# conference on grey literature and repositories

proceedings 2017

# CONFERENCE ON GREY

# LITERATURE AND REPOSITORIES

## Proceedings

2017

National Library of Technology, 2017

conference on
grey literature
and
repositories

NTK
50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Library of Technology

English conference website

(**https://nrgl.techlib.cz/conference/10th-conference-on-grey-literature-and-repositories/**)

Czech conference website

(**https://nusl.techlib.cz/konference/10-rocnik-konference/**)

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

Editor: Mgr. Hana Vyčítalová

ISSN: 2336-5021

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

## 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

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

## List of Reviewers:

RNDr. Miroslav Bartošek, CSc., Masaryk University

Ing. Lukáš Budínský, Tomas Bata University in Zlín

PhDr. Ladislav Cubr, National Library of the Czech Republic

Drs. Elly Dijk, DANS

Dr. Jan Dvořák, Charles University

Dr. Dominic Farace, Greynet

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

PhDr. Judita Matějová, Moravian Gallery in Brno

Mgr. MgA. Jakub Míšek, Masaryk University

Mgr. Lenka Němečková, Czech Technical University of Prague

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

Mgr. Pavla Rygelová, VŠB – Technical University of Ostrava

Christiane Stock, The Institute for Scientific and Technical Information

Mgr. Václav Stupka, Masaryk University

Marcus Vaska, University of Calgary

Mgr. Jan Zibner, Masaryk University

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: http://nrgl.techlib.cz/conference/conference-proceedings.

# Table of contents

**Rethinking the Role of Grey Literature in the Fourth Industrial Revolution.................... 6**

Dobrica Savić

**Digital Repository(-ies) at Charles University .................................................................19**

Jakub Řihák

**From the Dissemination of Electronic Theses and Dissertations to Their Long-term Archiving.........................................................................................................................33**

Eliška Pavlásková

**Research and Development in the Field of Research Data and Dissertations...............45**

Joachim Schöpfel, Hélène Prost, Cécile Malleret

**Attitudes of Charles University Academic Staff to Data Sharing....................................57**

Adéla Jarolímková, Barbora Drobíková, Martin Souček

**Data Deposit into the ASEP Repository............................................................................67**

Zdeňka Chmelařová, Jana Doleželová

**Institutional Rules and Policies for Sharing and Storing Research Data.......................77**

Michal Koščík

**Orphan and Out-of-Commerce works after the Amendment of the Czech Copyright Act ...........................................................................................................................................85**

Matěj Myška

**Changes in the Area of Extended Collective Management in Relation to Memory and Educational Institutions in the Light of the Czech Amended Copyright Act..................94**

Lucie Straková

**Grey abART ......................................................................................................................102**

Jiří Hůla

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: http://nrgl.techlib.cz/conference/conference-proceedings.

# RETHINKING THE ROLE OF GREY LITERATURE IN THE FOURTH INDUSTRIAL REVOLUTION

**Dobrica Savić**

**d.savic@iaea.org**

**International Atomic Energy Agency (IAEA), Vienna**

## Abstract

The world is at the dawn of a new industrial revolution that will fundamentally change the way we live and work. Many consider this the Fourth Industrial Revolution (4IR). While the First Industrial Revolution (1IR) mechanized production using water and steam power, the second one brought mass production using electric power, and the third one was characterized by automation and digitization, mainly using electronics and information technology.

The 4IR is building upon the third one, but the difference, and its main contribution, is the fusion of technologies that are blurring the lines between the physical, digital, and biological worlds. This is further enhanced by the emerging progress of technology in fields such as quantum computing, machine learning, artificial intelligence, robotics, virtual assistants, the Internet of Things, self-driving cars, drones, 3-D printing, nanotechnology, biotechnology, traffic and security monitoring systems, and renewable energy. This paper examines the potential impact of the emerging 4IR on grey literature (GL) and is based on analysis of the most prevalent current trends and developments in “cyber-physical systems” that connect machines, computers and people. It will examine the need to rethink the definition of GL, its creation and publication types, processing, sustainability and usability. Given the magnitude of the potential impact of the 4IR on GL, the question is what challenges the 4IR will pose to GL managers. One could assume that the acquisition of new knowledge and skills, and the revamping of existing processes and methods will be necessary. Becoming aware of this new phenomenon is only the beginning. It needs to be followed up by professional development and adequate training. Finally, the job of GL professionals will be to promote and publicize the usefulness and importance of GL, not only in their daily work, but also in research and science.

## Keywords

Grey Literature; Industrial Revolution; Information Technology; Information Management

## Introduction

The last 230 years, known as the '*industrial age*', started with the use of steam-powered machines in textile production and the introduction of the first mechanical loom in 1784. The introduction in 1870 of electrical energy, mass production and assembly lines marked the transition to the 2IR. The second half of the 20th century, brought us computers and electronics, which for many indicated the 3IR. Their massive spread was brought about by an increase in speed and functionality, along with a decrease in price and size. Machines became interconnected, were able to 'talk' to each other, and could do many jobs previously reserved only for people. For many, the introduction of these cyber-physical systems marked the beginning of a new era, the Fourth Industrial Revolution.

Although the 4IR is building upon the 3IR, the difference, and its main contribution, is the fusion of technologies that is blurring the lines between the physical, digital, and biological worlds. The 4IR already connects billions of people through powerful communication networks and smart mobile devices, offering access to an immense amount of data and information through high-speed internet access and unlimited storage. This affects our lives, our identities and the way we govern our societies, manufacture products and deliver services.

All of this is further enhanced by the emerging progress of technology in fields such as quantum computing, machine learning and artificial intelligence, robotics, virtual assistants, the Internet of Things, self-driving cars and drones, 3-D printing, nanotechnology, biotechnology, traffic and security monitoring systems, and renewable energy.

This paper examines the potential impact of the emerging 4IR on GL and it is based on analysis of the most prevalent current trends and developments in "cyber-physical systems" that connect machines, computers and people. It does that by looking into the historical content of the 4IR, the various terms used for the same concept, the basic pillars of 4IR and its overall impact on the way we manufacture products, manage companies and processes, and run our daily lives. It will examine the need to rethink the definition of GL, the creation and types of GL, processing, sustainability and usability. Given the magnitude of the potential impact, the question is what challenges the 4IR will pose to GL managers. It can only be assumed that it will demand the acquisition of new knowledge and skills, and the revamping of existing processes and methods. Becoming aware of this new phenomenon is only the beginning. It needs to be followed up by professional development and adequate training of GL users. Finally, the job of GL professionals will be to promote and publicize the usefulness and importance of GL, not only in their daily work, but also in research and information science.

In conclusion, the paper summarizes the future of GL, its volume and formats, a possible new definition refocusing on quality, intellectual property, curation and sustainability, the need for increased knowledge and visibility, and its improved relevance to our work.

## History of Industrial Revolutions

Around 230 years ago, the world progressed from the agricultural to the industrial age (IA). During the *agricultural age*, wealth came from the land and farming. With the introduction of technology, namely water mills, hydraulics, steam engines and coal, the agricultural age gave ground to a more superior industrial age that no longer depended on the land. The IA started with the use of steam-powered machines in textile production and the introduction of the first mechanical loom in 1784, which marked the birth of the factory. This became known as the *First Industrial Revolution*. Power from water ran all the machinery in mills that were placed near rivers and streams. This was a great improvement, however, limited mobility, together with the need for a steady flow of water, became a limiting factor for development. The introduction of steam engines, which used coal, was the turning point in revolutionizing the production of iron, railroads, textiles, and the printing press.

The introduction of electrical energy, mass production, conveyer belts and assembly lines, which started in 1870, marked the transition to the *Second Industrial Revolution*. Steel and petroleum became the major products that changed or enabled many other improvements and developments in transportation, construction, lightning, communication, and new materials such as plastic. The 2IR, also known as the 'Technological Revolution', lasted until the start of World War I in 1914.

The second half of the 20th century, brought us computers and electronics, which resulted in the digital automation of production using automation and IT. This, for many, indicated the *Third Industrial Revolution*. It is often called the computer or digital revolution because it was catalysed by the development of semiconductors, mainframe computing (1960s), personal computing (1970s-1980s), and the Internet (1990s). (Schwab, 2016). The introduction of industrial robots and robotics affected factories and industrial production.

It should be noted that there are some authors that do not accept the difference between the third and the fourth industrial revolutions, categorizing them both under the Third Industrial Revolution (e.g. Rifkin, J. 2011; Anderson, 2012; Dosi, 2013).

The increase in speed and functionality and the speed of computers, along with a decrease in price and size, brought us to a stage where machines became easily interconnected, 'talking' to each other, 'talking' to humans, and doing many jobs previously reserved only for people. For many, the introduction of 'Cyber-Physical Systems' (CPS) marked the beginning of a new era, the era of the *Fourth Industrial Revolution*. Robots, intelligence, automatons, the reduction of human labour and mediation via tools, appliances, machines, industrial automation and office automation are becoming widespread (Bloem et al., 2014). Highly intelligent CPS can autonomously perform end-to-end activities along the value chain.

Figure 1 visually represents the historical time-line of the industrial revolutions, listing the basic characteristic elements, while, at the same time, indicating the degree of complexity.

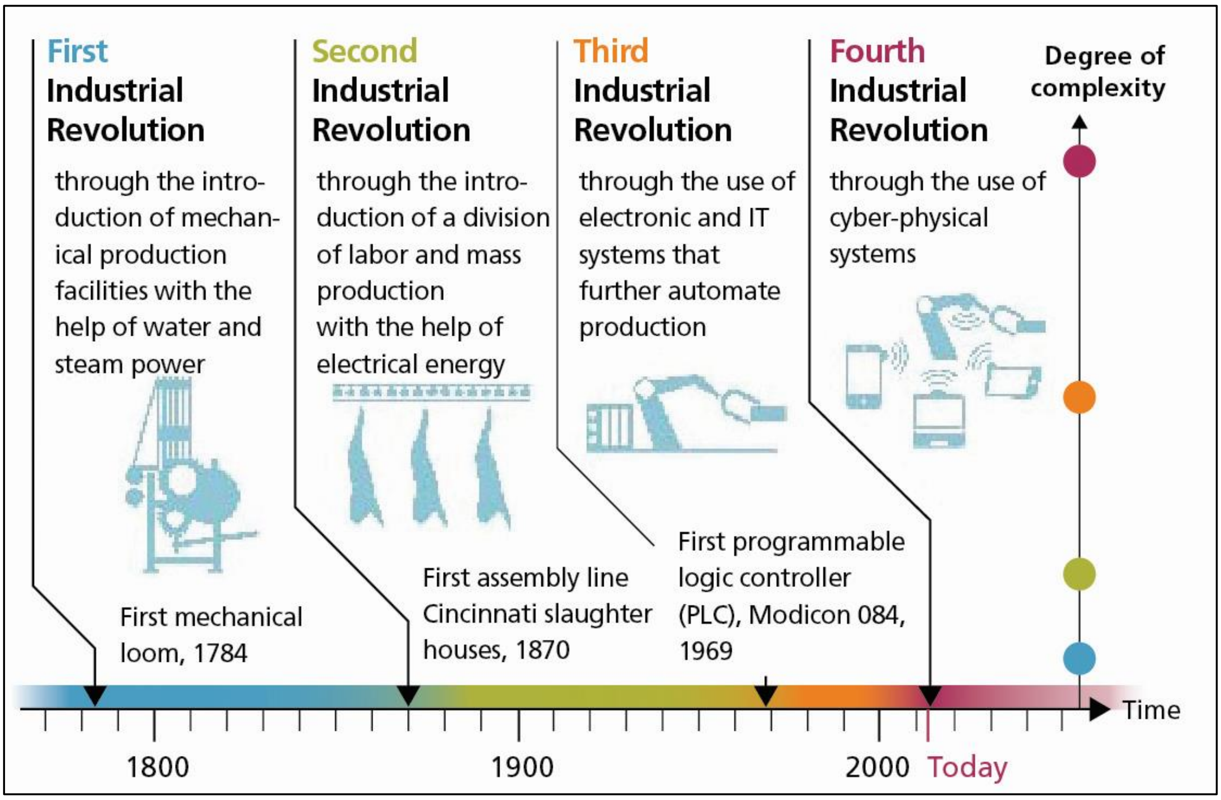

First Industrial Revolution
through the introduction of mechanical production facilities with the help of water and steam power
Second Industrial Revolution
through the introduction of a division of labor and mass production with the help of electrical energy
Third Industrial Revolution
through the use of electronic and IT systems that further automate production
Fourth Industrial Revolution
through the use of cyber-physical systems
Degree of complexity
First mechanical loom, 1784
First assembly line Cincinnati slaughter houses, 1870
First programmable logic controller (PLC), Modicon 084, 1969
Time
1800
1900
2000
Today

Figure 1: History of industrial revolutions (DFKI)

## Definition of the Fourth Industrial Revolution

There are a number of similar terms and corresponding definitions used to describe this new period of industrial development. Some of the most popular are Industry 4.0, the second machine age, the Fourth Industrial Revolution, smart factory, Industry X.0, and digital workplace.

The term *Industry 4.0* originates from Germany's 2011 Hannover Fair. It was a project of the German government to promote the computerization and innovation of manufacturing, in particular the reorganization of the global value chains. The essence of Industry 4.0 lies in a modern and modular structured factory, where physical processes are controlled by cyber physical systems that create a virtual world for making decentralized decisions.

*The Second Machine Age* indicates a stage when digital technologies (e.g. hardware, software and networks) are becoming more sophisticated and integrated and are transforming societies and the global economy. According to Erik Brynjolfsson & Andrew McAfee (2014), the world is at an inflection point where the effect of these digital technologies will manifest with 'full force' through automation and the making of 'unprecedented things'.

Professor Klaus Schwab, founder and Executive Chairman of the World Economic Forum, is the creator and the strongest proponent of studying the phenomena and using the term *Fourth Industrial Revolution*. He believes that we are at the beginning of a revolution that is fundamentally changing the way we live, work and relate to one another. A range of new technologies that are fusing the physical, digital and biological worlds characterizes this new revolution, affecting all disciplines, economies and industries, and even challenging ideas about what it means to be human. (Klaus Schwab 2016).

The *Smart Factory or Smart Manufacturing*[1] is an environment where machinery and equipment are able to improve processes through automation and self-optimization. 'Smart', because of the combination of production, information, communication technologies, sensors, motors and robotics, connecting the 'shop floor' to the 'top floor'.

Accenture[2] favors the term *Industry X.0*, the cyber-physical production system that combines communications, IT, data and physical elements. Machines "talk" to products and other machines, objects deliver decision-critical data, and information is processed and distributed in real time resulting in profound changes to the entire industrial ecosystem.

Gartner[3], another major world consulting company, talks about the *Digital Workplace* which enables new, more effective ways of working; raises employee engagement and agility; and exploits consumer-oriented styles and technologies.

## The Pillars of the Fourth Industrial Revolution

Just as there are many takes on the definition itself, there are also many opinions about the main pillars of the 4IR. Klaus Schwab talks about three groups of pillars or drivers, namely physical, digital and biological, with each one of them having related products and innovations. The World Economic Forum talks about 13 signs of the Fourth Industrial Revolution[4]. The European Union talks about 'Nine Pillars of Industry 4.0'[5], while the United Arab Emirates launched an unprecedented six-pillar plan to prepare for the Fourth Industrial Revolution[6].

Figure 2 lists some of the major drivers and pillars of the 4IR. It includes big data, artificial intelligence and machine learning, real-time analysis, robots, sensors, nanotechnology, 3D printing, Internet of Things, numerous smart devices, cyber security and visualization. The most important and fundamental of these are probably processing power, communication speed, artificial intelligence, augmented reality, and robotics.

[1] The National Institute of Standards and Technology (NIST) defines Smart Manufacturing as systems that are "fully-integrated, collaborative manufacturing systems that respond in real time to meet changing demands and conditions in the factory, in the supply network, and in customer needs."

[2] Accenture PLC is a global professional services company providing a range of strategy, consulting, digital, technology & operations services and solutions. **www.accenture.com**

[3] Gartner, Inc. is one of the world's leading research and advisory companies. The company helps business leaders across all major functions in every industry and enterprise size with the objective insights they need to make the right decisions. **www.gartner.com**

[4] **https://goo.gl/pyCK8m**

[5] **https://goo.gl/ZwzVm1**

[6] **https://goo.gl/BtzyJF**

Figure 2: The Fourth Industrial Revolution pillars

## The General Impact of the Fourth Industrial Revolution

The prediction is that the impact of the 4IR will be felt by all parts of society and through all of its activities and it will not be a small tremor. Every single activity and every industry will be affected in some way. The three main activities that will be impacted are:

- The way we manufacture products;
- The way we manage processes and companies;
- The way we run our personal lives.

*The impact of the 4IR on the way we manufacture products* is already present in many of the leading factories and production facilities. The impact can be noticed through:

- Reduced manual labour;
- Increased use of robots, sensors, artificial intelligence (AI) and machine learning;
- Automated supply chain management;
- Reduced level of stock;
- Stronger link between customer demands and production;
- Highly individualized and personalized products.

*The impact on the way processes and companies will be managed* is still not perfectly clear, although some indications are already present. They include:

- Horizontal and vertical integration through companies and entire industries;
- Removal of organizational silos, insistence on self-run and self-managed teams, building the ‘system of systems’;
- Real-time monitoring and planning;
- Introduction of ‘lean concepts’ (i.e. eliminating anything useless) ;
- Fast response to change and quick delivery using Agile;
- From reactive to predictive mode of operation and management.

*The impact of the 4IR on the way we run our personal lives* will be manifested in some, or even all, of the following ways:

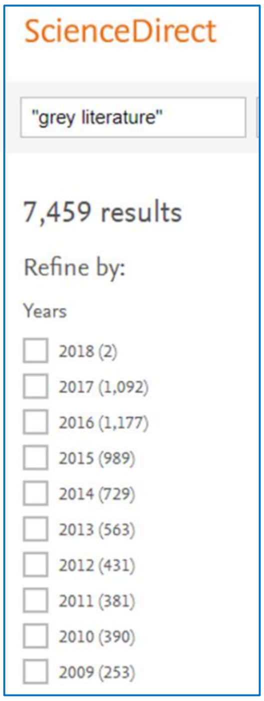
- The appearance of the almost omnipresent Internet of Things, including our households;
- The use of smart phones, need for constant communication and danger of spying; threats to our private lives through unauthorized use of security cameras and surveillance equipment;
- Unpredictable growth of society’s poor and rich parts;
- Shopping and retail industry (e.g. use of drones and already present online shopping);
- Work environment (remote/mobile work; 24/7 availability);
- Education (e.g. MOOCs, training for jobs vs. training for skills);
- The open access movement (e.g. the role of intellectual property, open science, crowd sourcing).

*"The challenges are as daunting as the opportunities are compelling. We must have a comprehensive and globally shared understanding of how technology is changing our lives and that of future generations, transforming the economic, social, ecological and cultural contexts in which we live.”* (Schwab, 2016).

## Impact of the 4IR on the Grey Literature Concept

A valid question to ask is one about the current use and the importance of GL, not as a source of information, but rather as a topic of research itself. In other words, is GL still a subject of scientific study and research? A quick look through ScienceDirect[7] using the phrase “grey literature”, results in 7,459 hits. As Figure 3 shows, the number of articles that either deal with or mention GL had a steady rise in the last 9 years, from only 253 references in 2009 to over a thousand in 2017. The two articles listed for 2018 are still in print. This is a good indication that interest is still there and that further exploration of the future and the role of GL is still valuable.

There have been many attempts to describe the concept of GL and to assign it a proper definition. The results achieved while doing this tell us that GL is much easier to *describe* than to *define* (Schöpfel, 2010).



"grey literature"

7,459 results

Refine by:

Years

- □ 2018 (2)
- □ 2017 (1,092)
- □ 2016 (1,177)
- □ 2015 (989)
- □ 2014 (729)
- □ 2013 (563)
- □ 2012 (431)
- □ 2011 (381)
- □ 2010 (390)
- □ 2009 (253)

Figure 1: ScienceDirect search results

[7] http://www.sciencedirect.com/

The 12th International Conference on Grey Literature (GL12), held in Prague in 2010, came up with the following definition:

> *"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, D. and Schöpfel, J., 2010).

Thanks to the hard work of the Prague definition authors, Dr. Farace and Dr. Schöpfel, in promoting grey literature and related research, and to the work done by GreyNet International[8], this definition is most widely accepted and followed.

Another interesting attempt to add an additional 'modern' twist to the definition of GL was to look at it from the perspective of traditional publishing, which usually goes through a peer-review process. Accordingly, GL is regarded as ***"****the diverse and heterogeneous body of material that is made public outside, and not subject to, traditional academic peer-review processes****"***. (Adams at al. 2016).

Although this definition brings into focus an interesting aspect of GL, it is very limiting, especially taking into consideration new challenges brought about by the IR.

The current concept of GL, as stated in the Prague definition, still has some challenges, especially from the 4IR perspective. The main challenges relate to multiple types of originators; humans and machines, volume and type, and the speed of GL creation. Therefore, the focus of the GL definition needs to shift more to quality, intellectual property, curation and sustainability. In its current form, the definition risks becoming obsolete due to its inability to differentiate GL from other types of documents.

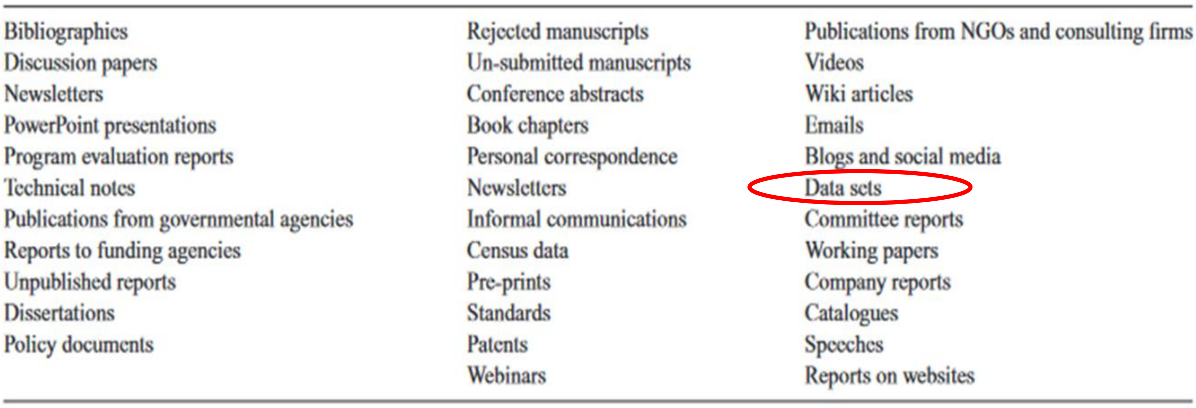
A proposed new definition, which might help meet some of the above-mentioned challenges, regards GL as *any recorded, referable and sustainable data or information resource of current or future value, made publically available without a traditional peer-review process.*

## Impact of the 4IR on Grey Literature Types

Let us examine just one of the facets of GL – its multitude of types and formats. Even a quick look at papers written about GL dealing with various formats and types, suggests a great variety. Figure 4 is a short list of possible types. However, a more complete list is available at the GreyNet International website[9]. It lists over 150 document types specific to GL.

[8] **http://www.greynet.org/**

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



Bibliographics
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

In order to illustrate the challenges already faced by GL, or that could bel faced with the progression of the 4IR, we will examine only one GL type, namely ‘data set’. This type typically includes a tremendous amount of data and information coming from the Internet of Things (IoT), the Internet of Everything (IoE), the Industrial Internet of Things (IIoT), Machine to Machine communication (M2M), self-driven cars, robots, sensors, security systems, and surveillance cameras. Estimates for the number of connected devices vary by billions. Gartner says some 20 billion 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. Such a huge number of devices, generating tons of data, mostly in an unstructured form, represents a considerable challenge for GL researchers, practitioners and managers.

## Impact of the 4IR on Grey Literature Processing

Wayne Balta, Vice President of the IBM Corporation, in his presentation regarding IBM’s concept of ‘smarter planet’ and the role of big data and sustainability (Balta, 2014), talks about three defining attributes that arise from the foundation of data. According to him, the world is becoming:

- **Instrumented** (ability to measure, sense, and see the exact condition of everything);
- **Interconnected** (people, systems and objects can communicate and interact with each other );
- **Intelligent** (we can respond to changes quickly and accurately, and get better results by predicting and optimizing for future events).

