP solution known as ARCLib on an open source basis that uses freely-available tools and systems. One part of the project, and at the same time a significant product of the project, is the creation of a methodology for the logical, long-term preservation of digital data that takes into consideration international standards in this area (reference model OAIS – ČSN ISO 14721 and ČSN ISO 16363 standards) and systems used to create digital data at Czech libraries and makes these accessible.[1] A methodology and solutions for the physical storage of data and the assurance of bit-level preservation will be prepared at the same time.

The whole project is the result of the changing situation in the sphere of Czech libraries (see, for example, Hutař and Melichar, 2014). The long-term management and preservation of digital documents (digital born and digitised) is becoming less and less a specialised matter that only those interested devote their attention to. It is necessary to ensure both "bit-level preservation" of data (safeguarding against physical loss, alteration or crashes involving digital files and carriers) and, at the same time, logical protection (safeguarding against the negative

[1] HUTAŘ, J., A. MIRANDA, E. PAVLÁSKOVÁ, Z. VAŠEK and Z. HRUŠKA. *Metodika logické ochrany digitálních dat.* 2018. Available from: **http://hdl.handle.net/11104/0282107**

impacts of changes and the ageing of information technologies and data formats on the availability and usability of digital information).

The ever-increasing spectrum of libraries and other institutions in the Czech Republic now has to preserve digital documents. Nonetheless, long term preservation and logical preservation in line with the concept of OAIS remain a costly business. First of all there is the need to acquire a software tool that enables storing and management of a large number of documents. At present, libraries are offered the option of acquiring a commercial solution or of using the progressive development of the Archivematica open source tool or other freely-available tools, although these invariably require extensive adaptation to meet the needs of a specific institution and the development of new parts. Commercial solutions are at present frequently based on a cloud basis, which is contrary to the standard policy of memory institutions in this area. A different path was chosen for the ARCLib project. [2] The aim is to develop an open-source LTP solution that is able to provide the required functionalities to ensure long term preservation within the environment of Czech libraries (whilst respecting all the international standards which are common in the community). While this is not a closed tool, the preservation possibilities are restricted to a pre-set group of data types (in light of the open-source nature of the tool, however, it is possible to broaden the set of data types as part of onward development). The aim of the new tool is to provide support for data preservation within the standards which are currently used at Czech libraries. These are primarily the standards in place at Národní digitální knihovna (National Digital Library), the Kramerius digital library, the ProArc production system and the repositories of DSpace. The ARCLib solution makes it possible to ensure the long-term protection of digital data at libraries of varying size and will be a freely-available alternative to accompany commercial solutions, the application of which in Central Europe is more common at large institutions such as national libraries and national archives.

Meanwhile, restriction to such major institutions is not solely based on the cost of acquiring the software tool involved. Indeed such costs are actually falling. An essential prerequisite for putting one's own policy of long term preservation into place is having sufficiently broad human resources, whereby the software tool is merely a necessary tool within a comprehensive LTP solution. The process of planning and subsequent operation of a trustworthy digital repository brings with it personnel and financial costs - for a clear description see, for example, Rosenthal (2009). Large teams remain the domain of large institutions and we cannot expect, even within the medium term, that smaller organisations such as regional libraries will have a comprehensive team of experts that would be able to cover the full range of issues involved, from the operation of hardware, through the management of content from the perspective of logical protection of stored data, to specialists that plan future steps in relation to long term preservation. Neither should we forget regular audits of the systems required for LTP, the evaluation of these and the preparation of documents for evaluation. Future users of the ARCLib solution will also have to fulfil the demands placed on evaluation. The implementation team is aware that regional and specialised libraries will not have the large team described at their disposal, but will need to ensure long term preservation all the same. As far as commercial products are concerned, they will be able to draw on the support of the supplier, but will also have to respect the policies issued by national institutions. As part of the ARCLib project, standard technical and user documentation is now accompanied by two

[2] Inspiration taken from, for example, the POWRR – Preserving Digital Objects With Restricted Resources project, conceived in a similar way - **http://commons.lib.niu.edu/handle/10843/13610**.

methodologies that provide users with sufficient knowledge of how to use the system in the right way and execute operation to ensure logical preservation with its assistance.