As pointed out by John Naisbitt[10], “We have for the first time an economy based on a key resource [Information] that is not only renewable, but self-generating. Running out of it is not a problem, but drowning in it is”. He went further to stress that, “We are drowning in information but starved for knowledge“. Following on Naisbitt’s thoughts, Wayne Balta developed a system of Four Vs of big data, which is important as well in understanding the role of GL.

[10] **https://en.wikipedia.org/wiki/John_Naisbitt**

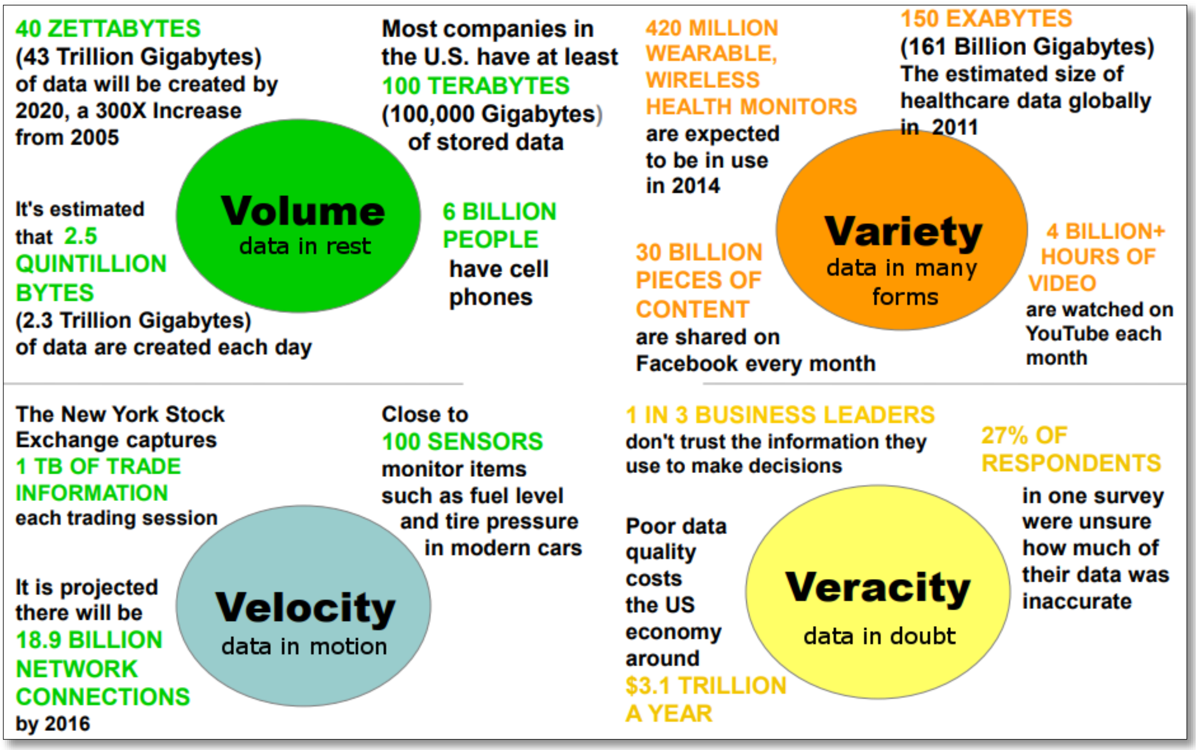

40 ZETTABYTES
(43 Trillion Gigabytes)
of data will be created by 2020, a 300X Increase from 2005
Most companies in the U.S. have at least 100 TERABYTES (100,000 Gigabytes) of stored data
It's estimated that 2.5 QUINTILLION BYTES (2.3 Trillion Gigabytes) of data are created each day
Volume
data in rest
6 BILLION PEOPLE have cell phones
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS are expected to be in use in 2014
150 EXABYTES (161 Billion Gigabytes) The estimated size of healthcare data globally in 2011
Variety
data in many forms
30 BILLION PIECES OF CONTENT are shared on Facebook every month
4 BILLION+ HOURS OF VIDEO are watched on YouTube each month
The New York Stock Exchange captures 1 TB OF TRADE INFORMATION each trading session
Close to 100 SENSORS monitor items such as fuel level and tire pressure in modern cars
It is projected there will be 18.9 BILLION NETWORK CONNECTIONS by 2016
Velocity
data in motion
1 IN 3 BUSINESS LEADERS don't trust the information they use to make decisions
27% OF RESPONDENTS in one survey were unsure how much of their data was inaccurate
Poor data quality costs the US economy around $3.1 TRILLION A YEAR
Veracity
data in doubt

Figure 3: Big data (Source IBM, Balta, 2014)

## Impact of the 4IR on Grey Literature Sustainability

The above-mentioned four Vs are also important for the long-term sustainability of GL. The Oxford dictionary defines sustainability as “the ability to be maintained at a certain rate or level.”[11] However, the most famous definition comes from the Brundtland Report (1992) that states “Development that meets the needs of the present without compromising the ability of future generations to meet their own needs.”

Sustainability of GL can be examined from three main aspects:

- **Environmental/technical**
  - Long-term preservation; organization and management; operability;
- **Economic/Financial**
  - Level and duration of support; Return on Investment (ROI); future value;
- **Social/Organizational**
  - Audience; information ownership & governance; freedom of access to information.

Each of the aspects mentioned here represents, by itself, a research topic. For this paper, it should be sufficient to note that sustainability represents the biggest challenge to the existence and future use of grey literature. Without functional sustainability, there will hardly be future for GL.

[11] https://goo.gl/OAW1JT

## Impact of 4IR on Grey Literature Usability

Closely connected to sustainability is GL usability. Designing the means, tools and methodologies for the future use of GL could become a breaking point for further industrial and social interest and in investing additional efforts to secure, process and maintain GL repositories. If its future usability cannot be guaranteed, there will not be much concentrated effort to do anything with it the present. Therefore, the question of usability needs to be examined from the following angles:

- **Tools for analysis**
  - Old vs. new tools and technology; different software functionality, concepts, expectations; dynamic vs. static information and documents;
- **Visualization**
  - 2-D and 3-D; virtual and augmented reality; requirement levels and technical skills;
- **Intellectual property**
  - Over protectionism; open access and open science; doubts about IP helping development, health, innovation;
- **Privacy**
  - Protection of sensitive personal information; CCTV cameras in public; social media photos.

Tools for future processing, analysis and presentation of GL, especially data and data sets, are a breaking point for its long-term sustainability and usability. However, intellectual property and rising concerns regarding privacy protection could also become major determining factors for the future of GL.

## 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 GL. Based on the review of the 4IR and the related developments already in place, it can be concluded that GL will not disappear in the future, that its volume will probably experience exponential growth, and that the number of GL types will increase.

Taking into consideration the volume and speed of GL creation, there seems to be a need to revisit the old definition of GL by refocusing on quality, intellectual property, curation, sustainability and usability. The most important, and probably the most critical step, is to differentiate GL from other document types so that proper attention can be focused on relevant GL issues and solutions.

In order to increase knowledge, visibility and relevance of GL, more work needs to be done on theoretical research and practical applications; on the development of proper training courses and tutorials; on establishing cooperation with data and information specialists, librarians and archivists; on promotion; and on efforts to demonstrate the value of properly managed GL collections.

## References

ADAMS, Richard J., Palie SMART a Anne SIGISMUND HUFF, 2016. Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies. *International Journal of Management Reviews* [online]. **10**(4), 432 - 454 [Accessed 16 September 2017]. Available from: **http://onlinelibrary.wiley.com/doi/10.1111/ijmr.12102/full**

ANDERSON, Chris, 2012. *Makers: the new industrial revolution*. Random House. ISBN 978-030-7720-962.

BALTA, Wayne, 2014. IBM, Big Data, and Sustainability. In: *Wharton Initiative for Global Environmental Leadership* [online]. [Accessed 16 September 2017]. Available from: **https://igel.wharton.upenn.edu/wp-content/uploads/2013/11/Wayne-Balta.pdf**

BLOEM, Jaap, Menno VAN DOORN, Sander DUIVESTEIN, David EXCOFFIER, René MAAS a Erik VAN OMMEREN, 2014. *The Fourth Industrial Revolution: Things to Tighten the Link Between IT and OT* [online]. Sogeti VINT [Accessed 16 September 2017]. Available from: **https://www.fr.sogeti.com/globalassets/global/downloads/reports/vint-research-3-the-fourth-industrial-revolution**

*Brundtland Report of the World Commission on Environment and Development: Our Common Future,* 1992 [online]. [Accessed 16 September 2017]. Available from: **http://www.un-documents.net/our-common-future.pdf**

DOSI, Giovanni, GALAMBOS, Louis, ed, 2013. *The Third Industrial Revolution in Global Business* [online]. Cambridge: Cambridge University Press [Accessed 16 September 2017]. ISBN 9781139236706. Available from: **https://doi.org/10.1017/CBO9781139236706**

FARACE, Dominic a Joachim SCHÖPFEL, 2010. *Grey literature in library and information studies* [online]. New York: De Gruyter Saur [Accessed 16 September 2017]. ISBN 978-3-598-44149-3. Available from: **https://doi.org/10.1017/CBO9781139236706**

RIFKIN, Jeremy, 2011. *The Third Industrial Revolution: How Lateral Power is Transforming Energy, the Economy, and the World* [online]. St. Martin's Press [cit. 2017-11-16].

SCHÖPFEL, Joachim, 2016. Towards a Prague Definition of Grey Literature. In: *Twelfth International Conference on Grey Literature: Transparency in Grey Literature* [online]. GreyNet, p. 11-26 [Accessed 16 September 2017]. Available from: **https://goo.gl/Jr2Fg1**

SCHWAB, Klaus., 2016 *The Fourth Industrial Revolution*. Penguin Random House. ISBN 978-1-944835-00-2.

The Nine Pillars of Industry, 2017. *Together for manufacturing* [online]. LCR 4.0 [Accessed 16 September 2017]. Available from: **http://lcr4.uk/2017/01/19/nine-pillars-industry-4-0/**

UAE launches unprecedented six-pillar plan to prepare for Fourth-Industrial-Revolution, 2016. *United Arab Emirates: The Cabinet* [online]. Ministry of Cabinet Affairs & The Future, 2017 [Accessed 16 September 2017]. Available from: **https://uaecabinet.ae/en/details/news/uae-launches-unprecedented-six-pillar-plan-to-prepare-for-fourth-industrial-revolution**

13 signs the fourth industrial revolution is almost here, 2015. *World Economic Forum* [online]. World Economic Forum [Accessed 16 September 2017]. Available from: **https://www.weforum.org/agenda/2015/09/13-signs-the-fourth-industrial-revolution-is-almost-here/**

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: http://nrgl.techlib.cz/conference/conference-proceedings.

# DIGITAL REPOSITORY(-IES) AT CHARLES UNIVERSITY “WHERE ARE WE NOW AND WHERE ARE WE HEADING?”

**Jakub Řihák**

**jakub.rihak@ruk.cuni.cz**

**Central Library, Charles University**

## Abstract

This paper describes recent activities of the Central Library of Charles University (based in Prague, Czech Republic) in regards to providing access to digitized and digital-born content, in particular theses and habilitation theses as well as additional varieties of electronic content. The paper also describes the process behind the creation of the digital repository of Charles University, current tasks and plans for the future development of this service. We attempt to answer two “simple” questions: “Where are we now?” and “Where do we want to be in the future?”

## Keywords

DSpace; Digital Repositories; Automation; Library

## Introduction

Since 2010, Charles University has had an internal regulation[1] that specifically targets the submission of theses in electronic form and makes it mandatory to submit theses to Study Information System (SIS) in the form of an electronic document. It also specifies that this electronic thesis has to be published online in the university repository. This task was previously fulfilled by ingesting theses into the Qualification works Repository system[2] created as a part of SIS.

In previous years, Charles University had provided access to most of its digitized and digital-born documents (small portion of digitized theses among them) in the DigiTool system developed by ExLibris. Even though this system is still running and is used to store and provide access to various types of digitized and digital-born materials, there was a demand for a change. The main reasons for this change were the following:

- high annual support fees
- licensing fees based on the number of digital objects stored in the repository
- demand for an open-source solution with big community support, both in the Czech Republic and abroad

The first analyses on the possibility to use a different repository system were carried out between the years 2014 and 2015. A special committee consisting of the university management, the faculty library management and a specialist from the field of librarianship and information science was established and entrusted with the task of comparing various digital depositories and digital library systems with the prospect of choosing the best possible solution to replace the expensive proprietary system with a more modern one with open source licensing.

In the meantime, it was decided that a new electronic thesis repository is needed, because the Qualification works Repository system didn’t satisfy all the requirements for interoperability between other library systems (with the exception of the library catalogue) and services, e. g. the discovery system, the National repository of Grey Literature and other international indexes, databases, information services and service providers.

It was decided that a new digital repository will be created using DSpace repository system, which is used by many Czech universities[3], has an established international community[4] and is developed as open-source software[5]. As for the annual support fees and licensing, there is no additional cost for using this system, as its support and development is community driven, with the possibility of voluntary memberships[6].

After more than six months of work, the Charles University Digital Repository[7] (CU Digital Repository) was created. It was decided that it would be used primarily as a repository for

[1] Available from: **http://www.cuni.cz/UK-3470.html**

[2] Available from: **http://is.cuni.cz/webapps/zzp/**

[3] **http://www.dspace.cz/dspace-v-cr**

[4] **http://registry.duraspace.org/registry/dspace**

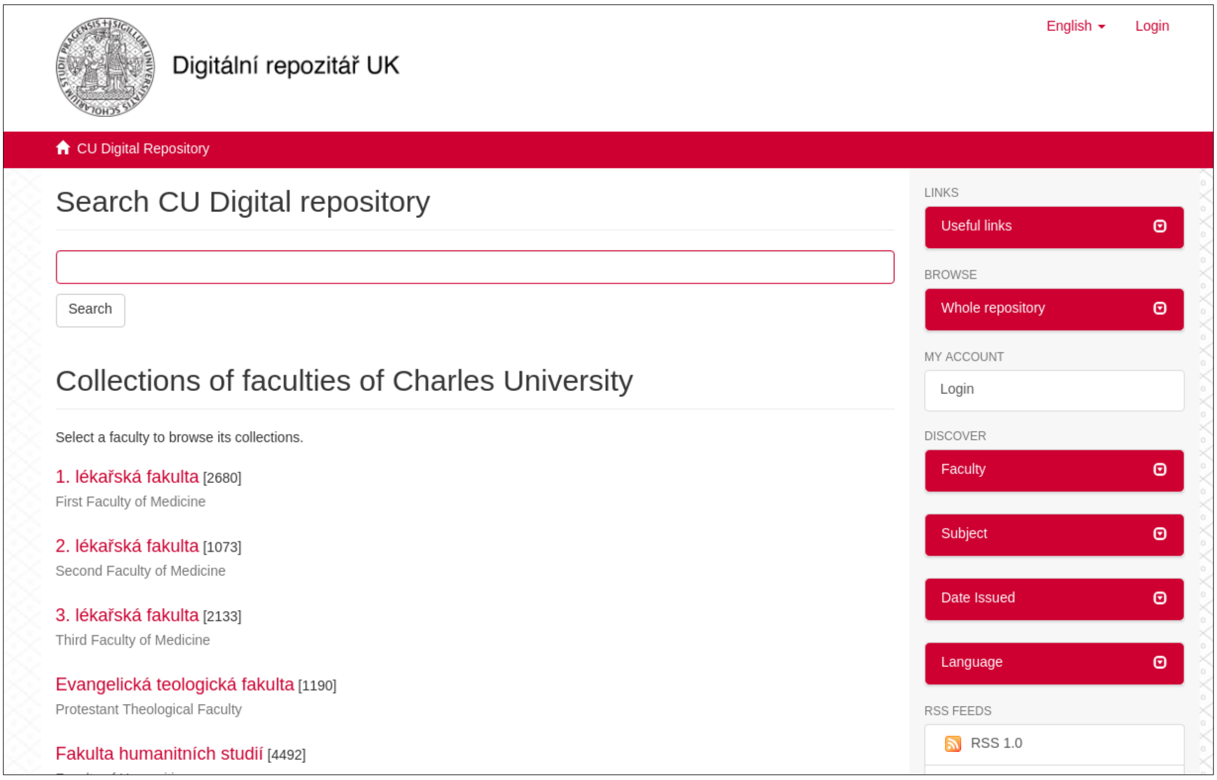
[5] **https://github.com/DSpace/DSpace**

[6] **http://duraspace.org/all_members/dspace**

[7] Available from: **https://dspace.cuni.cz**

newly defended theses due to the demand from university management and because theses offer a steady flow of new content to the repository. After nearly a year of successful operation, the Central Library now works on transferring other collections of digitized and digital-born documents from the DigiTool system and prospectively ending the use of the DigiTool system for storing and publishing digital materials.

In this article, I will try to describe the whole process by which the CU Digital Repository was created and the way it went from being an idea to a system that now stores and provides access to all publicly available theses of Charles University.



English Login

Digitální repozitář UK

CU Digital Repository

Search CU Digital repository

Search

Collections of faculties of Charles University

Select a faculty to browse its collections.

1. lékařská fakulta [2680]
First Faculty of Medicine

2. lékařská fakulta [1073]
Second Faculty of Medicine

3. lékařská fakulta [2133]
Third Faculty of Medicine

Evangelická teologická fakulta [1190]
Protestant Theological Faculty

Fakulta humanitních studií [4492]

LINKS
Useful links

BROWSE
Whole repository

MY ACCOUNT
Login

DISCOVER
Faculty
Subject
Date Issued
Language

RSS FEEDS
RSS 1.0

Figure 1: CU Digital Repository Homepage (https://dspace.cuni.cz)

## Creation of the CU Digital Repository

Works on the new digital repository began in early 2016, and the whole repository should be ready to ingest, store and publish newly defended theses from 1 January 2017. The Central Library of the Charles University wanted to implement the following principles in order to minimize the time between submission of the finalized thesis to the Study Information System and its publication in the digital repository and reduce the possibility of any human error in the ingestion workflow:

- The thesis should be ingested into DSpace directly from the Study Information System (SIS)
- There should be no unnecessary user interaction
- The ingested thesis has to have a permanent identifier and URL that won't change when the new version is ingested
- The ingested theses have to be accessible from the electronic catalogue (OPAC)
- The ingested theses have to be accessible from the discovery system

SIS does not provide an Application Programming Interface (API) of any kind, so the idea was to connect directly to the underlying database and gather all the necessary data (bibliographic metadata, thesis files and embargo information) from there.

Together with discovery system, OPAC is one of the main resources for finding an electronic thesis in Charles University, so there has to be a process that would allow adding links to digital objects in the repository to the correct record in library information system. Links in OPAC have to be permanent so that they don't change in cases where a new version of a particular thesis is ingested or transferred to another location. This could be done with the support of handle identifiers that have built-in support in the DSpace system.

A huge emphasis has also been placed on automation. With an average of 8,274 graduates in the academic year 2015-2016 (HÁJEK & BOJAR, 2017), there is the prospect of large number of theses that need to be published in a digital repository each academic year. It was also decided that the CU Digital repository will have the following structure:

→ *faculties (community level)*

→ *document (work) types (collection level)*

→ *items*

This structure is common in several Czech DSpace repositories[8], and it allows the content to be structured in a logical way that copies the organizational structure of the university and allows the user to access all existing document types of each faculty which can be also used for promotional purposes by the university faculty, as a link to the faculty's own collection and can be provided to students on the faculty's website or in other promotional materials.

[8] For example: CTU DSpace repository (**https://dspace.cvut.cz/**), Pardubice University DSpace repository (**http://dspace.upce.cz/**) or VŠB – Technical University of Ostrava DSpace repository (**https://dspace.vsb.cz/**)

## Defining workflow

After discussions with our library system administrators, it was finally decided that an existing SIS - Aleph workflow will be used to get a set of theses available for ingestion. This existing workflow is used to insert, update or delete (or rather hide) the record of the thesis bibliographic when a new thesis is available for publication. The DSpace thesis processing workflow could be inserted between those two steps with minimal changes in the existing SIS and Aleph processes. Dspace processes SIS exports, providing additional information about ingested theses to the Aleph library system. Aleph then processes the same metadata exports to insert, update or hide thesis records and the bibliographic record of each processed thesis[9] is enriched with the URL to the digital object in DSpace. The URLs and system numbers of processed theses are then passed back to SIS and stored in its database for future use. With the workflow set up in this manner, we can also ensure that all necessary data are identical in each of the connected systems as shown in Figure 2.[10]

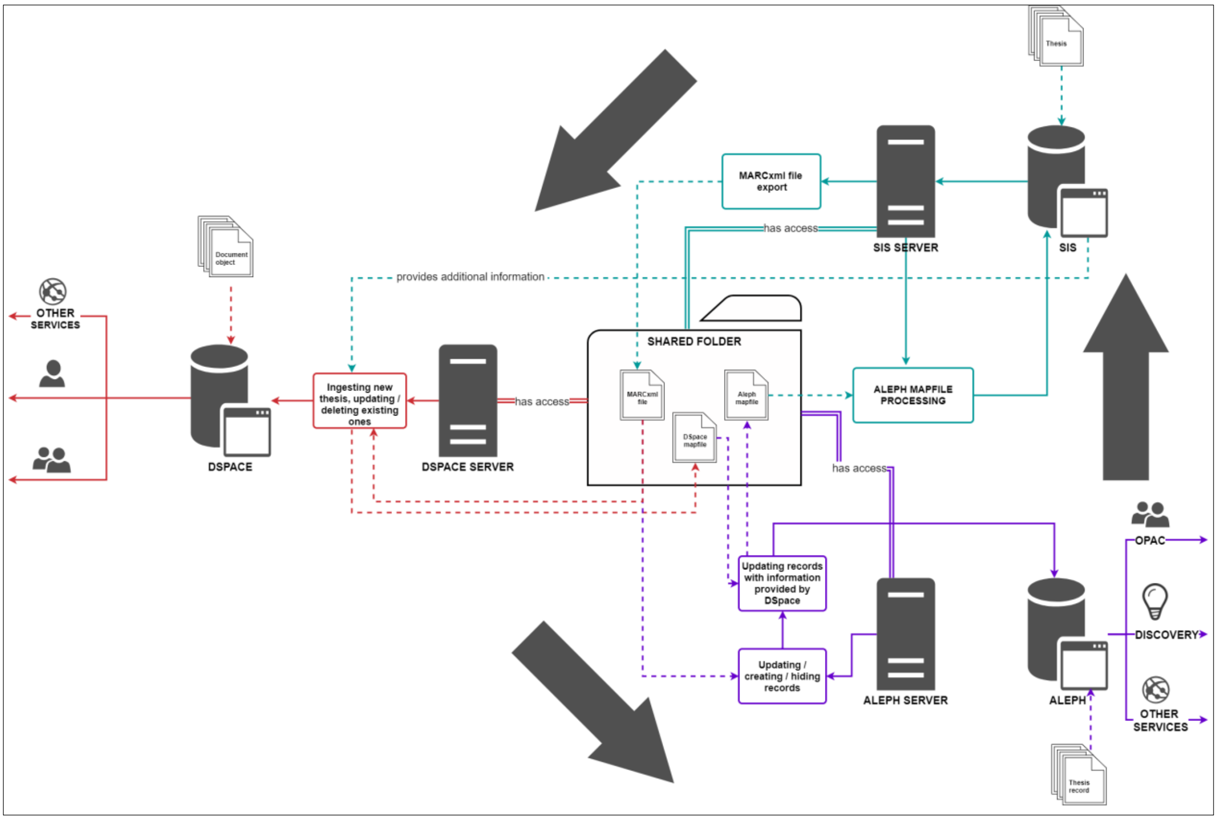


Figure 2: Thesis processing workflow diagram

[9] Of course, this does not apply to theses marked for deletion.

[10] Except for Aleph system number (unique bibliographic record identifier). This identifier can now only be added to the thesis record in DSpace after it is updated, since newly submitted theses are first processed by DSpace, not Aleph, which creates system numbers during the creation of the bibliographic record. This issue will be addressed in the future.