The methodology for the logical preservation of digital data was created first and was this year certified by the Ministry of Culture of the Czech Republic. The methodology describes the whole concept of the proposed LTP solution and explains the individual functions which make up the whole and, based on these, presents users with detailed instructions on how to use the procedures which the tool makes possible in ensuring the long term preservation of digital documents. The methodology describes in detail the structure of an archival information package and fundamental metadata sections, and the information which the system itself generates is explained here (for example, about validations, the method of version control, etc.). Instructions are also found in the methodology on how to assess the risks of stored data, how to prepare an institution that uses ARCLib for basic certification, etc. The project implementers also suppose that, once the system has been expanded, a certain community of users will develop and collectively maintain the knowledge base required for qualified decisions in the long term preservation of information content within the developed system - this should involve decision-making and recommendations of how to approach the database of formats, rules and services provided, decision-making on format migrations and chosen tools - and execute the functions required by the OAIS standard in the sphere of preservation planning. The second of the planned methodologies was also created this year, i.e. the methodology for bit-level preservation of digital data, which on the contrary focused on ways of safeguarding the "physical" preservation and cohesion of stored data using the methods described in the ARCLib tool. This methodology was submitted for certification in September 2018. It is envisaged that both of these methodologies will be regularly updated in the future to take into account changes in the tool itself and the procedures recommended within the international community.

The aim of the project is to develop a software tool and extremely detailed recommendations on how to use it so that it is also possible to use these methodological documents to carry out the basic tasks involved in long term preservation at smaller institutions with a limited number of workers. However, the scope of both methodologies goes beyond demarcation for system users alone. The general sections are the first attempt at normatively summarising recommendations and good practice for the processes of long term preservation of digital data at libraries in the Czech Republic. They can be used to plan activities for other LTP systems or by digital data producers and such as they are established with regard to the need for their long-term protection in the future. The universal applicability of the methodologies is also based on the involvement of all relevant participants that share in the digitisation and management of digital documents at libraries. Knihovna Akademie věd ČR, v.v.i. (Library of the Czech Academy of Sciences), Národní knihovna ČR (National Library of the Czech Republic), Moravská zemská knihovna (Moravian Library) and Masarykova univerzita (Masaryk University) are all participating in the project. The involvement of the National Library of the Czech Republic guarantees interoperability with the standards of the National Digital Library. Such interoperability will not only be at a general level: it will also be ensured that the AIP created according to National Digital Library regulations can also be preserved in ARCLib. Cooperation between both archiving solutions will significantly increase the level of protection of stored digital data. The openness of the solution, after minor modifications and developments of the system, particularly in the sphere of data schema makes it possible to engage other memory institutions or other producers of digital born documents.

The ARCLib tool is designed for the management and preservation of digitised and digital born documents. Procedures for the processing of both types have been prepared. The prerequisite is that access for a specific producer and its data format is invariably regulated. It is conditional on adherence to the master format for the storage of metadata, which in the case of ARCLib is the METS standard. ARCLib is not a production system and is not capable of creating submission information packages: it merely preserves them and creates archive alternatives from them.

Description of the system

ARCLib is a system for the logical and bit-level preservation of digital data that has been designed in line with the requirements inferred from the ČSN ISO 14721 (OAIS) standard. It uses tools that are already in existence, such as ProArc and Archivematica, to the maximum possible extent, mainly for the creation of SIP packages. It validates the prepared SIPs, converts them to archival packages (AIP) and preserves them in accordance with OAIS. When identifying individual modules, this paper is based on the actual naming of modules established during the definition of functional requirements and is still maintained. *Figure 1* illustrates the clearly-arranged schema of the ARCLib system and its modules and its relationship to the outside world. The modules of the system correspond to the modules specified in the ČSN ISO 14721 (OAIS) standard: however, they are adapted to meet the character of the system, i.e. to the fact that this is a dark archive that is not intended for end users and the fact that the system envisages the input of SIP already having been processed in other systems.