## Workflow automation – basic considerations

As has already been mentioned, preferably the whole thesis ingestion workflow should be automated to prevent possible human errors and to save time. There were the three following premises regarding thesis processing:

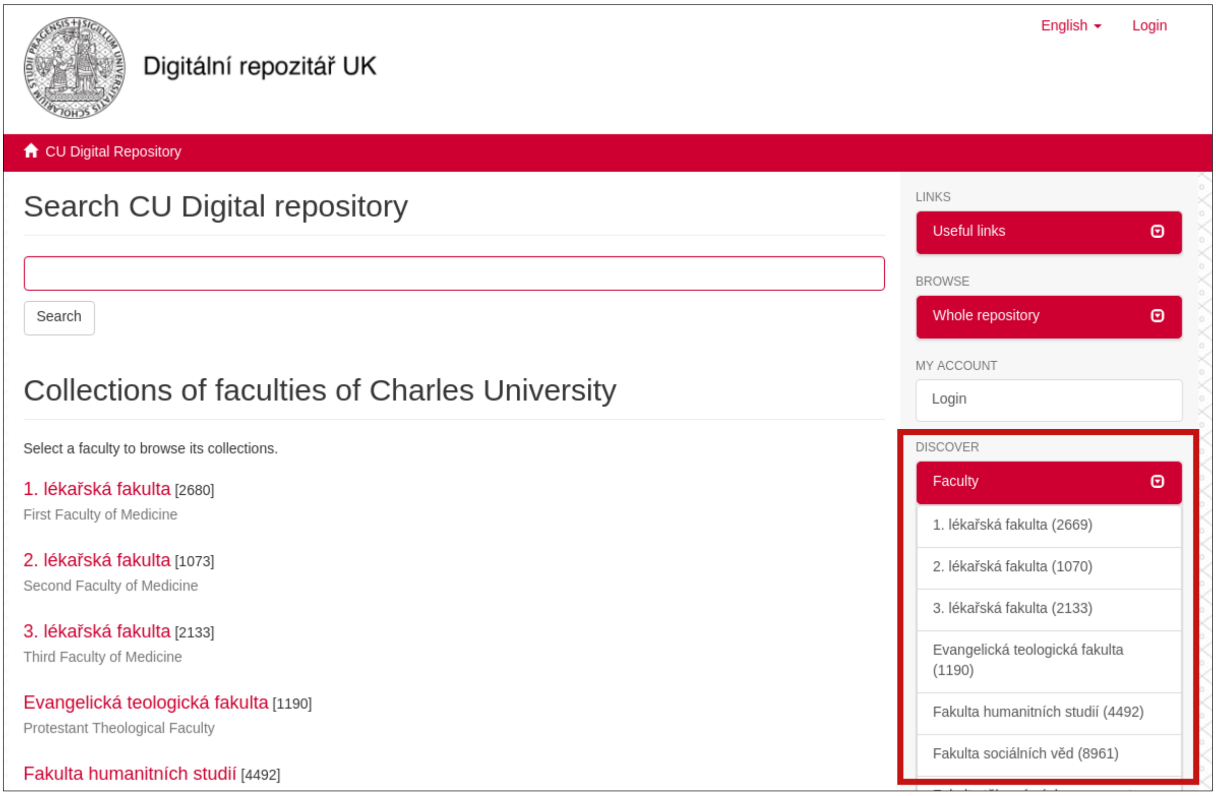
- thesis processing should take place at least once a day, but the program should check for new exports regularly several times a day
- preferably, ingestion should be done via command line tools or DSpace API
- automated ingestion should use resources that already exist if possible

For the purpose of workflow automation, the Python3 programming language is used. However, before the programming work started, it was necessary to consider which metadata we would like use to describe an electronic thesis in DSpace, which DSpace ingestion method we should use and what changes in DSpace will be necessary to ensure sufficient accessibility of the final digital object in DSpace.

## Metadata selection

The DSpace 5 system uses Dublin Core metadata format by default. There are two existing metadata schemas available[11] for item description in DSpace. Those schemas can be extended, or a new metadata schema can be created. This was the case with the CU Digital Repository, as additional metadata was required for creating custom search fields and sidebar facets that would help in making ingested theses more accessible and the whole DSpace user interface more user-friendly.

[11] Available at **https://goo.gl/BsX8hH**



English ▾ Login

Digitální repozitář UK

CU Digital Repository

Search CU Digital repository

Search

Collections of faculties of Charles University

Select a faculty to browse its collections.

1. lékařská fakulta [2680]
First Faculty of Medicine

2. lékařská fakulta [1073]
Second Faculty of Medicine

3. lékařská fakulta [2133]
Third Faculty of Medicine

Evangelická teologická fakulta [1190]
Protestant Theological Faculty

Fakulta humanitních studií [4492]

LINKS
Useful links

BROWSE
Whole repository

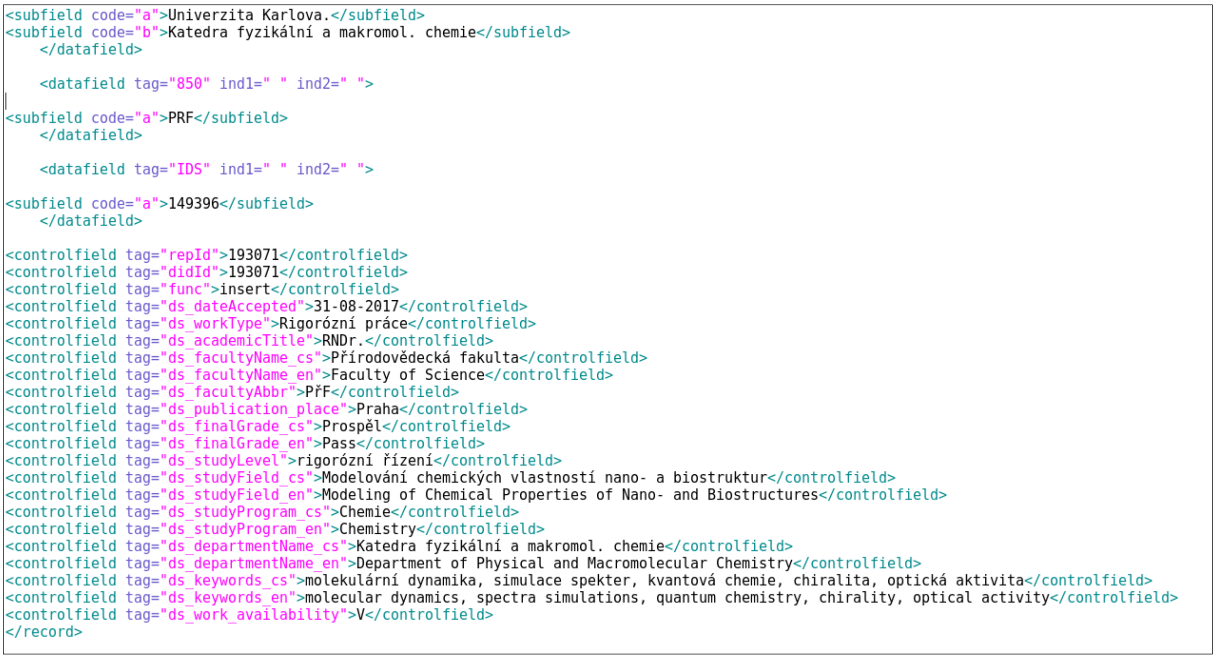
MY ACCOUNT
Login

DISCOVER
Faculty

1. lékařská fakulta (2669)
2. lékařská fakulta (1070)
3. lékařská fakulta (2133)
Evangelická teologická fakulta (1190)
Fakulta humanitních studií (4492)
Fakulta sociálních věd (8961)

Figure 3: Example of custom metadata used in sidebar facet

For custom descriptive metadata that are not part of the standard bibliographic record, control fields are used. These are not used as a data source for the document's bibliographic description during Aleph processing and are generated just for the purpose of the DSpace ingestion workflow. An example of this part of the metadata export is shown in Figure 4.

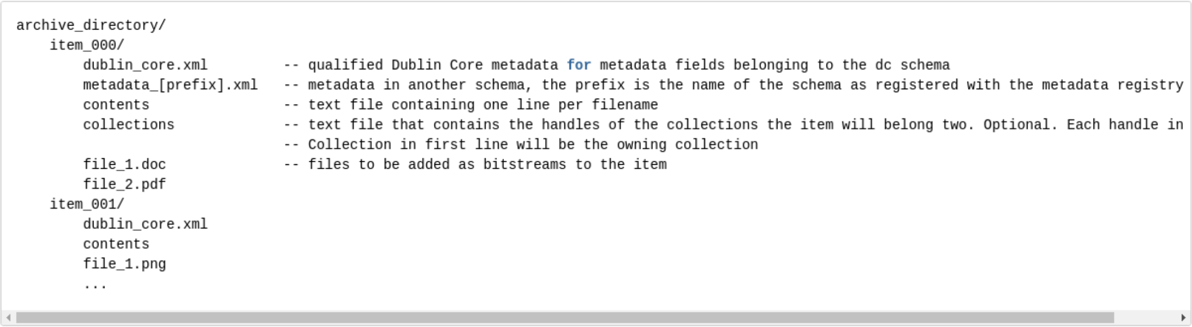


```
<subfield code="a">Univerzita Karlova.</subfield>
<subfield code="b">Katedra fyzikální a makromol. chemie</subfield>
    </datafield>

    <datafield tag="850" ind1=" " ind2=" ">

<subfield code="a">PRF</subfield>
    </datafield>

    <datafield tag="IDS" ind1=" " ind2=" ">

<subfield code="a">149396</subfield>
    </datafield>

<controlfield tag="repId">193071</controlfield>
<controlfield tag="didId">193071</controlfield>
<controlfield tag="func">insert</controlfield>
<controlfield tag="ds_dateAccepted">31-08-2017</controlfield>
<controlfield tag="ds_workType">Rigorózní práce</controlfield>
<controlfield tag="ds_academicTitle">RNDr.</controlfield>
<controlfield tag="ds_facultyName_cs">Přírodovědecká fakulta</controlfield>
<controlfield tag="ds_facultyName_en">Faculty of Science</controlfield>
<controlfield tag="ds_facultyAbbr">PřF</controlfield>
<controlfield tag="ds_publication_place">Praha</controlfield>
<controlfield tag="ds_finalGrade_cs">Prospěl</controlfield>
<controlfield tag="ds_finalGrade_en">Pass</controlfield>
<controlfield tag="ds_studyLevel">rigorózní řízení</controlfield>
<controlfield tag="ds_studyField_cs">Modelování chemických vlastností nano- a biostruktur</controlfield>
<controlfield tag="ds_studyField_en">Modeling of Chemical Properties of Nano- and Biostructures</controlfield>
<controlfield tag="ds_studyProgram_cs">Chemie</controlfield>
<controlfield tag="ds_studyProgram_en">Chemistry</controlfield>
<controlfield tag="ds_departmentName_cs">Katedra fyzikální a makromol. chemie</controlfield>
<controlfield tag="ds_departmentName_en">Department of Physical and Macromolecular Chemistry</controlfield>
<controlfield tag="ds_keywords_cs">molekulární dynamika, simulace spekter, kvantová chemie, chiralita, optická aktivita</controlfield>
<controlfield tag="ds_keywords_en">molecular dynamics, spectra simulations, quantum chemistry, chirality, optical activity</controlfield>
<controlfield tag="ds_work_availability">V</controlfield>
</record>
```

Figure 4: Custom thesis metadata in MARCxml export

## Ingestion method

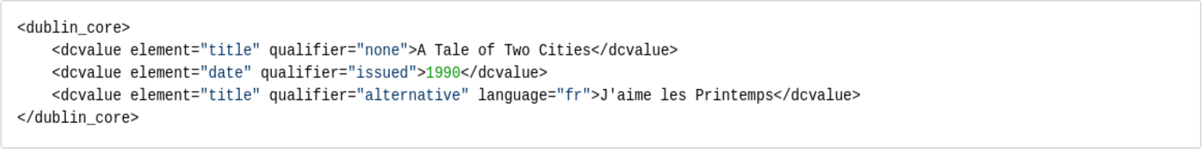
DSpace offers multiple methods of content and metadata ingestion.[12] After discussions and meetings with colleagues from other universities that are using DSpace as their repository system (mainly Tomas Bata University in Zlín and Pardubice University), it was decided that Simple Archive Format packages will be used. A Simple Archive Format package is "an archive which is a directory containing one subdirectory per item. Each item directory contains a file for the item's descriptive metadata, and the files that make up an item." (DONOHUE, 2017) The basic structure of the DSpace Simple Archive Format is shown in Figure 5. (DONOHUE, 2017)



```
archive_directory/
    item_000/
        dublin_core.xml         -- qualified Dublin Core metadata for metadata fields belonging to the dc schema
        metadata_[prefix].xml   -- metadata in another schema, the prefix is the name of the schema as registered with the metadata registry
        contents                -- text file containing one line per filename
        collections             -- text file that contains the handles of the collections the item will belong two. Optional. Each handle in
                                -- Collection in first line will be the owning collection
        file_1.doc              -- files to be added as bitstreams to the item
        file_2.pdf
    item_001/
        dublin_core.xml
        contents
        file_1.png
        ...
```

Figure 5: Simple archive format structure example

The Simple Archive Format package can be used for batch import of new items to DSpace, similarly to CSV import, but offers easy navigation in the content of each item and its descriptive metadata. Its simplistic nature is helpful in the development of an automation tool, because it allows possible errors in the package structure or content to be checked and corrected in very simple way, as can be seen in the following Figure 6, depicting a sample metadata file in the Dublin Core metadata schema.



```
<dublin_core>
    <dcvalue element="title" qualifier="none">A Tale of Two Cities</dcvalue>
    <dcvalue element="date" qualifier="issued">1990</dcvalue>
    <dcvalue element="title" qualifier="alternative" language="fr">J'aime les Printemps</dcvalue>
</dublin_core>
```

(Note the optional language tag attribute which notifies the system that the optional title is in French.)

Figure 6: Simple Archive Format metadata example

## Automation tool

The workflow automation tool was developed in 4 months. It uses the PostgreSQL database, where information on processing individual export files and theses is stored. The database is used for the purpose of determining whether or not the given export file or thesis entered the workflow in the past, to determine its processing 'direction' based on this information, and to store information on the processing state. Metadata exports are processed once a day, and each metadata export file represents a 'batch'. However, the automation tool checks for new metadata export files every 15 minutes and is able to process failed 'batches' or just individual theses for which the processing has failed.

[12] https://goo.gl/pFv9vF

The automation tool is able to gather the necessary bibliographic and other descriptive metadata and thesis files and to create a Simple Archive Format package and import it to DSpace using a standard command line importer[13].

The test of actual live data revealed an issue with an improper character escaping during metadata export file creation, resulting in the metadata export file not being processed. There were also some minor issues with displaying the additional metadata values in the DSpace user interface. However they were solved by customizing the affected parts of the DSpace user interface using a combination of XSLT, HTML and CSS. With these issues solved, the ingestion of theses to the production repository began in December 2016.

### Current state

The CU Digital Repository grows nearly every day. New these are ingested regularly and a small amount of habilitation works is already stored and published. There are currently over 90 000 items stored and available to the public. This also includes theses previously published in the Qualification works Repository that were moved to the CU Digital Repository during this year.

[13] See **https://goo.gl/j1vEph** for details.

In March 2017, the CU Digital Repository also began to receive habilitation works from individual faculties. At the beginning of February 2017, the Central Library was tasked with providing access to habilitation works according to Act no. 11/1998 Coll., on universities[14], and the CU Digital Repository had to be ready for their ingestion in one month.

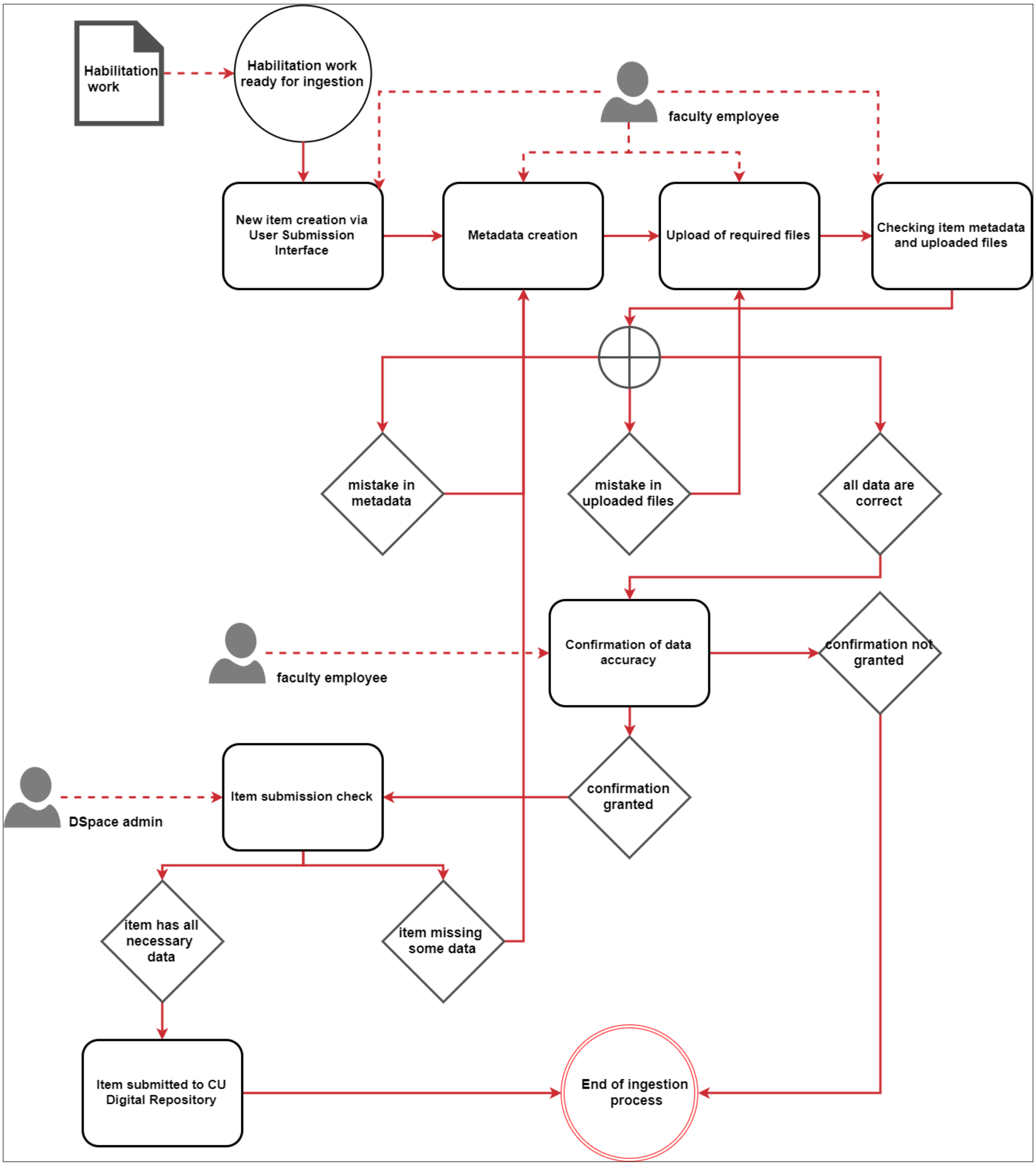


Figure 7: Habilitation works submission workflow

[14] Available at http://www.msmt.cz/vyzkum-a-vyvoj-2/zakon-c-111-1998-sb-o-vysokych-skolach.

As habilitation works are not stored in any electronic system, an ingestion workflow similar to the one used for theses could not be set up. Instead, it was decided that the internal DSpace tool - User Submission Interface[15] - will be used to gather all necessary metadata and files and publish habilitation works through the standard DSpace submission workflow.

New collections were created within the existing CU Digital Repository structure to hold habilitation works, and authorized faculty employees were given administrative rights to these collections, allowing them to submit new items and change items that have already been published. The CU Digital Repository administrators have the right to accept or reject submitted items. This provides repository administrators with a way to check submitted works and make it impossible to a submit habilitation work that does not follow the defined standards of bibliographic description or other content described in the Habilitation work submission methodology.[16] The habilitation work submission workflow is described in Figure 7. This workflow is not ideal for ingesting large amount of items, because it relies on manual work to a great extent, which could be very time-consuming when done for large quantities of documents. It is also prone to human error. However, it was designed with that in mind and offers a way to control the data quality of ingested items.

### Connecting to the National Repository of Grey Literature and OpenDOAR

The CU Digital Repository was connected to the National Repository of Grey Literature (NRGL) through the OAI-PMH protocol in April 2017. Thanks to this, Charles University is the biggest data provider for NRGL, with nearly 90,000 available records. This allows the CU Digital Repository to be more discoverable and allows Charles University to fulfil its vision of “taking active part in the development of the branches and subjects it teaches; [to be] a modern university open to the world” (Charles University, 2015) and also Strategic plan of Cantral Library of Charles University to a greater extent.

The CU Digital Repository is also registered in OpenDOAR – Directory of Open Access Repositories[17] and is indexed by Google Scholar on a regular basis. Registration in OpenDOAR is also one of the prerequisites for becoming a data provider for the OpenAIRE repository.

### Automatically generated citations

The most recent change in the CU Digital repository is the addition of the item citation to the item record view. The item citation is generated using a built-in OAI-PMH provider and Citace.com API. When the user displays an item record, a query is sent to the OAI-PMH provider, which returns the necessary data in a Dublin Core format and sends it to Citace.com. This data is coverted to the correct citation format according to the ČSN ISO 690 standard and then embedded in item record page. To implement this feature, it was necessary to create a customized OAI-PMH metadata schema that would hold all the necessary information, and it was done in cooperation with Citace.com employees.

[15] See **https://wiki.duraspace.org/display/DSDOC5x/Submission+User+Interface for details**.

[16] **https://knihovna.cuni.cz/rozcestnik/repozitare/metodika-vkladani-habilitacnich-praci-do-repozitare/**

[17] Repository record available at: **http://opendoar.org/id/3873/**

## Short-term and long-term plans

Short-term plans include:

- Enabling user authentication using Shibboleth connected to Central Authentication Service (CAS) identity provider,
  - Allowing the CU Digital Repository to dynamically assign roles to its users based on their user attributes provided by CAS and thus grant access rights to special collections in the CU Digital Repository.
- Ingestion of Open Access scientific publications from the Horizon 2020 programme,
  - fulfilling the requirements for research projects financed by the Horizon 2020 programme, according to which "each beneficiary must ensure open access to all peer-reviewed scientific publications relating to its results" (European Commission, 2017) by depositing publications in repositories. This is currently not possible on the institutional level, because the Register of Research Publications (OBD) currently being used does not provide access to actual files, and by connecting the CU Digital Repository to OBD, this access to research publications can be granted.
- Providing access to electronic books for disadvantaged students of Charles University,
  - which is in compliance with the Strategic plan of the Central Library of Charles University for the years 2015 – 2018. The Central Library is now working in close cooperation with Information and Advisory Services Centre (IASC) to provide access to these study materials and e-books via the CU Digital Repository.
- Transferring collections from the DigiTool repository,
  - collections of historical value, mainly digitized monographs, periodicals and maps, should be moved to the Kramerius digital library, which is currently being tested.
  - other collections, mainly of digital-born documents, could be moved to the CU Digital Repository. In the case of the collection of digitized theses, this transfer has already begun and is currently 80 % finished.
- Creating a digital library for historical monographs, periodicals and maps using the Kramerius digital library system.
  - The Kramerius digital library18 is, in our opinion, more suitable for providing access to digitized historical materials then DSpace and with the addition of ProArc19 software. It also has some of the long-term preservation capabilities.

[18] More details available at **https://github.com/ceskaexpedice/kramerius**

[19] More details available at **https://github.com/proarc/proarc/wiki**

Long-term plans include:

- Carrying out an analysis on the current state of the digital repositories and digital libraries used at Charles University and on the current state of publishing and preservation of digitized and digital-born documents,
    - that will serve as a foundation for the creation of a strategic plan for the development of services for providing access and the long-term preservation of digitized and digital-born content at Charles University, and should allow the Central Library to determine what the right direction of further development could be.
- Creating a strategic plan for the development of services for providing access to the digitized and digital-born content of Charles University.
    - The idea behind this strategic plan is to create a singular access point to the digitized and digital-born content of the university that can be promoted to the public more easily and guide users to the content instead of confusing them. Another advantages might be: more focused allocation of financial, technical and 'human' resources and future investments and development of any kind.
- Creating a central installation of the Kramerius digital library.
    - The Central Library would also like to create a centralized Kramerius digital library installation in which the digitization outputs of individual faculties could be published and which would serve (together with the already-implemented DSpace repository system) as a basis for this 'singular access point'.

## Conclusion

The creation of the CU Digital Repository started in June 2016 after several years of discussions. Its primary objective was to provide access to electronic theses defended from January 2017 to date. This objective was fulfilled in time thanks to the emphasis that was placed on automated processing and the focus on extending the already-existing workflow and its resources. After nearly a year of successful operation, the content of the CU Digital Repository has grown both in size and in the variety of the content provided. CU Digital Repository now also provides access to habilitation works and is prepared for the ingestion of research publications from the Registry of Research Publications (OBD) and electronic books for the disadvantaged students of Charles University. Even though errors and mistakes were made during the creation of the CU Digital Repository, we would describe its development as successful.

The CU Digital Repository is connected to the NRGL repository and OpenDOAR, which makes it possible to share the information stored in this repository with a broader audience. The repository will be continuously developed to provide better services for its users. The Central Library also aims to create a dedicated repository for digitized historical monographs, periodicals and maps. These two repositories should, in time, replace the DigiTool repository system currently being used to store the majority of digitized and digital-born materials and provide a basis for the creation of a singular access point to the digitized and digital-born materials of Charles University. In doing so, they will provide users with better access to these materials,enable the better promotion, the better allocation of financial, technical and human resources and make the long-term preservation of digitized and digital-born materials possible.

## References

DONOHUE, Tim. Importing and Exporting Items via Simple Archive Format. In: *DuraSpace Wiki: DSpace 5.x Documentation* [online]. San Francisco (CA): Atlassian, 2017 [Accessed 3 October 2017]. Available from: **https://wiki.duraspace.org/x/0QK3Ag**

*Charles University Strategic Plan 2016–2020* [online]. Prague: Charles University in Prague, 2015 [Accessed 3 October 2017]. Available from: **http://www.cuni.cz/UKEN-110-version1-charles_university_strategic_p.pdf**

*H2020 Programme: Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020* [online]. Brussels: European Commission, 2017 [Accessed 3 October 2017]. Available from: **http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf**

HÁJEK, Václav and Štěpán BOJAR, ed. *Výroční zpráva o činnosti Univerzity Karlovy v Praze za rok 2016* [online]. Praha: Univerzita Karlova, 2017 [Accessed 3 October 2017]. ISBN 978-80-246-3726-6. Available from: **http://www.cuni.cz/UK-8533-version1-vzc_2016_web.pdf**

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

# FROM THE DISSEMINATION OF ELECTRONIC THESES AND DISSERTATIONS TO THEIR LONG-TERM ARCHIVING

**Eliška Pavlásková**

**eliska.pavlaskova@ruk.cuni.cz**

**Institute of history and Archive of Charles University, Czech Republic**

## Abstract

Since 2006 it has been mandatory for Czech universities to make electronic theses and dissertations accessible on the Internet. Nevertheless, theses and dissertations are also historical archival materials of fundamental historical value, and need to be treated as such. In the year 2016, the Archive of Charles University initiated a change of the current policy on thesis submission. The emphasis was on using formats specifically suitable for long-term preservation (the format PDF/A, in particular). The objective was to collect theses in a form which may discontinue the practice of submitting printed versions and facilitate the use of electronic versions as the original archival materials. The presentation focuses on the historical development of thesis collection (including an analysis of files submitted during the years 2006-2016), submission policy description, and its implementation into the submission process.

## Keywords

Electronic Theses and Dissertations; Digital Preservation; Archiving; PDF/A; Format Policy

## Introduction

In general, theses or dissertations are materials with a significant value for the history of science and culture. As an outcome of university education and (mainly in the case of doctoral dissertations) research, these resources have a lasting value and significance and therefore constitute a heritage that should be protected and preserved for current and future generations.

Today, theses and dissertations are created and processed mostly in electronic form. Digital institutional repositories drastically change the ways in which theses are accessed, disseminated, and internally processed. Electronic versions of theses and dissertations (ETDs) include texts, databases, still and moving images, audio, graphics, software, and web pages, among a wide and growing range of formats.

Digital preservation has turned into a pressing challenge for institutions with the obligation to preserve digital objects over years. It is the collective term for actions that will ensure access to digital content in the future. The method of preservation is defined by a philosophical and practical understanding of the digital content (Digital Preservation Strategy, 2011). The Institute of the History of Charles University and Archive of Charles University is responsible for the long-term preservation of theses and dissertations in analogue (paper) form. With the advent of ETDs, their curation has become an obligation of the Archive as well. The first step in planning and executing the digital preservation strategy is the formulation and implementation of a new format policy. The policy was formed with regard for the needs of students and digital preservation and with practices internationally recognized as being the best, and it takes the recommendations of the National Archives into consideration.

## Background

Since 2006, Charles University has been accepting electronic versions of student theses and storing them in an institutional repository. By 2010, all students were required to deposit their ETDs via the web interface of the Student Information System (SIS). SIS creates a simple submission information package for ingest into the institutional repository, and it provides a mechanism for the identification and validation of deposited files. This policy is sufficient for dissemination and access to ETDs. Nevertheless, theses and dissertations are also historical archival materials. Until now, theses and dissertation have been archived in physical form (mostly on paper). Students of Charles University finish approximately 17 000 theses and dissertations every year and storage of physical materials become uneconomical and impractical. Archiving of digital data instead of paper is logical but not simple solution.

At Charles University, theses and dissertations are considered as archival materials under Act No. 499/2004 Coll., on archive and record management. There are several possible ways in which digital archival materials can be handled in compliance with the Act. Nevertheless, all of the variants demand that the archives have the ability to create submission information packages (SIPs) according to the structural and formal rules set by the National Archives. Consequently, any format policy issued by Charles University needs to take the recommendations of the National Archives into consideration.

## Development of the format policy

The format policy of Charles University is influenced by the concrete demands of Czech archival legislation. As a specialized archive under Act No. 499/2004 Coll., on archive and record management, the Archive of Charles University must follow the recommendations of the National Archives and their National Digital Archive (NDA) project. The NDA distinguishes between three groups of formats (Bernas, 2009):

- Preferred formats (e. g. plain text, XML, CSV, TIFF, Wave…)
- Accepted formats (e.g. PDF, JPEG2000, GIF…)
- Formats with a low durability (e.g. MS Word, internal formats of graphical applications…)

For still text and image documents, government resolution no. 1338 of 3 November 2008 demands the use of PDF/A-1a (ISO 19005-1 – Portable Document Format – Electronic document file format for long-term preservation), PNG (ISO/IEC 15948:2004 - Portable Network Graphics) and TIFF (Tagged Image File Format - revision 6 - Uncompressed) for use as output formats of electronic record management systems (ERMS) (Bernas, 2009).

The electronic theses and dissertations at Charles University are not managed by EMRS, so the resolution requirements are not mandatory. Nevertheless, the university should take account of them while creating its format policy.

The international community of digital preservation experts can base format assessment and policy creation on factors and categories used by leading institutions in the field. However, the British Library advises caution in the use of pre-existing documents. “Published guidelines, policies and assessments have a ripple effect and are often reused without considering the underlying evidence or the influence of unique organizational requirements. Meta assessments that make recommendations based on surveys of what other organizations do add a further level of obfuscation.” (Pennock et al, 2014)

Table 1: Format evaluation factorsTable 1 summarizes the criteria used in format evaluation and assessment by the British Library (Pennock et al, 2014), the Library of Congress (Sustainability Factors, 2017), MIT Libraries (File Formats for Long-term Access), and the National Library of the Netherlands (Rog a Van Wijk, 2009). The factors described influence the feasibility and cost of preserving information content in the face of future change in the technological environment in which users and archiving institutions operate.

| British Library | Library of Congress | MIT libraries | National Library of the Netherlands |
|---|---|---|---|
| Documentation and Guidance | Disclosure | Open, documented standard | Openness |
| Adoption and Usage | Adoption | Common usage by research community | Adoption |
| Complexity | Transparency | Standard representation (ASCII, Unicode)<br>Unencrypted<br>Uncompressed | Complexity |
|  | Self-documentation |  | Self-documentation |
| External Dependencies | External dependencies |  | Dependencies |
| Legal Issues | Impact of patents | Non-proprietary |  |
| Technical Protection Mechanisms | Technical protection mechanisms |  | Technical Protection Mechanism (DRM) |
| Development Status |  |  |  |
| Software Support |  |  |  |
| Embedded or Attached Content |  |  |  |
| Other Preservation Risks |  |  |  |
|  |  |  | Robustness |

Table 1: Format evaluation factors

The characteristics mentioned above should be used as a theoretical basis for the process of choosing preferred formats for theses and especially their annexes. Nevertheless, the selection of file formats for ETDs should be considered in the wider strategic context of Charles University, its Archive, its digital preservation needs, and its abilities. As mentioned in the Digital Preservation handbook: “At all times, the answer to digital preservation issues is not to try and “do everything”. Your strategy ought to move you towards simple and practical actions rather than trying to support more file formats than you need. (File formats and standards, 2017)”. From a practical point of view, it is crucial to reduce the complexity of data collection and storage only in necessary formats.

## Preliminary analysis

As mentioned above, Charles University has been collecting theses and dissertations for several years. ETDs were collected in PDF version 1.3 or newer, and students were allowed to deposit also an annex in the form of a single file or ZIP archive. Only version of the PDF and the occurrence of text information in the file were validated. Students were provided with no additional guidelines regarding the file creation. Deposited files represent a large set of files which allow us to run format analyses on a significant number of relevant objects.

Preliminary analyses of deposited files had two main objectives – to gain knowledge about the formats of annexes and to test identification tools. The analysis of PDFs with the main text of the theses was not relevant, as students were not required to deposit specific version of the PDF and we assume a change of policy from PDF 1.3 to PDF/A.

We tested a set of 481,396 files submitted as annexes of 2,528 theses and deposited between January 2015 and February 2016. The analysis identified 148 different formats consisting of 174 puids (PRONOM unique identifiers[1]). The distribution of formats in the set takes the form of standard "long tail" distribution. A large group of formats is represented only by the single file.

Table 2 shows the distribution of formats with an occurrence higher than 1 %. The most common format is the plain text format, which represents the source code of computer programs in most cases.

| **Format** | **Occurrence** |
|---|---|
| Plain Text File | 38.70 % |
| Portable Network Graphics | 12.28 % |
| [not able to identify] | 9.00 % |
| Hypertext Markup Language | 6.78 % |
| JavaScript file | 6.23 % |
| JPEG File Interchange Format | 4.98 % |
| Extensible Markup Language | 3.90 % |
| Java language source code file | 2.52 % |
| Extensible Hypertext Markup Language | 2.04 % |
| Apple Double Resource Fork | 1.70 % |
| ZIP Format | 1.26 % |
| Acrobat PDF 1.5 - Portable Document Format | 1.18 % |
| Windows Portable Executable | 1.04 % |

Table 2: The most frequent formats

[1] Unique identifiersof the file format developed by the National Archives (GB).

Approximately 80 % of ETDs with annexes (more than 2,000) has only one file attached as an annex. Figure 5 - Files per ETD (annexes including 10 or more files)shows the distribution of files per ETD for annexes including ten or more files. It is again “long tail” distribution. The largest sets of files are related to a small number of ETDs.

As a conclusion of the preliminary analysis, we determined that the format control in annexes must be flexible. There was a need for format guidelines in the case of standard format groups (e.g. images or video) and for a policy enabling students to deposit complex software objects with unknown characteristics.
Fido² was chosen as the main identification tool - the reasons for this decision include the demands placed on server administration and implementation requirements on the part of SIS.

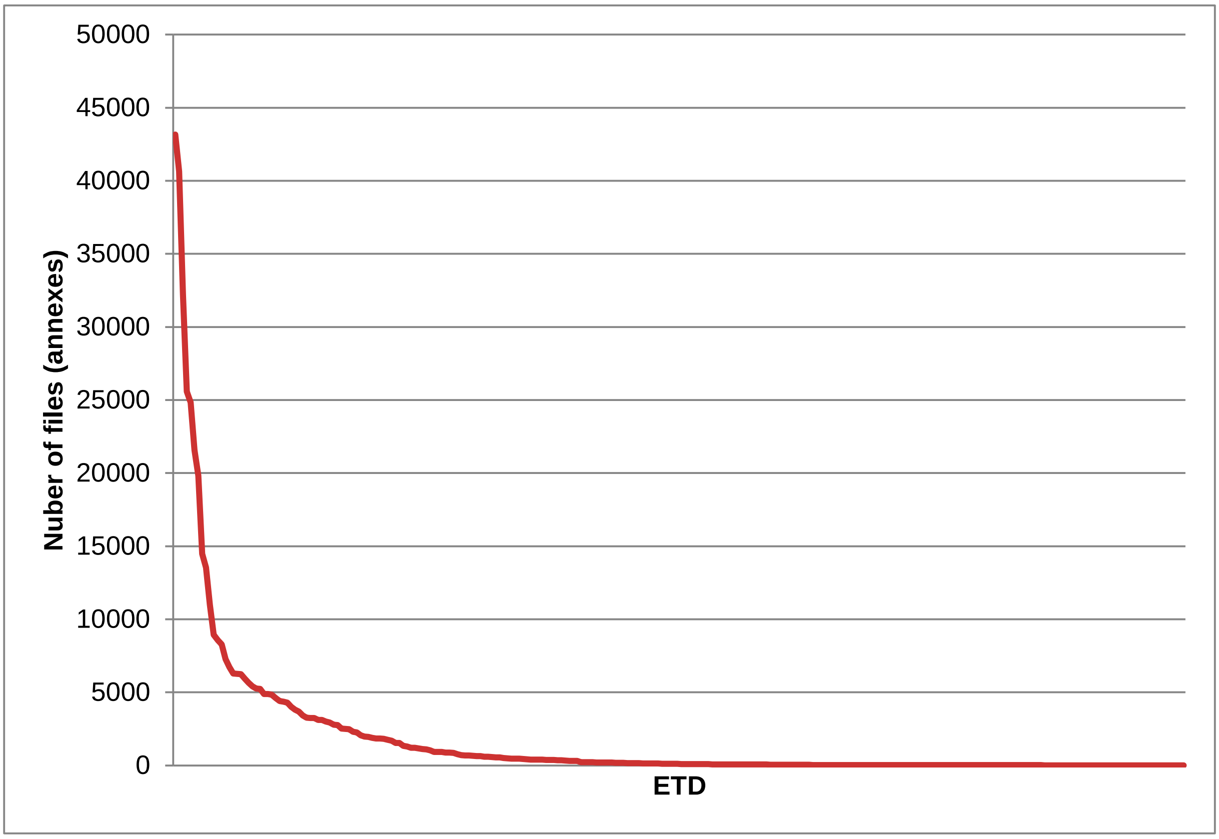


Figure 4: Files per ETD (annexes including 10 or more files)

## Structure of ETD

A thesis is usually perceived as one text document. However, from the point of view of the university administration, an ETD is a complex object consisting of several parts with different characteristics and needs in terms of the format used. The detail structure of the ETD is described by figure 2.

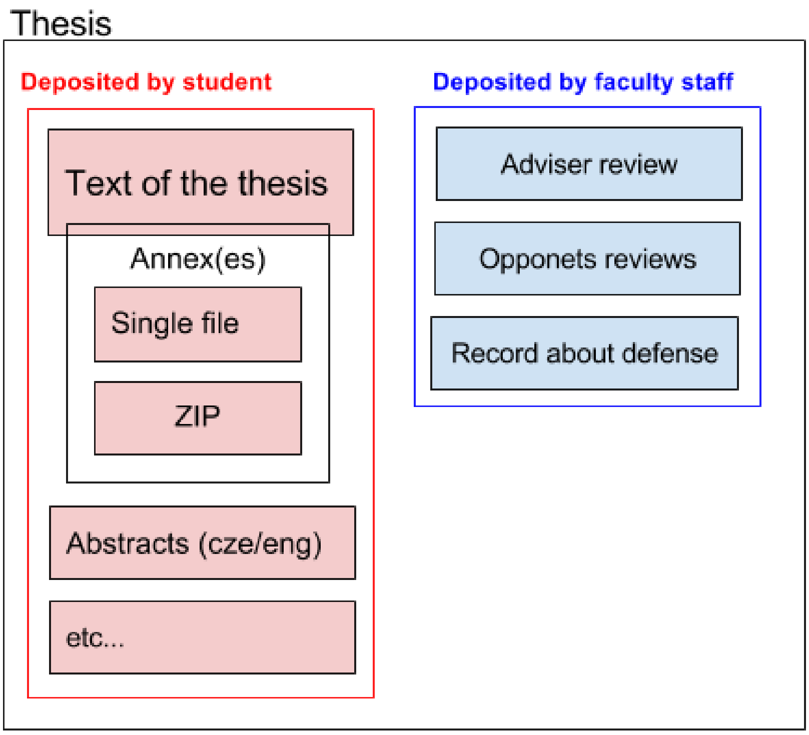


Figure 2: Structure of the ETD

**The text of the thesis** is the main part of the complex object. It is created by a student, and it has textual character. It is almost always a born-digital object, and common word editors or document preparation systems are used for its creation. There is no feasible way to accomplish a trustworthy conversion from the format used by the student to create the thesis to the format preferred for archiving. The content of the text is heterogeneous, and it is essential to check the output of the conversion manually, otherwise the content of the document can be significantly changed or even destroyed. For example, mathematical equations are frequently difficult to convert without errors and need to be checked by an expert in the field (ideally by their author).

An obvious choice for a text document is the PDF/A format. It is recommended by the National Archives as well as by the majority of leading institution in the field (see, for example, Rimkus, 2014). The word processing software that is common in the academic community (Microsoft Word, Libre and Open Office) is usually able to create a PDF/A format[2]. The Archive of Charles University contacted study departments of faculties with a request for further information on the software used. The most important response came from the Faculty of Mathematics and Physics with regard to documents created by the typesetting system TeX. The faculty provided the archive with a template for creatiing PDF/A 2u from TeX.

Given the problematic nature of PDF/A version 3 (Wheatley, 2015), only versions 1 and 2 were allowed. As regards the National Archives recommendation, level “a” was chosen and level “u” was later added on the request of the Faculty of Mathematics and Physics. Level "a" and level "u" both enforce the use of correct unicode mapping. Moreover, level "a" demands the use document structure.

[2] With exception of Pages and Word for Mac.

**Annex(es)** are a file or files created by students as an appendix to the main text. In textual form, it may be part of the main text. The inclusion of text annexes into the main text is encouraged, but in some cases, it is useful to use a separate file. The content of the annex can be a set of images, sounds, a video, tables, or computer programs. In some cases, annexes consist of research data and should be treated with consideration of their reuse. The size of the annex object can be significant.

Based on recommendations of the National Archives and after consultations with the Czech National Film Archives, the following formats were chosen for individual content types:

Text annex(es):

- PDF/A (version 1a or 2u)

Image annex(es) - the use of PDF/A (version 1a or 2u) is strongly recommended to students to for image annexes. If not possible, JPEG format can be used.
Audio annex(es):

- Waveform audio format (WAV, *.wav nebo *.wave)
- Moving Picture Experts Group Phase Audio Layer III (.mp3)

Audio-visual (video) annex(es):

- Moving Picture Experts Group Phase 2 (MPEG-2, *.vob)
- Moving Picture Experts Group Phase 4 (MPEG-4, *.mp4)

Annex(es) with character of data (tables):

- Comma-separated values (CSV, *.csv)
- Extensible Markup language (XML, *.xml) – the submission package must contain the relevant XSD or DTD
- Plain text file (*.txt)

The formats listed above have the status of an allowed format. All other formats are considered as non-approved. The exception was created with regard for the scientific or research data, software, computer application or simulations. The submission must be accompanied by an application containing a simple description of the data characteristics.

**Other documents – errata, abstracts, summaries and thesis proposition (in Figure 2 as "etc.")** are mostly textual[3] objects used for administrative purposes. From a format point of view, they have the same (even simpler) character as the text of the thesis. Therefore, PDF/A (version 1a or 2u) was chosen.

**Reviews and record about defence** – set of materials in textual form, created usually by the faculty staff. It can be born-digital, or in some cases, digitalized. The output formats of scanners and equipment used by staff need to be taken into consideration when making a format policy. PDF/A (version 1a or 2u) was chosen for the pilot stage of implementation.

[3] Charles University is considering the use of non-textual video files as abstracts in sign language.

## Pilot stage of implementation, analysis and policy changes

The pilot stage for the implementation of the new format policy started in March 2017 and ended in June 2017. During this period, we were able to collect a set of more than 3,000 ETDs and more than 5,000 files submitted as annexes.

Students were provided with an information site[4] containing provisional guidelines for the creation of PDF/A and basic guidelines for submission of annexes. An electronic help desk was created for answering specific problems with the ETD submission.

During the pilot period, we identified several areas that need to be improved or customized. We encountered problems regarding the behaviour of the validation tool, problems with PDF/A conversion in word processing software, and errors in the workflow for annex processing.

As mention above, we use the open source software VeraPDF[5] as a validation tool. VeraPDF is currently the only existing open source PDF validation tool, and it is able to validate all versions of PDF/A against a set of rules based on the PDF/A specification (ISO 19005). We also created our own version of the validation profile (a set of rules used for validation).

The second challenging area was user behaviour and the use of an information site with guidelines. During the whole pilot period, we were constantly analysing user queries in the help desk application and updating the information site with guidelines. From a format policy point of view, the most serious problems were caused by conversion in word processing or typesetting software. Approximately 11 % of queries concerned Unicode mapping in Microsoft Word (all versions). Nevertheless, a student used glyphs with no representation in Unicode only in one extremely specific case. In all other cases, an error occurred during file conversion. The conversion errors were usually independent of the software (Microsoft Word) and font versions used. The most common problem was the use of "□" as a bullet point. Approximately 3 % of problems reported by students were caused by processing transparency in images or graphs. We developed strategies to avoid this type of error and published them as a part of the guidelines. Table 3 shows the total number of users' queries.

[4] **https://www.cuni.cz/UK-7987.html**

[5] **http://verapdf.org/**

| **Contens of the query** | **Occurence** |
|---|---|
| Misunderstanding of guidelines | 66 |
| Unclear queries (student did not react to request for more information) | 33 |
| Errors of interface (timeouts, etc.) | 33 |
| LaTeX (transferred to MFF) | 32 |
| Non-relevant to format policy (e.g. requests to change the title) | 31 |
| Unicode mapping | 28 |
| Vera 1.4 - error in profile (critical failure in system, eliminated after two hours) | 18 |
| Use of Office 2007 | 14 |
| Submission form for annexes and its use | 14 |
| Use of Pages or Office for Mac | 11 |
| Processing of transparency | 10 |
| Digitalized reviews | 9 |
| Validation profile malfunction | 5 |
| Obsolete guidelines (from the website of the faculty) | 3 |
| Misuse of Adobe Acrobat | 2 |
| Indesign | 1 |
| Request for additional information | 1 |
| Personal opinion about PDF/A | 1 |
| Total number of queries | 312 |

Table 3: Users' queries analysis

Specific set of problems constitute a typesetting system TeX. We closely collaborate with the Faculty of Mathematics and Physics, where we were able to find an expert with knowledge and experience in the use of TeX.

The number of files attached as annexes to 75 ETDs totalled 5,834 files, and 53 different file formats were identified. Ninety percent of the files were attached to only two ETDs. Unfortunately, this distribution prevents us from carrying out a reliable analysis of the formats used. It is safe to assume that the authors of both ETDs use formats specific to their work. There is, therefore, no way to interpret the data correctly.

During the pilot stage, we also encountered numerous problems with documents deposited by the faculty staff (reviews and records about defence). It was decided that the forced use of

PDF/A for these types of documents will be stopped and that the practice of archiving them in analogue form as part of the student's file will be preserved.

## Conclusion

The long-term preservation of ETDs at Charles University can be done only on the condition of an existing and preserved format policy. Different approaches must be chosen for the main text of the thesis, annexes and supplementary documents. According to the practices internationally recognized as the best and according to Czech legal requirements, PDF/A is probably the only possible choice for submitted texts. Viable implementation of PDF/A collection must be based on format validation and comprehensive guidelines for students.