ARCLib does not have means of displaying archived data (image servers, browsers, etc.). ARCLib is used by the managers of archive digital data and data is used, following export, by dissemination systems and, where appropriate, by other digital library systems (DAM systems). The updating of AIP and the generation of new versions of AIP proceeds in large part through data editing in external systems (ProArc, DSpace) and subsequent re-ingest in ARCLib.

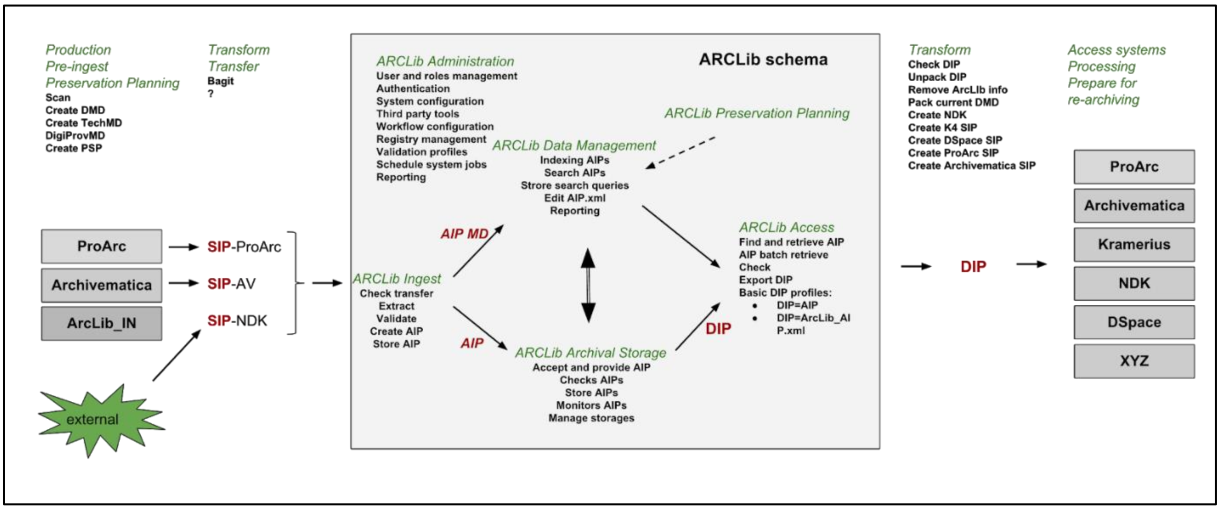


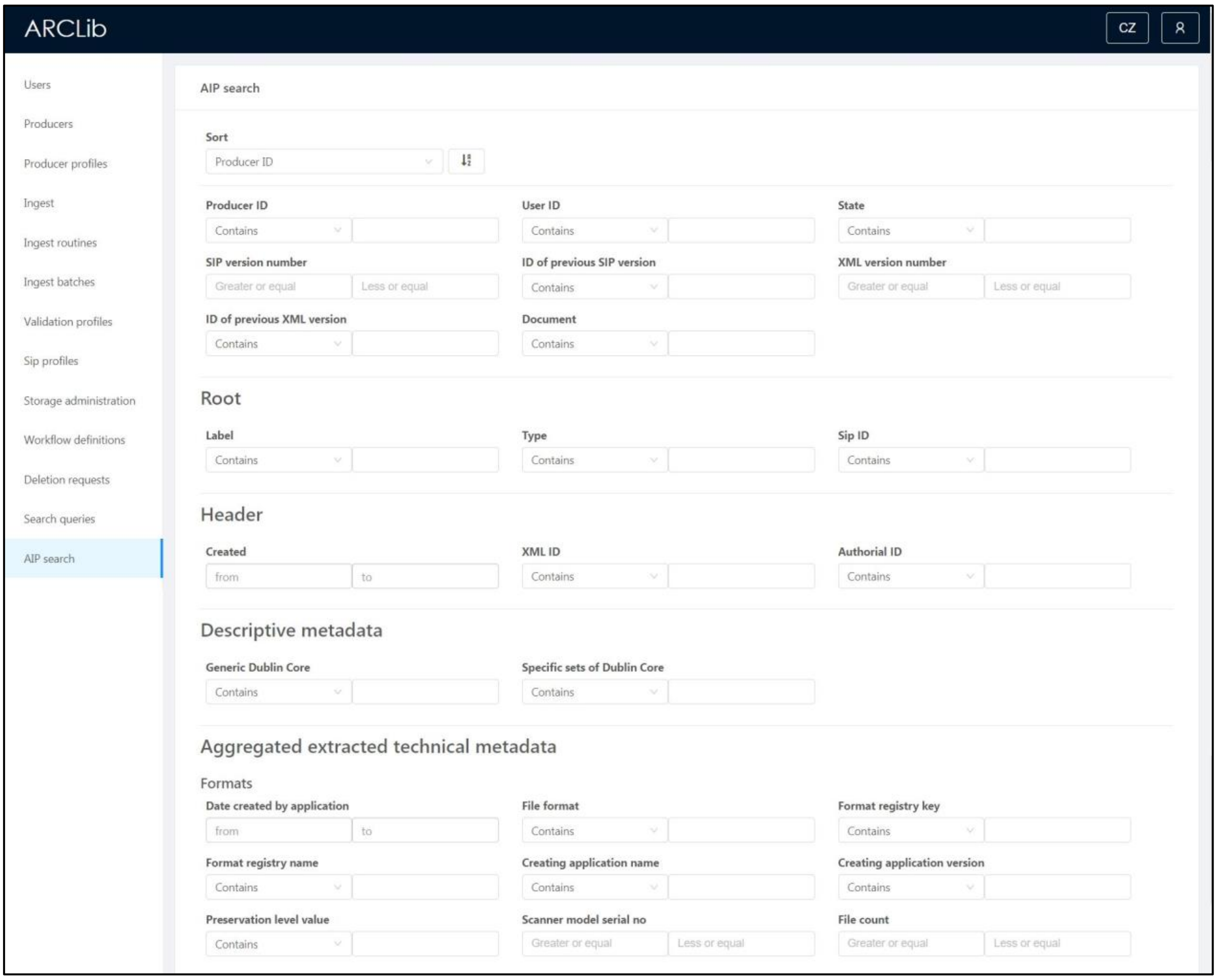
Figure 3: ARCLib schema

The ARCLib Ingest module

The ARCLib system envisages the input of data in the form of fully-fledged SIP created according to the pre-set standard employed by the institutions or external system. The functions of the module are primarily as follows: validation of input SIP according to validation templates provided by the producer, the extraction of metadata from SIP and the creation of new metadata. Metadata information is stored in a record in ARCLib AIP XML format, which is based on the METS and PREMIS international standards. The original metadata record of SIP is always preserved within the package. SIP processing proceeds according to the profile that is specific for the concerned data type of the relevant producer.

The ARCLib Ingest module is able to process the following structures of input data:

- ProArc National Digital Library monographs;
- ProArc National Digital Library periodicals;
- ProArc native monographs and periodicals;