The format policy regarding annexes should be flexible and enable the submission of large and heterogeneous sets of files. Two possible ways of submission were facilitated - submission of files in allowed formats and submission of non-approved files supplemented with short additional information about the data deposited.

Supplementary materials such as reviews or record of defence should be part of the student file in an analogue form. Alternatively, digitalization equipment with the ability to produce PDF/A should be provided for the administrative staff of Charles University.

## References

BERNAS, Jiří. Národní digitální archiv. *Knihovna* [online]. 2009, **20**(1), p. 22-29 [Accessed 16 September 2017]. ISSN 1802-8772. Available from: **http://knihovna.nkp.cz/knihovna91/bernas.htm**.

*Digital Preservation Strategy* [online]. Wellington: Archives New Zealand Te Rua Mahara o te Kāwanatanga: National Library of New Zealand Te Puna Mātauranga o Aotearoa. 2011 [Accessed 25 September 2017]. Available from: **http://archives.govt.nz/sites/default/files/Digital_Preservation_Strategy.pdf**

File formats and standards*. Digital Preservation Handbook* [online]. Glasgow: Digital Preservation Coalition, 2017 [Accessed 23 September 2017]. Available from: **http://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards**

*File Formats for Long-term Access. MIT Libraries* [online]. Cambridge (MA): Massachusetts [Accessed 2 October 2017]. Available from: **https://libraries.mit.edu/data-management/store/formats/**

MCGUINNESS, Rebecca, Carl WILSON, Duff JOHNSON and Boris DOUBROV. VeraPDF: open source PDF/A validation through pragmatic partnership. In: *14th International Conference on Digital Preservation* [online]. [Accessed 23 September 2017]. Available from: **https://ipres2017.jp/wp-content/uploads/28Rebecca-McGuinness.pdf**

PENNOCK, Maureen, WHEATLEY, P., MAY, P. Sustainability assessments at the British Library: Formats, frameworks and findings. In: *Proceedings of the 11th International Conference on Digital Preservation*. 2014. p. 141-148. Available also from: **https://fedora.phaidra.univie.ac.at/fedora/get/o:378110/bdef:Content/get**

RIMKUS, Kyle, Thomas PADILLA, Tracy POPP and Greer MARTIN. Digital Preservation File Format Policies of ARL Member Libraries: An Analysis. *D-Lib Magazine* [online]. 2014, **20**(3/4), - [Accessed 23 September 2017]. DOI: 10.1045/march2014-rimkus. ISSN 1082-9873. Available from: **http://www.dlib.org/dlib/march14/rimkus/03rimkus.html**

ROG, Judith; VAN WIJK, Caroline. Evaluating file formats for long-term preservation. *Data Analysis and Knowledge Discovery*, 2008, 24.1: p. 83-90. Available also from: **https://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf**

Sustainability Factors. *Sustainability of Digital Formats: Planning for Library of Congress Collections* [online]. Washington: Library of Congress, 2017 [Accessed 23 September 2017]. Available from: **https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml**

WHEATLEY, Paul, Peter MAY, Maureen PENNOCK and Simon WHIBLEY*. PDF Format Preservation Assesment* [online]. Version 1.3. London: British Library, 2015 [Accessed 23 September 2017]. Available from: **http://wiki.dpconline.org/images/e/e8/PDF_Assessment_v1.3.pdf**

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: http://nrgl.techlib.cz/conference/conference-proceedings.

# RESEARCH AND DEVELOPMENT IN THE FIELD OF RESEARCH DATA AND DISSERTATIONS

## Joachim Schöpfel

**joachim.schopfel@univ-lille3.fr**

**GERiiCO laboratory, University of Lille SHS, France**

## Hélène Prost

**helene.prost@inist.fr**

**INIST (CNRS), France**

## Cécile Malleret

**cecile.malleret@univ-lille3.fr**

**Academic library, University of Lille SHS, France**

## Abstract

The paper presents the research project D4Humanities conducted by the GERIICO laboratory at the University of Lille in the field of research data management (RDM). In particular, it describes the development of a local workflow for the submission of research data related to PhD dissertations and the connection to the national RDM infrastructure Huma-Num (deposit, preservation and dissemination of research data via the NAKALA service), along with the RDM training program for PhD students provided by the Graduate School in Social Sciences and Humanities at the University of Lille.

## Keywords
Research Data; Electronic Theses and Dissertations; Data Management; Data Sharing; Data Literacy; PhD Training Program; Social Sciences and Humanities

## Introduction

Research data management (RDM) has been described as a "wicked problem" without easy answers, perhaps even insoluble, at least temporarily (Awre et al., 2015). However, each research performing institution must define its own RDM strategy in order to provide optimal work conditions for its faculty and scientists. So even if (or just because) there may be no "best model" for RDM, institutions can learn from each other's experiences and successful initiatives.

Two years ago, at the 2015 Conference on Grey Literature and Repositories in Prague, we presented empirical results on research data in electronic theses and dissertations (ETDs) and on RDM behaviours and needs of scientists and PhD students on the social sciences and humanities (SSH) campus of the University of Lille (Schöpfel et al. 2015). The basic assumption of our research was (and still is) the observation that research results produced by PhD students can contribute to data-intensive scientific discovery (Schöpfel et al., 2014). They are "hidden treasures" (Prost et al. 2015); however, a few issues must be addressed in order to make them visible, available and, moreover, reusable for scientific research.

In 2015, we published an institutional approach to RDM in the field of PhD dissertations as a White Paper (Chaudiron et al. 2015), promoting five leading principles for the development of campus-based research data support services (see Schöpfel et al. 2015, figure 5). After validation by our research department, we started to implement this approach on the campus, together with the SSH graduate school, the GERiiCO research laboratory[6], the academic library and the ANRT service (National Centre for the Reproduction of PhD Dissertations). The implementation will take three years (2015-2018); one part of the implementation was integrated in a research project called *D4Humanities*[7] and received funding from the Regional Council (Conseil Régional Hauts-de-France) and the European Institute of Social Sciences and Humanities in Lille (MESHS).

Where are we now? What have we learned? Our paper will present the actual advancement of the implementation process, in the particular context of our institution and country, and it will make some statements and recommendations for similar initiatives.

## The context

To foster uptake and increase efficiency and outcomes, the design and implementation of such a program must evaluate the specific conditions of the immediate and wider environment,

[6] Information and communication sciences, see **http://geriico.recherche.univ-lille3.fr/**

[7] Deposit of Dissertation Data in Social Sciences and Humanities – A Project in Digital Humanities, see **http://d4h.meshs.fr/**

including the expressed needs of the community being served. Here are the essential context features for the Lille program.

**National context**

- Since 2006, the French universities have a centralized system for the deposit, indexing and preservation of ETDs called STAR[8]. The ETDs are reported in the French academic union catalogue SUDOC and on the platform for PhD dissertations[9]. The author can opt for open access via an institutional or other academic open repository.
- In 2018, the digital deposit of PhD dissertations on STAR will become mandatory for all universities, faculties and departments.
- STAR allows the deposit of supplementary files including datasets but does not provide specific tools for their curation and publishing.
- With funding from the CNRS (National Centre for Scientific Research) and some universities, the consortium Huma-Num is developing an infrastructure for the SSH research communities. This infrastructure includes a platform for the deposit, curation, recording, publishing and preservation of datasets (NAKALA).
- The CNRS is also developing online services to improve the data literacy of scientists and professionals and to facilitate the preparation of data management plans[10].

**Local context**

- The University of Lille has a mandatory ETD policy; all PhD dissertations must be submitted in digital format, for deposit in the national STAR system.
- A new institutional repository is under development (Dspace). It is uncertain to which extent it will accept the deposit of datasets, especially from PhD students.
- The SSH campus provides a solution for the sharing of files in the cloud; however, the storage capacity is limited, and the server does not guarantee long term preservation

**Needs**

- Following our own and other survey results (Schöpfel & Prost 2016), scientists express above all a need for storage and long term preservation solutions, along with advice and assistance for RDM, for both the deposit as well as for the description and legal aspects. Data sharing, especially open access, is not a priority.
- PhD students have less experience with RDM but are more motivated than other scientists in data sharing. They are also motivated by the RDM related criteria of project calls and funding agencies' programs, especially by the EU Framework Programme for Research and Innovation[11] which requires the submission of a data managemen plan (DMP) along with the project proposal[12]; they know they must be compliant with these criteria in order to get funding for their research.

[8] http://star.theses.fr/

[9] http://www.theses.fr/

[10] Platform DoRANum http://doranum.fr/ and service DMP OPIDoR https://opidor-preprod.inist.fr/

[11] Horizon 2020 or H2020, see https://ec.europa.eu/programmes/horizon2020/

[12] See the H2020 guidelines http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

## The development of a local workflow

Compliant with our surveys on the emerging environment of data repositories[13] and, in particular, on similar projects[14] we decided to create a local workflow for the deposit of research data by PhD students. Our assumptions:

- No new data repository (institutional/local or disciplinary) but a connection to existing infrastructure in SSH.
- A solution linked to the national ETD system by metadata and identifiers but independent of this system.
- A separation between data and dissertations from the beginning and onwards (separate deposit).
- A "by default" solution, complementary to specific data repositories.
- An integrative, complete solution covering all needs (recording, preservation, dissemination...).

The guiding principle was to provide an interface (with technical assistance) on our campus for the deposit of research data on the NAKALA platform of the national infrastructure for SSH communities. Figure 1 presents the main aspects of our solution.

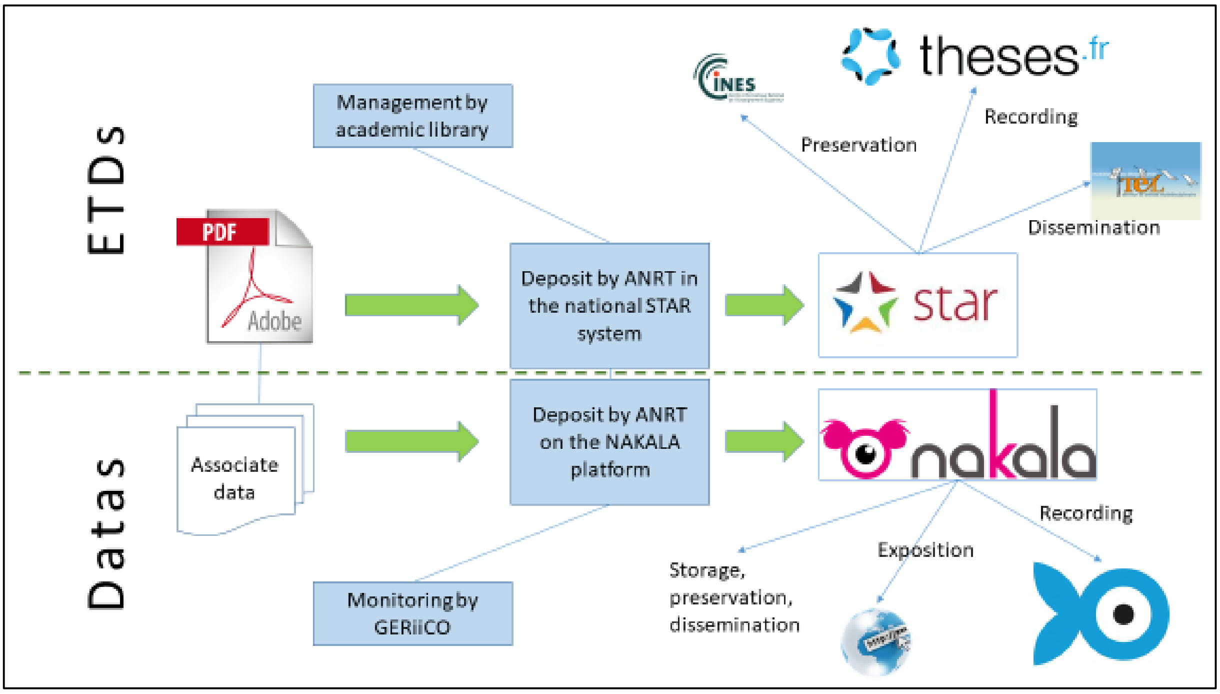

ETDs
Datas
Management by academic library
PDF
Adobe
Deposit by ANRT in the national STAR system
star
theses.fr
CINES
Preservation
Recording
Dissemination
Associate data
Deposit by ANRT on the NAKALA platform
nakala
Monitoring by GERiiCO
Storage, preservation, dissemination
Exposition
Recording

Figure 1: Local ETD/data workflow

As before, the ETDs will be submitted to the national STAR system by the academic library which in 2018 will integrate the actual ANRT staff; via the STAR system, the ETDs will be preserved by the national CINES agency, with their metadata being disseminated by the national academic union catalogue SUDOC and the national ETD portal Theses.fr.

[13] Nearly 2,000 sites indexed by the international directory re3data http://www.re3data.org/

[14] For instance, the ETDplus project funded by Educopia https://educopia.org/research/grants/etdplus and the workflow at the University of Bielefeld, see Vompras & Schirrwagen (2015)

Following the PhD students' choice, the text of the dissertation can be published on the national open access TEL server, the Lille institutional repository and/or on another platform.

The same staff will deposit the associated datasets on the NAKALA platform, create the metadata and link them to the dissertation on STAR. After formal validation and acceptation of the files, NAKALA will guarantee the preservation, the dissemination (following the students' choice), the exposition of the metadata on the web and the indexing by the Huma-Num discovery tool ISIDORE[15]. Like the ETD deposit, the academic library will supervise the submission of datasets together with the research laboratory GERiiCO[16], which will be in charge of the scientific follow-up of the workflow.

Thus, our main problem was not the creation of a new system but the connection between existing systems, with questions related to compliance and interoperability. The discussion with the NAKALA team identified eleven specific issues where action has to be taken:

Content/coverage

- Granularity: what exactly should be defined as a dataset for deposit? We have discussed this question in two communications (Schöpfel et al. 2016, 2017). There are no clear rules or guidelines. The pragmatic solution is to accept datasets on a granularity level, which makes sense for understanding (validation) and reuse, and to allow deposit of dataset collections with a hierarchical structure.
- Data format: which formats can be accepted? While the national ETD system only accepts PDF files, the NAKALA data repository supports all file formats that can be accepted by the national academic digital archive in Montpellier[17] so that we can use their checklist FACILE as a filter for the validation of acceptable file format[18].
- Database: how should larger databases (surveys, inventories, text samples etc.) be dealt with? What are the limits for deposits on the NAKALA platform? This issue is part of the tests with the Huma-Num team.

Metadata

- Indexing: who should do the indexing? Our idea is that the indexing should be done and supervised by information professionals, based on the basic metadata provided by the PhD students for the national ETD system STAR.
- Data structure: how should data be described and structured? Our option would be to apply the Metadata Encoding & Transmission Standard of the Library of Congress[19] but we still have to assess the compliance of METS with the NAKALA platform.
- Referential: we decided to index five elements of the Dublin Core following a qualified metadata schema (file name, data type, creator, date, title). This means that we have to prepare precise descriptions and term lists and determine what is acceptable for these DC elements. These metadata together with the ETD and data identifiers will be used for the connection between dissertations on STAR and data on NAKALA.

[15] ISIDORE combines a search engine and a metadata harvester for all kind of SSH data from the Huma-Num infrastructure **https://www.rechercheisidore.fr/**

[16] Information and communication sciences, **http://geriico.recherche.univ-lille3.fr/**

[17] CINES **https://www.cines.fr/**

[18] **https://facile.cines.fr/**

[19] METS **http://www.loc.gov/standards/mets/**

- Identifier: which unique identifier should be used for the datasets? Even if France is part of the DataCite consortium for the assignment of DOIs20, we opt for the moment for the handle system which is applied by the Huma-Num infrastructure, but we remain open for future adoption of the DOI.
- Source code: a last issue is how to describe sources code-related to datasets. How can this information be included in metadata? So far we have no solution. Perhaps, this is out of scope, at least for the moment and/or for this project.

Other issues

- Legal aspects: we anticipate legal issues like copyright, third party rights, privacy etc. Our approach is twofold: we provide basic legal advice as part of the library's data service, and we ask the students to provide a declaration (template) that they have the permission to upload the datasets on NAKALA.
- Deposit: who has access to the NAKALA platform? Who is an authorized user? Our first choice is to limit access to the project team (i.e. information professionals of the academic library, with a generic address and identification via the national academic IT network RENATER) and to prohibit self-archiving. However, this may change in the future.
- Data size: actually, we don't know exactly what will be the potential data volume. In average, 60-80 PhD dissertations are submitted per year on our campus, representing roughly 2 GB. But except for very few dissertations, these deposits in the national ETD system do not contain data files. So all we can do is try to make some estimations, perhaps also with our international partner projects.

Four other issues have been raised but they are not directly linked to the development of the workflow:

- Long term preservation: Up to now, the NAKALA platform does not guarantee long-term preservation of submitted datasets. But they have an agreement with the national CINES agency which ensures long term preservation of backups of the different Huma-Num platforms' content, which means that the NAKALA datasets could be recovered if necessary.
- Quality: the question was raised about the quality of datasets. Should all datasets provided by PhD students be accepted? Should we set up a kind of validation procedure? If so, which criteria should be applied? Who should evaluate? For the moment, we will not filter submitted data files otherwise than by formal criteria (size, format...), similar to other projects and data repositories. But the question remains open.
- Promotion: we have already discussed how to promote the new data service - who should do this, what the best communication vectors are, and what should be the message. As mentioned before, the main message will not be "PhD students must share their data" but rather "we can provide a solution for the preservation of your research data". And the message will be communicated via the Graduate School, the Research Department and research laboratories and the academic library. Also, our intention is not to make the deposit of datasets mandatory but to promote and incite data deposit as a form of good scientific practice.

[20] **http://www.inist.fr/?DOI-Assignment&lang=en**

- Technical documentation: after the launch of the new workflow, we will have to write the technical documentation, a procedure on two levels, one for the professional staff, the other for the students in the form of guidelines or recommendations to facilitate the process of submission and deposit.

The tests of the new workflow started at the end of September 2017. The workflow will be operational in 2018. We will perhaps customize the Huma-Num interface for the submission of datasets and the creation of metadata but this is not essential for the project.

We mentioned above that the University of Lille has started to develop a new institutional repository (Dspace). A priori, this would not modify the workflow for ETD related datasets, as it would not modify the submission of ETDs to the national STAR system. As Dspace is able to harvest metadata and to integrate different identifiers and outbound links, connecting the NAKALA datasets to the institutional repository should not be a problem.

## The PhD training program

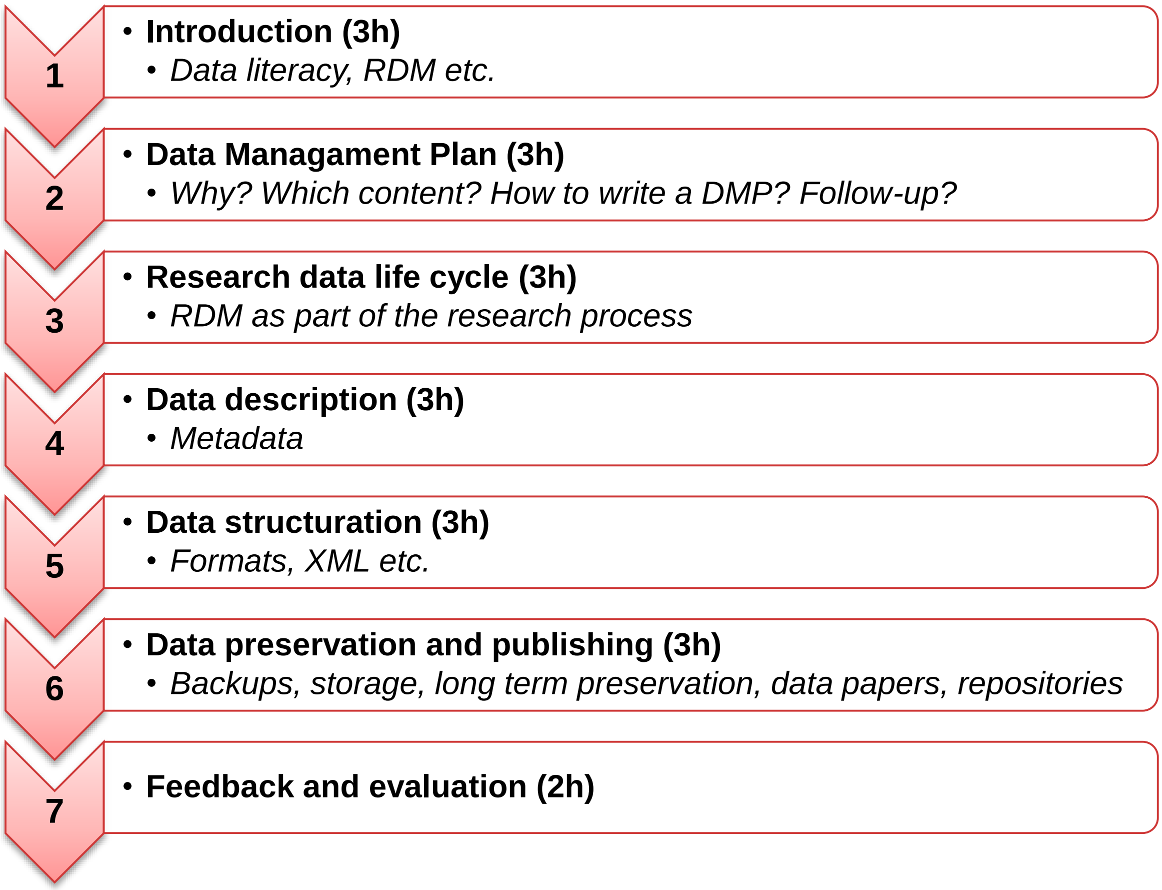
Our assumption is that RDM will become a part of basic scientific skills and good practices of research work, like sampling, statistics, surveys or systematic reviews. PhD students will have to obtain data literacy as part of their scientific education and training program. We started to work with PhD students on RDM nearly three years ago. Our experience is that most of them get some elements of data literacy through their research practice (e.g. privacy issues, ethics, backups) but they lack an overall ability of data curation, including description, preservation, and sharing[21].

Following this assumption we launched a training program to develop and enhance RDM skills for PhD students in SSH, together with the research support service of the academic library but as part of the Graduate School training program, not as a library course. The integration in the official program of the Graduate School and the mixed pedagogic team (four scientists and one academic librarian) partly explains the success of the program, together with the newness and interest of the topic itself. Since the beginning, about 40 PhD students have participated.

Our first training program (2015) consisted of three seminars (3 x 6 hours), organized together with scientists from the University of Lille and other institutions, on research data management, legal aspects and potential reuse and exploitation of data, including content mining.

In 2016 and 2017, we organised the program in a more traditional way, as a seminar with seven sessions (6 x 3 hours and 1 x 2 hours), one session per month from January to June, on different topics (figure 2).

[21] Concerning the global concept of RDM, see the overview by Neuroth et al. 2013



- **Introduction (3h)**
  - *Data literacy, RDM etc.*
- **Data Managament Plan (3h)**
  - *Why? Which content? How to write a DMP? Follow-up?*
- **Research data life cycle (3h)**
  - *RDM as part of the research process*
- **Data description (3h)**
  - *Metadata*
- **Data structuration (3h)**
  - *Formats, XML etc.*
- **Data preservation and publishing (3h)**
  - *Backups, storage, long term preservation, data papers, repositories*
- **Feedback and evaluation (2h)**

Figure 2: RDM training modules

We try to cover the most important elements of RDM, with a double objective: provide theoretical and operational knowledge on research data and data management (data literacy), and develop practical skills (data behaviour). Based on the feedback of the first two years, our 2017 seminar focuses on the PhD students' own scientific experience and data behaviour; in each session, the introduction and overview is followed by discussion, practice and individual follow-up. Each session takes place in a computer room.

The operational goal of the seminar is the writing of a data management plan for each PhD project on the OPIDoR platform[22], which is the French adaptation of the JISC DMPonline service[23]. Each student creates his/her personal account on OPIDoR, writes a DMP with the European Commission's H2020 template and shares the DMP with the seminar's training staff, which provides individual follow-up, comments, suggestions etc. directly on the platform. At the end of the seminar, the staff downloads the final version of each DMP for evaluation and direct feedback.

Concretely, each student completed the H2020 initial DMP template, i.e. a general description of the research project followed by five issues with specifications for each dataset:

- Dataset reference and name
- Dataset description

[22] Hosted by INIST https://dmp.opidor.fr/

[23] Hosted by the JISC Digital Curation Centre https://dmponline.dcc.ac.uk/

- Standards and metadata
- Data sharing
- Archiving and preservation (including storage and backup)

For instance, a student preparing an anthropological and ethnographic study on the perception of the 2016 Olympic Games by the people from Rio de Janeiro described three different data types, based on questionnaires, interviews and photos, and explained how he will index the data sets, with whom he will share them, and how he will store and preserve them.

Our experience with this approach is promising: with personal guidance, follow-up and feedback, the PhD students not only learn to write their own DMP but, in doing so, they learn to anticipate the essential issues of research data curation, like standard description, systematic back-up and secure storage and, in the end, long-term preservation and publishing. They also learn to prepare DMP in good and due form, compliant with the H2020 criteria, which will be an essential asset for future research work and project submission. Also, their DMP will contribute, as didactic material, to further training.

Nevertheless, we also observed that students who are just beginning their PhD have different questions and needs from those in second or third year of graduation who already have their methodology and often also datasets. In the future (2018), we will probably divide the seminar in two parts, the first one (modules 1-3, see figure 2) for "beginners" and the second (modules 4-7) for "advanced". Even so, the writing of a DMP will be part of both.

## Lessons learned

We must keep in mind that our primary project target is not to foster or increase open access to/in research data. Our main goal is to help young scientists in SSH develop their data literacy, i.e. their RDM related skills, to raise awareness about the challenges of data curation and publishing, and to provide a solution by default for their datasets. Our project is an investment in the future, and we expect the young scientists to have efficient and modern data behaviours, just as we expect them to develop new methodologies and conceptual approaches in their research field. Data sharing is part of the game, but not the only or most important purpose.

1 - Our experience confirms that in many respects, RDM is not just a technical problem but a “people problem” (Ward et al. 2011). “Improving tools are not the only steps necessary to overcome barriers. The next steps will likely involve training for scientists (...)” (Tenopir et al. 2015). In fact, solutions are often available (cf. Kindling et al. 2017). What is needed are new skills[24], information, promotion and incitation to use these solutions; paradoxically, we can say that our focus is not on data but on people. The most important decision of our project was perhaps the choice of a specific target group - young scientists and specifically PhD students in SSH. It was this choice that made our project intelligible and distinctive, on our campus as well as in the French HE landscape.

[24] Of course, this includes also the acquisition of new data skills by the project team.

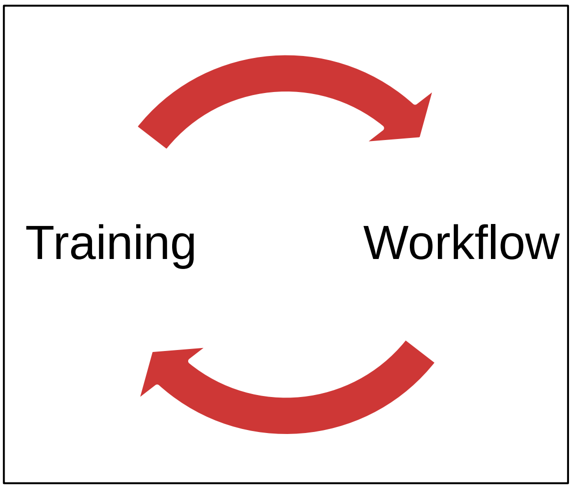


Figure 3: PhD training and ETD/data workflow

2 - Technical solution and education are related (figure 3). Data literacy (DMP, RDM) is more and more considered as good scientific practice, like usual scientific skills (sampling, statistics etc.). Our experience and conviction is that it is not enough to develop RDM tools if we do not teach scientists how (and why) to use them. Therefore, available infrastructure will shape the content of the training program (e.g. for file formats or dataset structuration); but then again, the discussion and feedback from the training program contribute to the further development of the campus-based RDM solutions. Our approach can be considered as an organizational learning process on the campus.

3 - The project is run together with the academic library research support team. However, it is NOT a library project. The library staff is part of the project, with specific tasks and skills. Their contribution is essential for the success of the project, and they are of course members of the project steering committee. Yet, from the beginning the project was designed as a research project with a doctoral training program, under the responsibility of the graduate school, with a scientific project management held by our research laboratory, and under the political and strategic leadership of our academic research department. There are at least two reasons: legitimacy (in the sense that scientists have everyday experience with RDM); and the fact that scientists do not usually consider RDM as a "library affair" but as part of their daily research work with other scientists and technical staff.

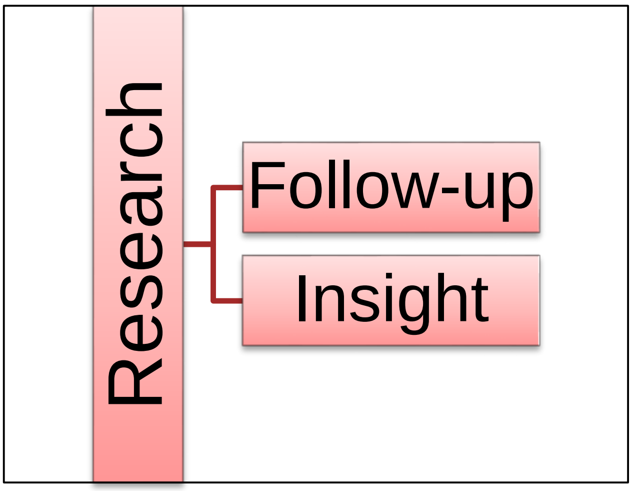


Figure 4: RDM related research

4 - The project is not limited to education and workflow. As said above, its main character is research. We can distinguish two different levels of RDM related research, i.e. follow-up and insight (figure 4). Follow-up means monitoring and assessment of the training program, including feedback and continuous adjustment; and it means evaluation of uptake and usage of the new data workflow. Insight covers a larger range of data related topics, such as type, format and content of datasets, the impact of data on format and content of ETDs and the text and data mining of dissertations and data. At present, we are working on the first issue, together with partners from different universities and institutions. In the future, we will focus on the second issue and prepare, together with German colleagues and ProQuest, an international research project on new formats of PhD dissertations.

## References

AWRE, Chris et al., 2015. Research Data Management as a “wicked problem”. *Library Review*. **64**(4-5), 356-371. ISSN 0024-2535.

CHAUDIRON, Stéphane, Catherine MAIGNANT, Joachim SCHÖPFEL and Isabelle WESTEEL, 2015. Les données de la recherche dans les thèses de doctorat - Livre blanc: Rapport de recherche [online]. Université de Lille 3, [Accessed 25 September 2017]. Available from: **http://hal-univ-lille3.archives-ouvertes.fr/hal-01192930/document**

KINDLING, Maxi et al, 2017. The Landscape of Research Data Repositories in 2015: A re3data Analysis. *D-Lib Magazine* [online] **23**(3/4). [Accessed 25 September 2017]. Available from: **http://www.dlib.org/dlib/march17/kindling/03kindling.html**

NEUROTH, Heike (EDS.), 2013. Digital curation of research data experiences of a baseline study in Germany [online]. Glückstadt: Vwh, Hülsbusch, Fachverl. für Medientechnik und -wirtschaft [Accessed 25 September 2017]. ISBN 978-386-4880-544.

PROST, Hélène, MALLERET, Cécile & SCHÖPFEL, Joachim. Hidden Treasures. Opening Data in PhD Dissertations in Social Sciences and Humanities. *Journal of Librarianship and Scholarly Communication*. 2015, **3**(2), eP1230.

SCHÖPFEL, Joachim et al, 2014. Open Access to Research Data in Electronic Theses and Dissertations: An Overview. *Library Hi Tech*. **32**(4), 612-627.

SCHÖPFEL, Joachim, Hélène PROST and Cécile MALLERET. Making data in PhD dissertations reusable for research. In: *Conference on Grey Literature and Repositories* [online]. Prague: National Library of Technology, 2015 [Accessed 25 September 2017]. ISSN 2336 - 5021. Available from: http://repozitar.techlib.cz/record/993/files/idr-993_1.pdf

SCHÖPFEL, Joachim and Hélène PROST, 2016. Research data management in social sciences and humanities: A survey at the University of Lille 3 (France). *LIBREAS. Library Ideas* [online]. [Accessed 25 September 2017]. 29, p 98-112. ISSN ISSN: 1860-7950. Available from: **http://libreas.eu/ausgabe29/09schoepfel/**

SCHÖPFEL, Joachim, PROST, Hélène and Violaine REBOUILLAT, 2016. Research data in current research information systems. In: *CRIS 2016, St Andrews, 8-11 June 2016* [online]. [Accessed 25 September 2017]. Available from: **http://hdl.handle.net/11366/501**.

SCHÖPFEL, Joachim, KERGOSIEN, Eric and Hélène PROST, 2017. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse.In: *Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017, 31 mai 2017 Toulouse (France)* [online]. [Accessed 25 September 2017]. Available from: **http://hal.univ-lille3.fr/hal-01530937**

TENOPIR, Carol et al, 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE*. 2015, **10**(8), e0134826+.

VOMPRAS, Johanna and Jochen SCHIRRWAGEN. Repository workflow for interlinking research data with grey literature. In: *Conference on Grey Literature and Repositories* [online]. Prague: National Library of Technology, 2015 [Accessed 25 September 2017]. ISSN 2336 - 5021. Available from: **http://repozitar.techlib.cz/record/993/files/idr-993_1.pdf**

WARD, Catharine et al. Making sense: Talking data management with scientists, 2011. *International Journal of Digital Curation*. **6**(2), 265-273. ISSN: 1746-8256.

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

# ATTITUDES OF CHARLES UNIVERSITY ACADEMIC STAFF TO DATA SHARING

**Adéla Jarolímková**

**Adela.jarolimkova@ff.cuni.cz,**

**Institute of Information Studies and Librarianship, Faculty of Arts, Charles University**

**Barbora Drobíková**

**barbora.drobikova@ff.cuni.cz**

**Institute of Information Studies and Librarianship, Faculty of Arts, Charles University**

**Martin Souček**

**martin.soucek@ff.cuni.cz**

**Institute of Information Studies and Librarianship, Faculty of Arts, Charles University**

## Abstract

Data management and sharing are an integral part of contemporary research work. At Charles University, we carried out a survey of selected aspects of current data management practices and researchers' attitudes to data management and sharing. In our paper we present a part of its results focused on academic staff and comparison of their answers with the answers of doctoral students, interdisciplinary comparisons, selected comments and recommendations based on survey results.

## Klíčová slova

Research data management, research data sharing, open access, researchers, academic staff

The publication was created as part of the Progres Q15 programme at Charles' University entitled "Life paths, lifestyles and quality of life from the perspective of individual adaptation and relationships between actors and institutions".

## Introduction

Work with data is an integral part of contemporary science. However, skills and knowledge that are not currently a common part of higher education are required for the administration, storage and sharing of data. For this reason more and more research is focusing on the level of "data literacy", i.e. a set of skills that make it possible to search for, interpret, critically evaluate, administer and ethically use data (Calzada Prado 2013), the needs that scientists have in the area and proposed education programmes (for example, Carlson 2011, Haendel 2012, Sapp Nelson 2017, etc.). Within the Czech environment, only Pavlásková has dealt with research data in her dissertation (Pavlásková 2016).

For this reason we took the opportunity to take part in an international comparative study of data literacy and the management of research data[1] that was first conducted in France, Turkey and Great Britain and whose initial results were presented at the ECIL conference (Chowdhury 2016).

Within the bounds of this research, data is considered to be any information stored in digital format, including text, numbers, images, video or film, sound, software, algorithms, equations, animations, models, simulations, etc.

## Methods

The Czech version of the questionnaire was used in the survey. A description and selected results have already been published (Jarolímková et al. 2017, Drobíková et al. 2017). The questionnaire was sent out by e-mail to all academic workers and doctoral candidates at Charles' University.

In this article we concentrate on the part of the results to concern the attitudes of academic workers to the sharing of research data and their current approach to sharing and which consists of a total of six questions (see Table 1).

[1] Data from all participating countries will be published at a common website. Data from the Czech part of the survey is available as DROBIKOVA, Barbora, JAROLIMKOVA, Adela, and Martin SOUCEK, 2017. Data literacy and research data management survey [Data set]. Zenodo. **http://doi.org/10.5281/zenodo.997844**

| |
|---|
| **Which of the following applies to your research data? (My data is openly available to everyone/ My data is openly available only to my research team/ My data is available openly upon request/ My data has restricted access (e.g. only some parts of the dataset is accessible) My data is not available to anyone else)** |
| **Do you have any concerns for sharing data with others? (No concerns/ Fear of losing the scientific edge/ Legal and ethical issues/ Misuse of data/ Misinterpretation of data/ Lack of resources (technical, financial, personnel, etc.)/ Lack of appropriate policies and rights protection/Any other)** |
| **Do you collaborate with other researchers and share data? (No/ Yes, with researchers in the same team/ Yes, with researchers in the same university/ Yes, with researchers in other institutions/ Any other)** |
| **I am familiar with the open access requirements (Strongly agree/Agree/Neither agree nor disagree/Disagree/Strongly disagree)** |
| **I am comfortable and willing to share my research data with others (Strongly agree/Agree/Neither agree nor disagree/Disagree/Strongly disagree)** |
| **I perceive data ethics could be an issue when research data is shared with others (Strongly agree/Agree/Neither agree nor disagree/Disagree/Strongly disagree)** |

Table 1: Questions relating to the sharing of research data

Descriptive statistics were used when interpreting data and Pearson's chi-squared test to search for connections between questions at a reliability level of α= 0.05. The answers toquestions in which a 5-scale Likert scale was used were merged as follows to ensure clearer arrangement: strongly agree and agree as yes, disagree and strongly disagree as no.

## Results

A total of 2,381 responses were obtained, although only 1,434 questionnaires were completed in full. Given that only the section on demographic characteristics was completed in the incomplete questionnaires, these questionnaires were omitted from the analysis. A total of 603 complete questionnaires were obtained from academic workers at Charles' University. To ensure clear arrangement, questionnaires were divided into four basic specialisations - humanities, medicine, natural science and social science (see figure 1). Engineering and agriculture were also represented in 7 questionnaires, but these were omitted due to the low representation of those specialisations.

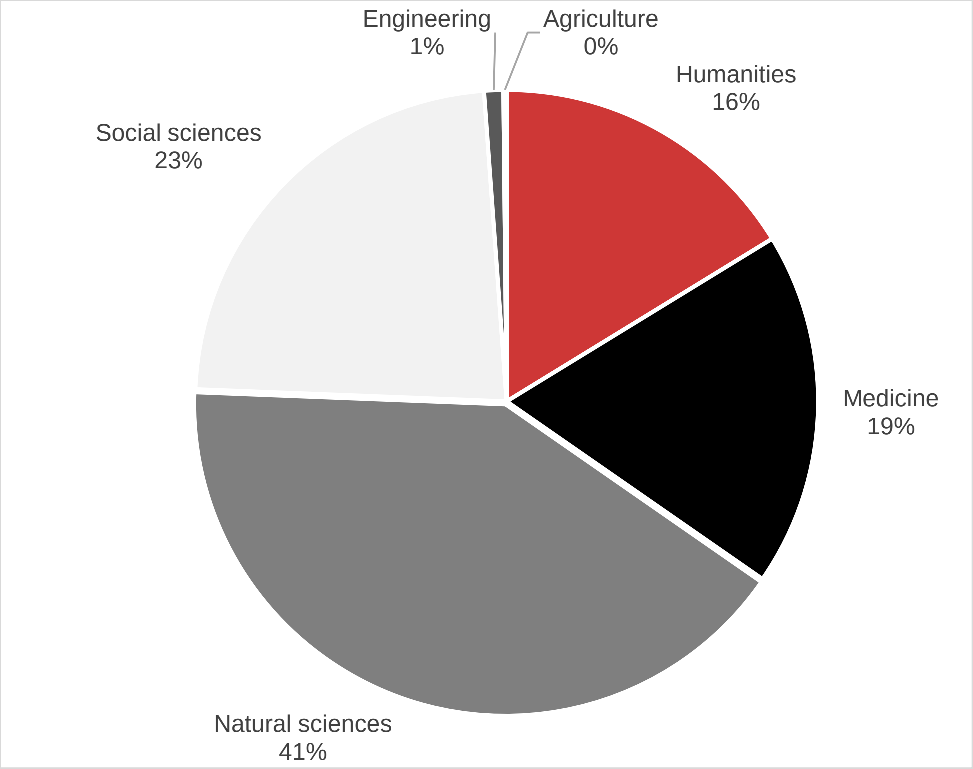


Figure 1: Respondents by specialisation

As far as the age structure of respondents is concerned, most respondents were between 36 and 45, with the category of 65 and over having the lowest representation (Figure 2).

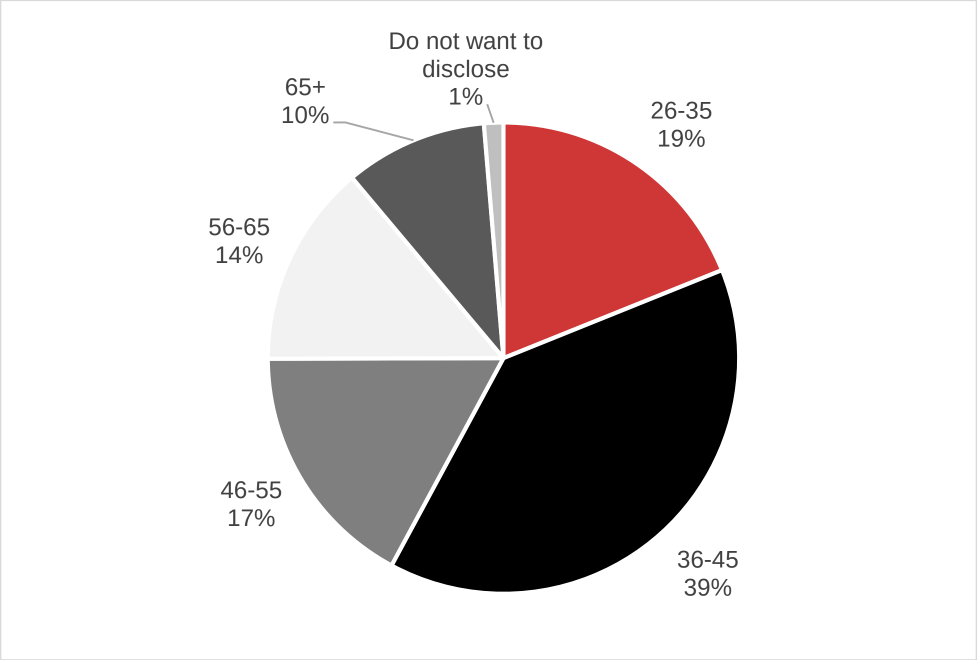


Figure 2: Respondents by age

Some 48 % of respondents from the ranks of academics are familiar with the requirements of open access (n - 596), which is higher than among doctoral candidates, only 36 % of which stated having knowledge of open access (n=826).

Most academics have already shared their data (see Figure 3). Only 13 % (n=596) stated that they did not share data in any way (Figure 3). The highest number of those who did not share came from social sciences (21 % n=140), while the percentage of those not sharing in natural sciences was only 9 % (n=247). Data is most commonly shared within a team. Some 45 % of respondents (n=596) shared their data with scientists from other institutions. In comparison with the results published in Drobíková et al, 2017, academics share their data more frequently than doctoral candidates (p<0.001) and differences were also found between age categories: 8 % in the youngest age category (26-35 years) do not share data, whereas this figure is 25 % of respondents in the oldest age category (65+).

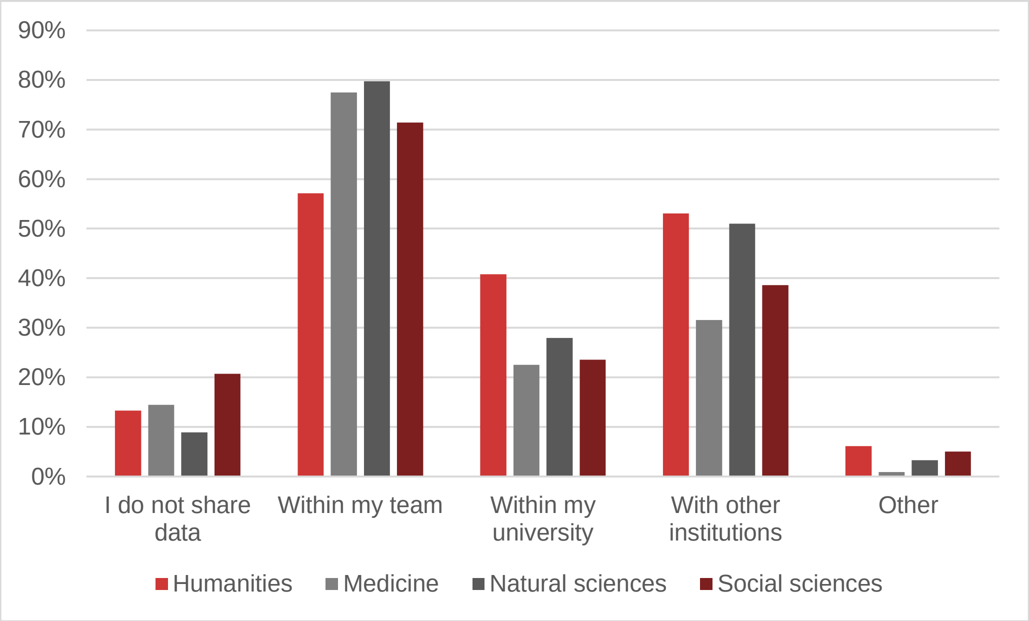


Figure 3: Current practice in data sharing

As was ascertained during the analysis of results for doctoral candidates, however, sharing does not automatically mean open access to data (Drobíková et al., 2017). On the contrary, open access is the least common method of academics sharing their data, in that there are no significant differences between specialisations here (see Figure 4). Those academics that are familiar with the principles of open access provide open access to their data more often than those who are not (p=0.002). Academics also provide their data openly to a greater extent than.

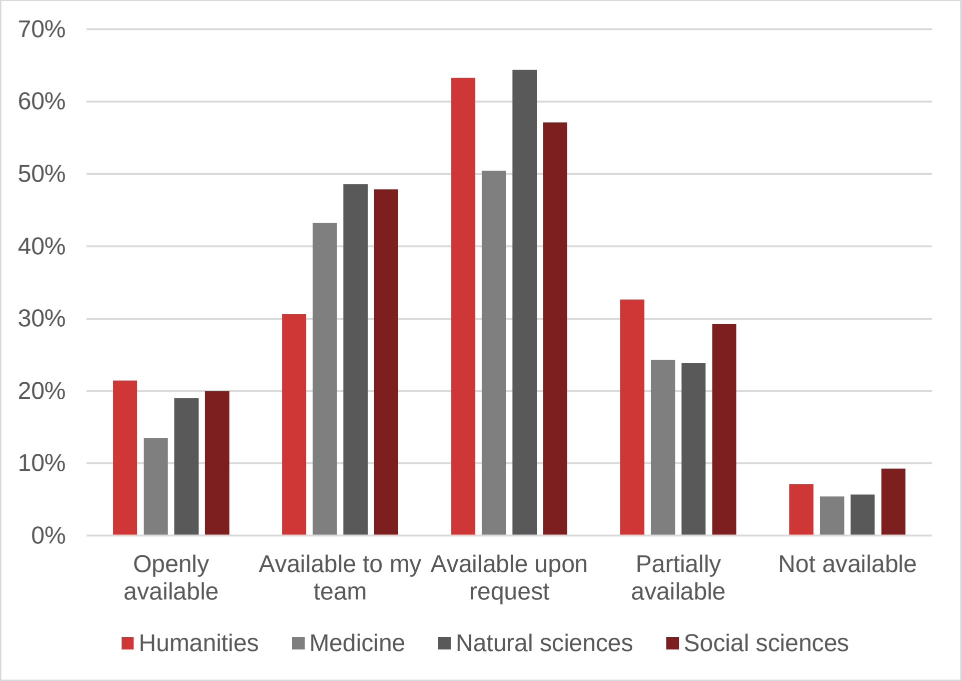


Figure 4: Method of making data available

The answers to the question on fears relating to sharing data brought some interesting results (Figure 5). Respondents were able to choose more than one answer to this question. More than a third of respondents/academics (36 %, n=596) have no fears about sharing data, whereby there is a significant difference between respondents from the humanities and natural sciences, where 41 % (n=98) and 45 % (n=247) respectively had no fears on the one hand and, on the other, medicine with (26 %, n=111) and social sciences (24 %, n=140) on the other. The highest number of academics fear incorrect interpretation of data (35 %, n=596), in that the differences between specialisations are not statistically significant. Thirty-one per cent of respondents fear legal and ethical problems, and there is again a difference here between medicine and social sciences, where respondent fears are more common, and other specialisations. Respondents from medicine also fear the misuse of data more than those from other specialisations (p=0.004). Only a small number of academics fear a lack of resources or the absence of guidelines in the sphere of research data management - on the contrary, it ensued from comments that respondents have greater fears over an excess of guidelines and regulations in this area.