- ProArc audio documents;
- National Digital Library periodicals and monographs;
- Archivematica DSpace;
- Archivematica General;
- National Digital Library electronic documents.



AIP search

Sort: Producer ID

Producer ID: Contains
User ID: Contains
State: Contains
SIP version number: Greater or equal / Less or equal
ID of previous SIP version: Contains
XML version number: Greater or equal / Less or equal
ID of previous XML version: Contains
Document: Contains

Root

Label: Contains
Type: Contains
Sip ID: Contains

Header

Created: from / to
XML ID: Contains
Authorial ID: Contains

Descriptive metadata

Generic Dublin Core: Contains
Specific sets of Dublin Core: Contains

Aggregated extracted technical metadata

Formats

Date created by application: from / to
File format: Contains
Format registry key: Contains
Format registry name: Contains
Creating application name: Contains
Creating application version: Contains
Preservation level value: Contains
Scanner model serial no: Greater or equal / Less or equal
File count: Greater or equal / Less or equal

Figure 2: The ARCLib Data Management module

ARCLib Data Management module

This module is primarily designed for the management of stored AIP. It contains information about the AIP stored in the system and makes it possible to search for and index such information. The module also makes it possible to browse the content of AIP and edit metadata, and provides the option of creating a new version. Tools for reporting are also part of this module.