Fears also affect the willingness to share data. Those that have no fears are simultaneously more often willing to share their data (p<0.001). Fears of misuse (p<0.001) and of legal and ethical problems (p=0.003) have a negative influence on the sharing of data.

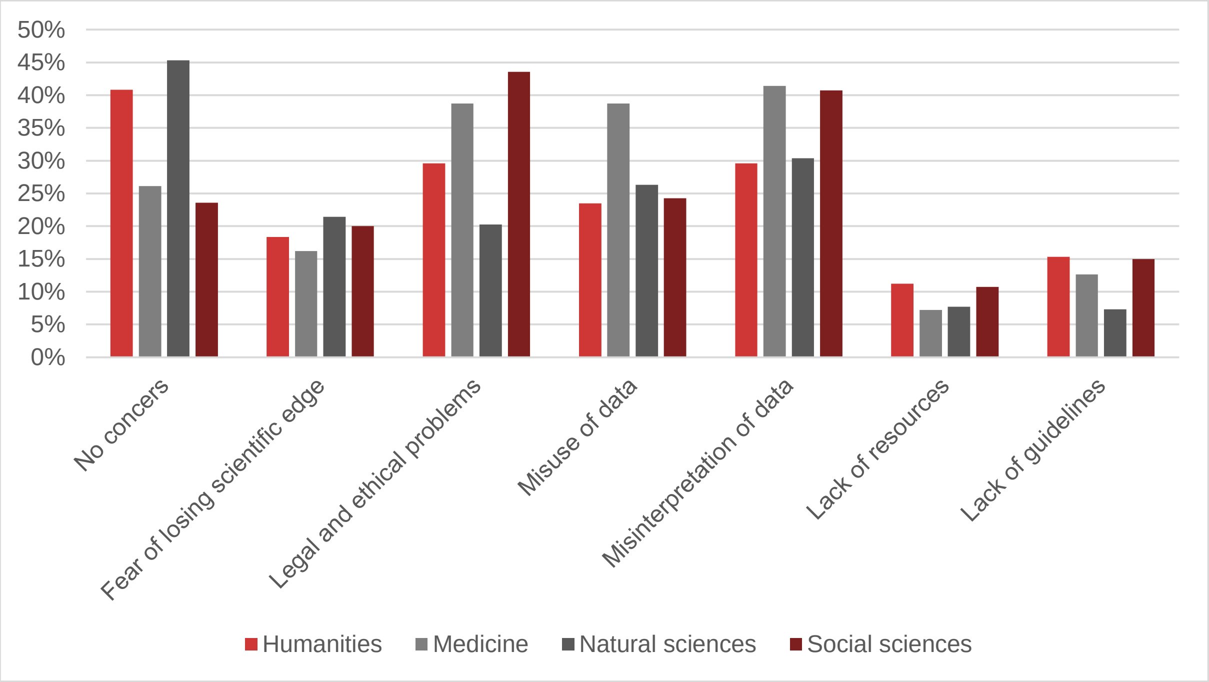


Figure 5: Fears associated with sharing data

A complementary question to the above was whether scientists are willing to share their data (Figure 6). Scientists could answer strongly agree, agree, neither agree nor disagree, disagree or strongly disagree. To ensure greater clarity, we combined the answers strongly agree and agree and the answers disagree and strongly disagree in the graph. A positive response to this question predominated in all areas of science. Scientists are willing to share their data. The answer of "agree" had strongest representation among scientists from the humanities (74 % agree to 12 % disagree, n=98). By contrast, it was lowest among scientists from the sphere of medicine (55 % agree to 23 % disagree, n=111). The attitude of scientists to this issue is consistent with the previous question. Nonetheless, the calculated chi-squared test value of p=0.13 does not confirm dependence of membership of a specialisation on willingness to share data.

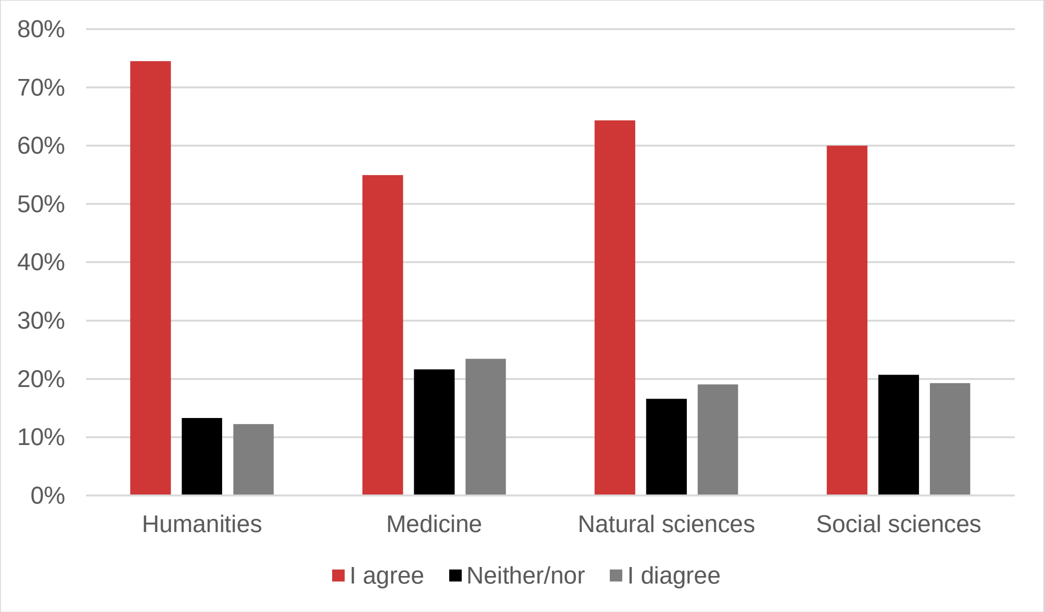


Figure 6: Willingness to share data

The questionnaire also asked the question of whether scientists would consider themselves to be exposed to ethical problems by sharing data (Figure 7). The majority of academics from all four areas of science represented invariably agreed that certain ethical problems could arise. In contrast to other areas of science, more than a fifth academics from the natural sciences (26 %, n=247) think that ethical problems cannot arise, which is a significant difference when compared with the opinions of academics from other areas of science. On the contrary, the vast majority of academics from the spheres of medicine (72 %, n=111) and social sciences (69 %, n=140) think that problems could arise.

The calculated chi-squared test value of p=0.004 confirms dependence of opinions in the sphere of ethical problems on scientific specialisation.

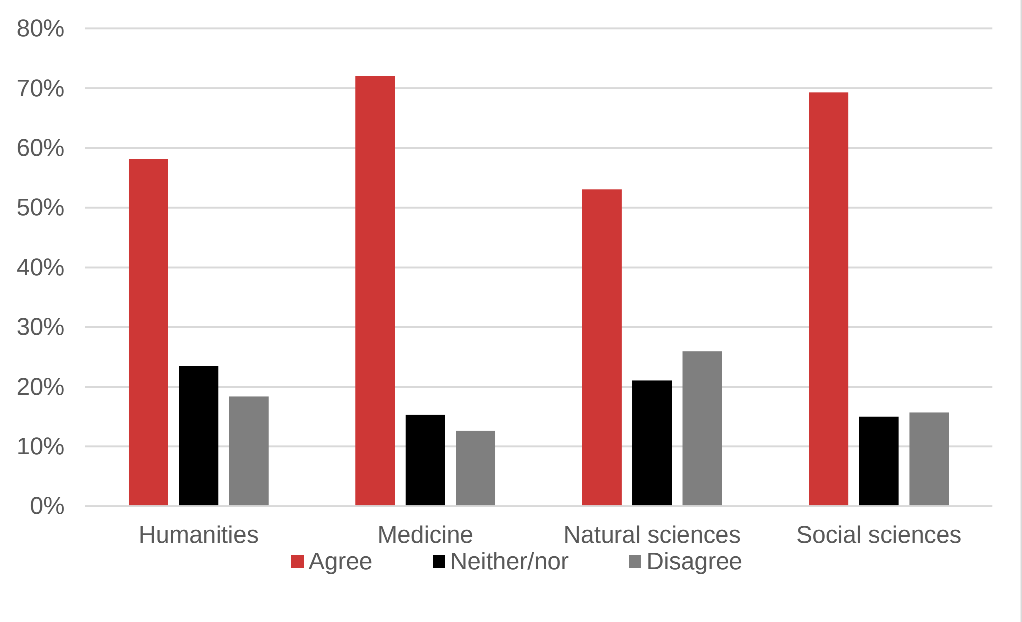


Figure 7: Ethical problems in sharing data

## Discussion

The conclusions of the survey brought an initial insight into the issue of research data at Charles' University and it can be said that they do not differ significantly from similar research abroad (for example, Tenopir 2015, Enwald 2017). Academic workers at Charles' University are willing to share their data and most of them do so at present. Their approach is not one of open access, however, a fact that might be influenced by several factors. First in line is that less than half of academics are familiar with the principles of open access. Secondly, there is no simple solution for storing data in the form of an open university repository. Some specialisations have their own area-specific repository (LINDAT/CLARIN, Czech Social Sciences Data Archive), while others are reliant on international services such as Zenodo and Dryad. In respect of the fact that most respondents had not undergone any training in data management, it is understandably more difficult for them to find their way around this area, to choose the right repository and to work with it. It also emerged from the comments to the survey that open access, or indeed data sharing in general, does not make sense in all specialisations.

The willingness to openly share data is also influenced by fears over possible problems, particularly legal and ethical problems, and fears regarding the misuse or incorrect interpretation of data. There are more significant inter-disciplinary differences evident here, clearly arising from the nature of the data in individual specialisations and showing the need to adjust any solution for the administration of data, training and other activities in his area to suit individual areas of science. The data shows differences in the approach to research data between the humanities and natural sciences on the one hand and medicine and social sciences on the other. It ensues from the comments to the questionnaire that respondents also fear an increase in administration associated with data management, which was not an option provided in the questionnaire. Some refer to the technocratisation of research, the obstruction of intellectual activity or even the danger of the idea of open data sharing and there were negative comments regarding experience of the misuse or direct theft of data. As to question of sharing, it is important to respondents whether data is shared before the publication of results in the standard way or after this, something which was not differentiated in the questionnaire. Some comments also refer to the difficulty of creating a central solution with regard to the differences between specialisation, which is confirmed by the results of the questionnaire, although there were comments calling for a central repository.

## Conclusions

The important findings of the research are that academics and scientists at Charles' University are willing to share their research data, but that they see a variety of risks associated with sharing, and particularly with open access, and associate a further increase in their administrative load with data management. In order to support open access to data at Charles' University, therefore, it will be necessary to create a secure infrastructure for data sharing that suits the particularities of individual specialisations and to ensure support for data management, for example in academic libraries, so that scientists are not burdened by further administrative duties.

It is also clear that more research is required to deepen our understanding of certain aspects of data sharing, research conducted using quantitative and qualitative methods, and to concentrate mainly on the specifics of individual areas of science.

## References

CALZADA PRADO, Javier and Miguel Ángel MARZAL, 2013. Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents. *Libri* [online]. **63**(2), 123-134 [Accessed 9 April 2017]. DOI: 10.1515/libri-2013-0010. ISSN 18658423. Available from: **http://www.degruyter.com/view/j/libr.2013.63.issue-2/libri-2013-0010/libri-2013-0010.xml**

CARLSON, Jacob, Michael FOSMIRE, C.C. MILLER and Megan SAPP NELSON, 2011. Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *Libraries and the Academy*. **11**(2), 629-657.

DROBÍKOVÁ, Barbora, Adéla JAROLÍMKOVÁ and Martin SOUČEK, 2017. Data literacy of Charles University PhD students : are they prepared for their research careers? In: ŠPIRANEC, Sonja, Serap KURBANOGLU, Joumana BOUSTANY, Esther GRASSIAN, Diane, MIZRACHI, Loriene ROY and Denis KOS. *The Fifth European Conferece on Information Literacy (ECIL) : abstracts*. Saint-Malo: Information Literacy Association, p. 41. ISBN 978-2-9561952-0-7.

ENWALD, Heike, Terttu KORTELEINEN and Maija-Leena HUOTARI. Research data management: experiences of scholars in Finland. In: ŠPIRANEC, Sonja, Serap KURBANOGLU, Joumana BOUSTANY, Esther GRASSIAN, Diane MIZRACHI, Loriene ROY and Denis KOS. *The Fifth European Conferece on Information Literacy (ECIL) : abstracts*. Saint-Malo: Information Literacy Association, p. 46. ISBN 978-2-9561952-0-7.

HAENDEL, Melissa A., Nicole A. VASILEVSKY and Jacquline A. WURZ, 2012. Dealing with Data: A Case Study on Information and Data Management Literacy. *Plos Biology* [online]. **10**(5) [cit. 2017-04-09]. DOI: 10.1371/journal.pbio.1001339.

CHOWDHURY, G., G. WALTON, S. KURBANOGLU, Y. UNAL and J. BOUSTANY, 2016. Information Practices for Sustainability: Information, Data and Environmental Literacy. In: *SPIRANEC, S., S. KURBANOGLU and H. LANDOVA. The Fourth European Conference on Information Literacy (ECIL)*. Prague: Association of Libraries of Czech Universities, p. 22. ISBN 978-80-270-0530-7.

JAROLÍMKOVÁ, Adéla, Barbora DROBÍKOVÁ and Martin SOUČEK, 2017. Výzkumná data na Univerzitě Karlově. In: *INFORUM 2017: 23. ročník konference o profesionálních informačních zdrojích, Praha 30.-31. května 2017* [online]. Praha: Albertina icome Praha [Accessed 23 September 2017]. ISSN 1801-2213. Available from: **http://www.inforum.cz/pdf/2017/jarolimkova-adela.pdf**

PAVLASKOVA, Eliska, 2016. *Analýza výzkumných dat na základě fondu disertačních prací Univerzity Karlovy v Praze s ohledem na dlouhodobé uložení digitálních objektů* [online]. Praha [Accessed 12 April 2017]. Available from: **https://is.cuni.cz/webapps/zzp/detail/103851/**. Dizertační práce. Univerzita Karlova. Filozofická fakulta. Vedoucí práce RNDr. Pavel Krbec, Ph.D.

SAPP NELSON, Megan R., 2017. A Pilot Competency Matrix for Data Management Skills: A Step toward the Development of Systematic Data Information Literacy Programs. *Journal of eScience Librarianship* [online]. **6**(1), [Accessed 9 April 2017]. Available from: **https://doi.org/10.7191/ jeslib.2017.1096**

TENOPIR, Carol, Elisabeth D. DALTON, Suzie ALLARD a Mike FRAME, 2015. Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. *PLos One* [online]. [Accessed 3 February 2017]. DOI: 10.1371/journal.pone.0134826. Available from: **http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0134826**

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

# DATA DEPOSIT INTO THE ASEP REPOSITORY

**Zdeňka Chmelařová**

**chmelarova@lib.cas.cz**

**The Czech Academy of Sciences, Library**

**Jana Doleželová**

**dolezelova@knav.cz**

**The Czech Academy of Sciences, Library**

## Abstract

The ASEP repository has provided bibliographic records deposit for the results of scientific research at the Czech Academy of Sciences since 1993. In 2012, the database was expanded to include a repository in which full text documents are stored. Since 2017, ASEP also supports data records and data file storage. The bibliographic records of the results can be linked to metadata records so that the user gets not only the full text of a result, but also the data on which the result is based. Each dataset has its own description and metadata follows international standards. In the paper, we will introduce the repository workflow, describe dataset deposit and international standards used as well as different types of user interfaces.

## Keywords

Data Repository; ASEP; Czech Academy of Sciences; Library of the Czech Academy of Sciences

## Introduction

The Library of the Czech Academy of Sciences (the "Library")[1] has, since 1993, been the administrator of the ASEP[2] (Automatizovaný systém evidence publikací – (Register of Publication Activity of the CAS) bibliographical database, in which bibliographical records and the full texts of documents of all fundamental results of basic research produced at the institutions of the Czech Academy of Sciences (CAS) are stored. Bibliographical records are the fundamental data pillar for the internal evaluation of institutions conducted by the management of CAS and for international evaluation of the results of science and research produces with financial support from the budget of the Czech Republic[3].

The function of the Library is not passive. Its range of duties include the development of the entire system, from the structure of data to the user environment and, last but not least, passing on information to all users that use the system in a way which is understandable. The development and modification of the ASEP system mainly focus on users from institutions of the CAS – authors, managers of individual institutions and the management of CAS itself. The Library monitors international development in the sphere of science and research and subsequently develops and modifies the system. A superstructure to the ASEP database was created in 2012 in the form of a repository of complete texts, meaning that each bibliographical record in ASEP can be accompanied by the full text of the document and the full text of reviews, or responses. The practical use of this function was positively received in evaluation by institutions at the CAS in 2015, when evaluators had on-line documents at their disposal for peer review. The ASEP system was further expanded in 2017 to include another superstructure – a data repository.

## Storing and archiving data files

Storing data files and sharing them with the scientific public is nothing new – there are many open institutional, multi-discipline and area-specific repositories that many authors from the CAS have used for a long time now. Area-specific repositories have been established at institutions of the CAS themselves, for example the Czech Social Science Data Archive (ČSDA)[4] at the Institute of Sociology, while the Institute of the Czech Language was a partner to the creation of the Lindat/Clarin repository[5]. An internal survey was conducted at the CAS to concern the archiving of data at institutes of the CAS and the interest shown by institutes in storing data in a data repository. An analysis of this survey shows that awareness of secure archiving is not at the sort of level it would merit. Files containing scientific data are most commonly stored on local computers and servers, not the safest places for archiving. We come across similar experiences in international surveys on the approach of scientists to storing and

[1] *Library of the Academy of Sciences of the Czech Republic* [online]. Prague: Knihovna AV ČR, v. v. i., ©2017 [cit. 26.9.2017]. Available from: **https://www.lib.cas.cz**

[2] *Online catalogue of the ASEP database* [online]. Prague: Knihovna AV ČR, v. v. i., ©1993-2017 [cit. 26.9.2017]. Available from: **https://asep.lib.cas.cz/arl-cav/cs/rozsirene-vyhledavani/**

[3] More on the evaluation of research, development and innovation: R&D Council. Evaluation of research and development. *Research and development in the Czech Republic* [online]. Prague: Research and Development Council, ©2015 [cit. 26.6.2017]. Available from: **http://www.vyzkum.cz/FrontClanek.aspx?idsekce=18748**

[4] *Czech Social Science Data Archive* [online]. Prague: Sociologický ústav AV ČR, v. v. i., ©2005-2014 [cit. 26.9.2017]. Available from: **http://nesstar.soc.cas.cz/webview/**

[5] *Lindat/Clarin* [online]. Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles' University, ©2017 [cit. 26.9.2017]. Available from: **https://lindat.mff.cuni.cz/cs/**

sharing data conducted at, for example, the University of Tartu in Estonia[6], or in studies dedicated to research data, in which Charles' University[7] was also involved. The conclusion of the article, that "there is no simple, unambiguous institutional recommendation of how authors should work with their data, even though scientists suspect that this is an important area to them as well", is entirely apt. The ASEP repository would provide authors from the CAS with the opportunity to archive data securely and over the long-term. The majority of academic institutes expressed an interest in this in the internal survey. We consider a data repository to be an important superstructure to the ASEP database and are convinced that the scientific public will come to appreciate it over time. We see the role of the Library in this area as being one of a mediator that passes on information regarding why to archive and share data, provides a place for storage, advises on how to describe it and attends to long-term protection and archiving. The reasons for archiving and sharing data are described in a number of documents.[8] The opportunity to verify the validity of conclusions in published documents, the effective use of data obtained from public sources, preventing scientific errors, etc., are most commonly mentioned. New results in industry and in other projects can be created based on the use of archived data from original research. Certain scientific magazines (for example, Nature, Science, The American Naturalist) already lay down conditions for data storage, when a scientist is obliged to share data together with a publication. The Public Library of Science publishing house issues instructions for sharing data and presents a list of suitable repositories. Authors to have received a grant from the H2020 programme are obliged to store the full text of the document in open access and, from 2017 onwards, the data files on which their publications were produced[9].

When creating a data repository, we used a number of examples of good practice, as published in the international Registry of Research Data Repositories[10] and in an overview of open institutional repositories at the Technical University of Ostrava website.[11] The DataShare repository of the University of Edinburgh[12], tried and trusted for many years now, was inspirational to us in terms of international institutional repositories, as was the Zenodo[13] project initiated by the EU and CERN in terms of multi-discipline repositories and the Lindat/Clarin repository in terms of area-specific repositories.

[6] MUULI, Viktor. *Research Data in Estonia: collecting, storing, availability: some findings from questionnaire* [online]. Estonian Research Council, 2014. 23.10.2014 [cit. 26.9.2017]. Available from: **http://dspace.ut.ee/bitstream/handle/10062/44052/RD_questionnaire_eng_muuli_14.pdf?sequence=1&isAllowed=y**

[7] JAROLÍMKOVÁ, Adéla. Výzkumná data na Univerzitě Karlově. In: *INFORUM 2017: 23rd Annual Conference on Professional Information Resources, Prague, 30.-31.5.2016* [online]. Prague: AiP, 2016 [cit. 26.9.2017]. ISSN 1801–2213. Available from: **http://www.inforum.cz/pdf/2017/jarolimkova-adela.pdf**

[8] HRABAL, Jan. Repozitáře vědeckých dat. In: *Knihovna.cz* [online]. Brno: Division of Information and Library Studies, Faculty of Arts, Masaryk University, ©2013. 22. 2. 2016 [cit. 26.9.2017]. Available from: **http://ltp.knihovna.cz/?p=385**

[9] *REGULATION (EU) No 1290/2013 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 11 December 2013 laying down the rules for participation and dissemination in "Horizon 2020 - the Framework Programme for Research and Innovation (2014-2020)" and repealing Regulation (EC) No 1906/2006 [online].*
In: Official Journal of the European Union. L 347/81, 20. 12. 2013, 23 s. [cit. 26.9.2017]. Available from: **https://www.h2020.cz/cs/storage/87c59c7c965787c1deb7a7c85ee5d5be89fbf58b?uid=87c59c7c965787c1deb7a7c85ee5d5be89fbf58b**

[10] *Registry of Research data Repositories* [online]. Re3data.org Project Consortium. [Cit. 26.9.2017]. Available from: **http://www.re3data.org/**

[11] *Green Open Access* [online]. VŠB – TUO Central Library, ©1998-2016**.** Most recent update 13.3.2017 [cit. 26.9.2016]. Available from: **http://knihovna.vsb.cz/open-access/green-open-access.htm**

[12] *Datashare* [online]. University of Edinburgh. [Cit. 26.9.2017]. Available from: **http://datashare.is.ed.ac.uk/**

[13] *Zenodo* [online]. [Cit. 26.9.2017]. Available from: **https://zenodo.org/**

There are 53 institutes[14] at the CAS, divided into three areas of science: I. the area of Mathematics, Physics and Earth Sciences, II. the area of Life and Chemical Sciences, and III. the area of Humanities and Social Sciences, from which it is clear that the quantity, types and size of stored data files differ depending on individual specialisations and focus. A huge amount of data is produced during scientific research, but not all of it need be stored and archived and this is why authors should pay particular attention to file preparation. Many projects recommend, or directly demand, that the beneficiaries create a Data Management Plan, a document in which they plan and describe what data will be produced during research and how they will manage that data. A page is available to authors on the Library website that concentrates on the organisation of data, with links to guidelines and videos that might inspire them[15].