Figure 2 shows the search interface of the system prototype currently in existence. Searching is adapted to the needs of digital data managers and is possible in relation to descriptive administrative and technical metadata. Searching is accessible via API.

The ARCLib Administration module

Administration includes a function for the configuration of workflow for the processing of Ingest, the relevant registers relating to this (the register of steps of the ingest workflow, the scripts used as part of ingest, the register of validation profiles, etc.). This module also includes the administration of the tools of third parties used within the system. It is also here that the administration of users and their roles and authentication settings is done.

The ARCLib Archival Storage module

This module is derived from the system and is approached as a separate application, which means that it can even be used outwith the ARCLib system. Functions for data management are available via REST interface. The module makes it possible to ingest and disseminate AIP, maintain information about the location of packages, check integrity, preserve operational metadata, update metadata, connect to a specific location of preservation technology, replicate data in a number of locations, back up, administer preservation technologies and media and report.

The ARCLib Access module

In light of the fact that ARCLib is conceived as a back-end application that is not intended for end users, the possibility of accessing data is restricted to the possibility of AIP export. The content of export DIP is equal to the content of AIP and is primarily intended for data producers. For this reason ARCLib does not contain any tools for forcing a policy which limits access to data.

The ARCLib Preservation Planning module

The functions of preservation planning, as they are perceived by the ČSN ISO 14721 standard, are shifted outside the system, in light of the character of the system - this primarily involves functions of an organisational and research-based nature (for example, monitoring a designated community or monitoring technology). Here the project primarily envisages the expert and methodological activity of the National Library of the Czech Republic. ARCLib itself comprises the basic tools for, in particular, work with formats.

The ARCLib system was launched in test regime on the infrastructure of the Library of the Czech Academy of Sciences in the autumn of 2018. This is the first prototype to contain all fundamental functionalities (following prototypes of individual parts). The individual elements of the system will be verified in this and, depending on the results, work will continue on modifying the system. The development process as a whole should culminate in verification of the functionality of the ARCLib tool in the form of semi-operation in the year 2020.

Conclusion

Once it has been completed, software output from the ARCLib will become a valuable tool for libraries and other institutions engaged in the long term preservation of digital data. Ideally, it will become an alternative to commercial solutions, as well as a variant which can be used by institutions that do not have the human resources or funds to operate a comprehensive solution to ensure LTP. The ARCLib system is also adapted to the needs of the Czech library environment. Nevertheless, the open-source character of the system will enable further development and potential expansion into other, related areas, particularly for other memory institutions.

The project will also produce two methodologies that focus on logical preservation and the preservation of bit-stream. These are review materials that are used both for work with the system and for general familiarisation with an issue.

References

ČSN ISO 14721, 2014. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví.

ČSN ISO 16363. *Systémy pro přenos dat a informací z kosmického prostoru - Audit a certifikace důvěryhodných digitálních úložišť.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

HUTAŘ, Jan and Marek MELICHAR. České paměťové instituce a digitální data - historický exkurz, současný stav a předpokládaný vývoj III. *Duha* [online]. **28**(2) [Accessed 25 September 2018]. ISSN 1804-4255. Available from: **http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1**

HUTAŘ, J., A. MIRANDA, E. PAVLÁSKOVÁ, Z. VAŠEK and Z. HRUŠKA, 2018. *Metodika logické ochrany digitálních dat* [online]. [Accessed 25 September 2018]. Available from: **http://hdl.handle.net/11104/0282107**

ROSENTHAL, Colin, Asger BLEKINGE-RASMUSSEN and Jan HUTAŘ, 2009. *Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER)* [online]. 1. vyd. Praha: Národní knihovna ČR. 65 p. ISBN 978-80-7050-569-4. Available from: **http://www.ndk.cz/platter-cz**

THOMAS, Lynne M., Jaime L. SCHUMACHER, Drew VANDECREEK, et al., 2014. *From Theory to Action: Good Enough Digital Preservation for Under-Resourced Cultural Heritage Institutions* [online]. Washington (DC): Institute of Museum and Library Services [Accessed 26 September 2018]. Available from: **http://hdl.handle.net/10843/13610**