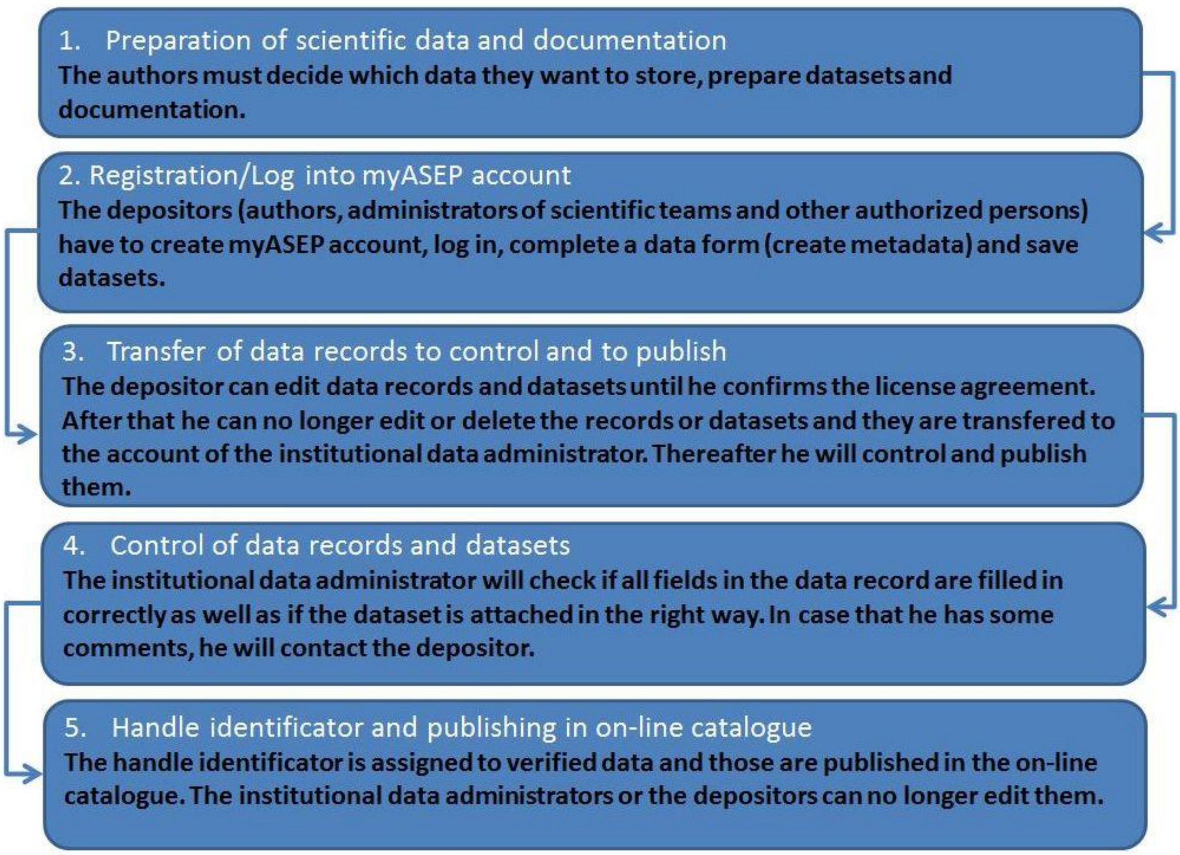
## Workflow

The creation of data records and the storage of data sets in the data repository of the CAS follow on from the method of processing to date. Even though it is possible for anyone else to store data in the repository on behalf of the author, which is the common practice in storing bibliographical records and the full texts of documents, we would recommend that the authors themselves be the depositors in the case of data. Filling in metadata forms and saving data sets is a simple matter from the technical perspective. When transferring data records and data sets to data administrators for checking, the depositor confirms that he agrees with the Agreement on the Storage of Data in the ASEP repository[16]. Fundamental requirements: 1. the author must have the necessary rights to store data (the consent of joint authors); 2. sensitive information may not be published (personal numbers, names, telephone numbers, etc.); and 3. a licence for handling data sets must be submitted. The relevant data administrator undertakes a formal check of data records and of stored data sets. If everything is in order, it publishes them in the ASEP online catalogue. The workflow of storing data records with data sets in ASEP is shown in Figure 1.

[14] More about the institutes of the Czech Academy of Sciences: *CAS institutions* [online]. CAS, ©2017 [cit. 26.9.2017]. Available from: **http://www.avcr.cz/cs/o-nas/struktura/pracoviste-av/**

[15] Knihovna AV ČR, v. v. i. ASEP. *Data preparation* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 31.10.2017]. Available from: **https://www.lib.cas.cz/asep/pro-autory/priprava-dat/**

[16] *Agreement on the Storage of Data in ASEP* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 26.9.2017]. Available from: **https://www.lib.cas.cz/podpora/data/asep/drasep/dohoda_vkladatel.pdf**



1. Preparation of scientific data and documentation
The authors must decide which data they want to store, prepare datasets and documentation.

2. Registration/Log into myASEP account
The depositors (authors, administrators of scientific teams and other authorized persons) have to create myASEP account, log in, complete a data form (create metadata) and save datasets.

3. Transfer of data records to control and to publish
The depositor can edit data records and datasets until he confirms the license agreement. After that he can no longer edit or delete the records or datasets and they are transfered to the account of the institutional data administrator. Thereafter he will control and publish them.

4. Control of data records and datasets
The institutional data administrator will check if all fields in the data record are filled in correctly as well as if the dataset is attached in the right way. In case that he has some comments, he will contact the depositor.

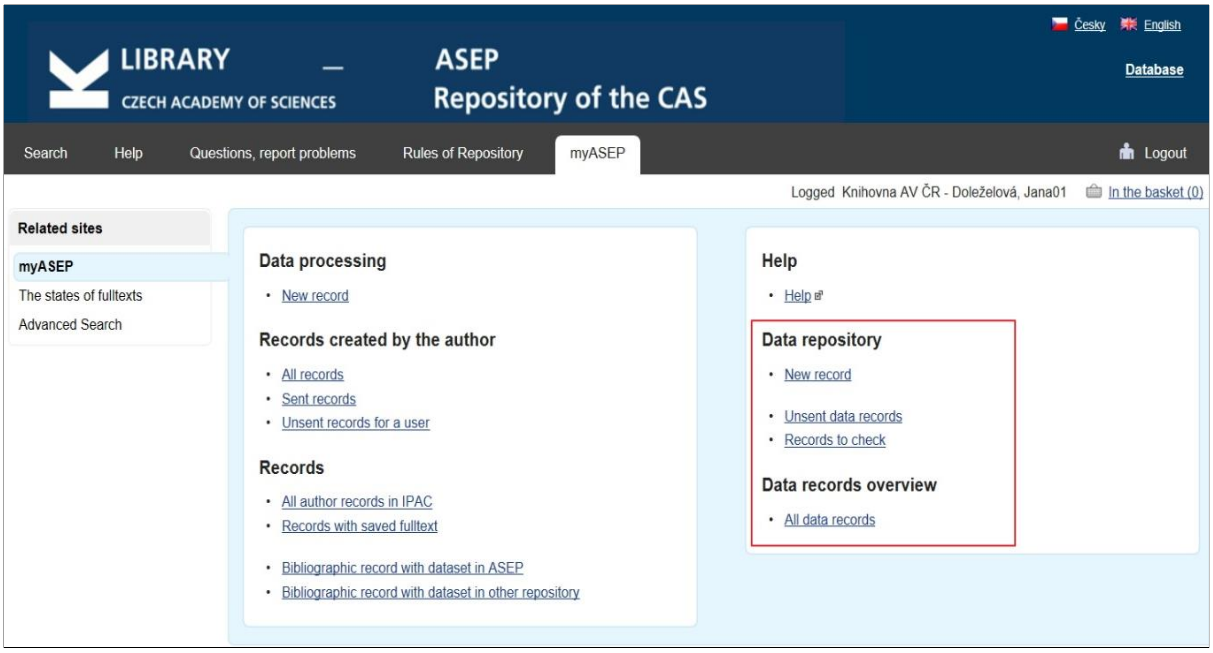
5. Handle identificator and publishing in on-line catalogue
The handle identificator is assigned to verified data and those are published in the on-line catalogue. The institutional data administrators or the depositors can no longer edit them.

Figure 1: Dataset storage workflow into the ASEP repository

## User environment - myASEP

The depositors (authors) and the administrators of the system have their own myASEP user account, from which they manage their data. Figure 2 shows the ASEP user environment for depositors. After logging in, they are able to work with their records, meaning enter new bibliographical records and citations, store the full texts of documents and reviews (the left-hand side of myASEP) and create new data records with data sets (the right-hand side of myASEP). The depositor has an overview of all records that are being processed, prepared for approval and approved and published in the online catalogue. The user account of the system administrators looks similar, but other links and functions are added, in the case of data records a link to records which depositors have submitted for checking and publication. Detailed instructions for use are available to authors and system administrators alike at the Library website[17].

[17] Knihovna AV ČR, v. v. i. ASEP. *For authors* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 31.10.2017]. Available from: https://www.lib.cas.cz/asep/pro-autory/



Logged Knihovna AV ČR - Doleželová, Jana01 In the basket (0)

Related sites

- myASEP
- The states of fulltexts
- Advanced Search

Data processing

- New record

Records created by the author

- All records
- Sent records
- Unsent records for a user

Records

- All author records in IPAC
- Records with saved fulltext
- Bibliographic record with dataset in ASEP
- Bibliographic record with dataset in other repository

Help

- Help

Data repository

- New record
- Unsent data records
- Records to check

Data records overview

- All data records

Figure 2: myASEP account - author

## Data records and data sets

In selecting a metadata set for the ASEP repository, we tried to ensure the maximum possible completeness of data without placing too much of a load on the scientists authors that create the metadata. We also considered individual fields according to the requirements of other systems such that cooperation would be possible in the future (for example, Data Citation Index, OpenAIRE[18]). We drew on the requirements placed on a data repository: metadata in English, information about financing – statement of projects, links to publications and other output relating to the data, description from the content and technical perspectives, statement of scientific disciplines and key words, determination of time and location. Metadata for data records is entered in an online form in which each field is provided with instructions, so that the depositor knows how to enter data in the field. Fields which are mandatory are highlighted in the form, in that it is not possible to publish a data record in the online catalogue without filling in these fields. The current metadata structure is published at the Library website.[19]

Mandatory fields include author statements, title of the dataset, stored file description, data set type, documentation language, keywords, license settings, and file access. When entering authors, we use the authority base, which enables an unambiguous identification of the author, his/her output and affiliation. Major emphasis is placed on the choice of an apt title for the data set and a description of the file/files in Czech and in English. If a longer description is required, we recommend attaching a readme.txt text file to the data set and to provide further detailed information there. The depositor determines and subsequently sets a Creative Commons licence for the item entered, or chooses his own licence, the wording of which he saves in relation to the data set. The choice of licence is entirely a matter for the author and we do not

[18] *OpenAIRE* [online]. Most recent update 22.9.2017 [cit. 26.9.2017]. Available from: https://www.openaire.eu/

[19] Knihovna AV ČR, v. v. i. ASEP. *Description of field – data* [online]. Knihovna AV ČR, v. v. i., ©2017 [cit. 31.10.2017]. Available from: https://www.lib.cas.cz/asep/pro-zpracovatele/manual/popis-poli-data/

recommend, but simply offer the choice of a CC licence, which we expect authors to use. If the depositor chooses open access or open access with time embargo, the data sets become accessible immediately after publication, or after the passing of the time embargo. If access on request is set, the user must request the data set from the author.

The structure of data in ASEP has been based on international standards since the very outset. We use library standard UNIMARC for the storage of data, ISO 369 for language coding, ISO 3166 for country coding and Unicode (UTF-8) for symbol coding. First name, surname and institution are accompanied in sets of authorities of authors with identifiers of the Web of Science (RID) and SCOPUS (AIS) systems and the ORCID identifier[20]. Authorities of projects are provided with numbers from the code list of the Central Register of Projects of the Czech Republic (CEP), the code list of European Commission projects (CORDIS) and the code lists of CAS programmes. Subject classification corresponds to the newly-created mapping of specialisations in the Information Register of R&D Results (RIV)[21]. Each data record has a unique HANDLE[22] identifier assigned to it, an OAI-PMH and the Dublin Core DCMI metadata standard are integrated.

The concept of a data set in ASEP entails a set of files that might contain research data, documentation in which there is important information for users, and perhaps the wording of a licence if the Creative Commons licence is insufficient and the user chooses a different licence. The maximum size of one saved file is 2 GB and the total maximum size of saved files for one data set is 20 GB. Larger files can also be saved subject to agreement with the repository administrator. When choosing the format of files, we recommend using the standard open formats that are supported by various systems and programmes and whose long-term protection is ensured. For text files, for example, we recommend txt, pdf, html or csv, for images jpeg, tiff or png and for media mp3, etc. We are aware that these formats might not be sufficient because different specialisations need to store data in formats that are better suited to their data and are tried and trusted in practice by a certain community.

## Links in records

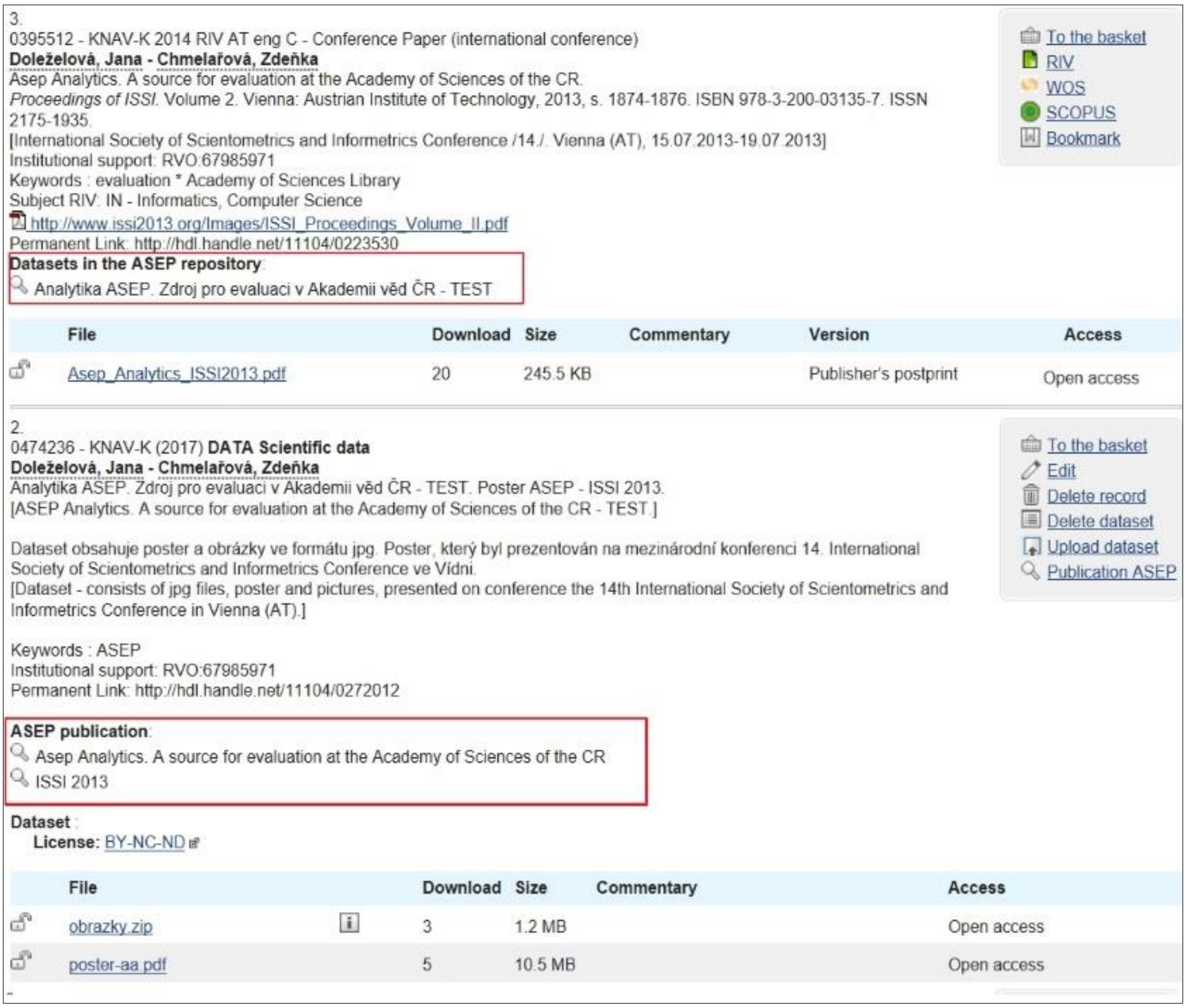
Links to publications and other scientific results (patents, applied research) that relate to the data can be entered in a data record and in the same way links to data records can be entered in bibliographical records. Figure 3 shows the interconnection of data and bibliographical records in ASEP. A data file may be attached to a data record (we favour this method), but we also make it possible for authors who have their data in, for example, an area-specific repository to create only a data record in ASEP with a link to the other repository or storage site. This might be useful in the case that such a repository does not allow the entry of metadata in the required format or to the required extent. Data is cited in much the same way as are bibliographical records, although this is not yet entirely common. Information on how the relevant data set is to be cited is available for each data record. Bibliographical citation in the ASEP database is governed by ČSN ISO 690 standard. For unspecified sources, including data files, the standard provides general rules of citation. Different citation styles and practices are used in data repositories and there is no uniform approach. We plan to introduce

[20] ORCID: **https://orcid.org/**

[21] Office of the Government of the Czech Republic. *R&D Information System 2.0* [online]. Prague: Office of the Government of the Czech Republic, ©2016-2017. Most recent update 19.9.2017 [cit. 26.9.2017]. Available from: **https://www.rvvi.cz/**

[22] HANDLE: **https://www.handle.net/**

the Citace.com service in the future to provide a number of options of how to cite data (APA, Harvard, Chicago, etc.).



3.
0395512 - KNAV-K 2014 RIV AT eng C - Conference Paper (international conference)
**Doleželová, Jana - Chmelařová, Zdeňka**
Asep Analytics. A source for evaluation at the Academy of Sciences of the CR.
*Proceedings of ISSI.* Volume 2. Vienna: Austrian Institute of Technology, 2013, s. 1874-1876. ISBN 978-3-200-03135-7. ISSN 2175-1935.
[International Society of Scientometrics and Informetrics Conference /14./. Vienna (AT), 15.07.2013-19.07.2013]
Institutional support: RVO:67985971
Keywords : evaluation * Academy of Sciences Library
Subject RIV: IN - Informatics, Computer Science
http://www.issi2013.org/Images/ISSI_Proceedings_Volume_II.pdf
Permanent Link: http://hdl.handle.net/11104/0223530
**Datasets in the ASEP repository**:
Analytika ASEP. Zdroj pro evaluaci v Akademii věd ČR - TEST

| File | Download | Size | Commentary | Version | Access |
|---|---|---|---|---|---|
| Asep_Analytics_ISSI2013.pdf | 20 | 245.5 KB | | Publisher's postprint | Open access |

2.
0474236 - KNAV-K (2017) **DATA Scientific data**
**Doleželová, Jana - Chmelařová, Zdeňka**
Analytika ASEP. Zdroj pro evaluaci v Akademii věd ČR - TEST. Poster ASEP - ISSI 2013.
[ASEP Analytics. A source for evaluation at the Academy of Sciences of the CR - TEST.]

Dataset obsahuje poster a obrázky ve formátu jpg. Poster, který byl prezentován na mezinárodní konferenci 14. International Society of Scientometrics and Informetrics Conference ve Vídni.
[Dataset - consists of jpg files, poster and pictures, presented on conference the 14th International Society of Scientometrics and Informetrics Conference in Vienna (AT).]

Keywords : ASEP
Institutional support: RVO:67985971
Permanent Link: http://hdl.handle.net/11104/0272012

**ASEP publication**:
Asep Analytics. A source for evaluation at the Academy of Sciences of the CR
ISSI 2013

**Dataset** :
**License:** BY-NC-ND

| File | | Download | Size | Commentary | Access |
|---|---|---|---|---|---|
| obrazky.zip | i | 3 | 1.2 MB | | Open access |
| poster-aa.pdf | | 5 | 10.5 MB | | Open access |

Figure 3: The example of link from data record to publication record

## Outlook

- The bibliographical records that have a full text stored in ASEP are regularly harvested for the OpenAIRE international database using OAI-PMH and in the future we are also counting on transferring data records.
- We want to include the data repository in the Re3d register of scientific data repositories.
- Another issue we wish to concentrate on is that of large data files, their storage and long-term protection using, for example, the CESNET[23] storage site. We can take inspiration from the Lindat/Clarin repository, which is also designed for storing large data sets, and archiving large language data is ensured in cooperation with CESNET[24].

[23] ANTOŠ, David. *Způsoby využití datových úložišť CESNET aneb čekání na velká data* [online]. CESNET, 2014 [cit. 2017-9-26]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/CESNET_Datova-uloziste.pdf**
[24] HAJIČ, Jan. *LINDAT/CLARIN* [online]. Brno: Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles' University, 2014. 26.11.2014 [cit. 26.9.2017]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/LINDAT-CLARIN.pdf**

- We will endeavour to comply with the conditions of exporting data records to the Data Citation Index, the database at WOS, which monitors the citation count of research data.
- We would like to obtain certification as a reliable repository. A certificate is not simply a formal document that the repository complies with the required criteria – it is a tool with which to check the proper functioning of the repository.

## Conclusion

A basic data structure is defined in the ASEP data repository that is based on international standards, a system of links is in place between the data and bibliographical records stored in ASEP and between data and bibliographical records stored in other systems. Depositors are able to store data sets provided with metadata or create data sets with respect to data stored at other storage sites. The CAS has an open system that can easily be modified and enlarged as required. The base is in place and we will modify and broaden this according to the practical experiences of users and offer new functions which the Library considers important. We will, in the forthcoming period, familiarise scientists with the system of storing and describing data sets in ASEP and will also listen, so that we are able to find an intersection point between the needs of scientists and the ideas of system administrators.

## References

ANTOŠ, David. *Způsoby využití datových úložišť CESNET aneb čekání na velká data* [online]. Praha: CESNET, 2014 [Accessed 26 September 2017]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/CESNET_Datova-uloziste.pdf**

HAJIČ, Jan. *LINDAT/CLARIN* [online]. Brno: Ústav formální a aplikované lingvistiky MFF UK, 2014. 26.11.2014 [Accessed 26 September 2017]. Available from: **https://www.cesnet.cz/wp-content/uploads/2014/10/LINDAT-CLARIN.pdf**

HRABAL, Jan. Repozitáře vědeckých dat. In: *Knihovna.cz* [online]. Brno: KISK FF MUNI, 2013. 22. 2. 2016 [Accessed 26 September 2017]. Available from: **http://ltp.knihovna.cz/?p=385**

HRABAL, Jan, HRUŠKA, Zdeněk. Úvod do problematiky dlouhodobé ochrany digitálních dokumentů – díl 2. In: *Knihovna.cz* [online]. Brno: KISK FF MUNI, 2013. [Accessed 26 September 2017]. Available from: **http://ltp.knihovna.cz/?p=249**

HRUŠKA, Zdeněk. Audit digitálních repozitářů. *Duha* [online]. 2013, **27**(4) [Accessed 26 September 2017]. ISSN 1804-4255. Available from: **http://duha.mzk.cz/clanky/audit-digitalnich-repozitaru**

JAROLÍMKOVÁ, Adéla. Výzkumná data na Univerzitě Karlově. In: *INFORUM 2017: 23. ročník konference o profesionálních informačních zdrojích, Praha, 30.-31.5.2016* [online]. Praha: AiP, 2016 [Accessed 26 September 2017]. ISSN 1801-2213. Available from: **http://www.inforum.cz/pdf/2017/jarolimkova-adela.pdf**

MUULI, Viktor. *Research Data in Estonia: collecting, storing, availibility: some findings from questionnaire* [online]. Estonian Research Council, 2014. 23.10.2014 [Accessed 26 September 2017]. Available from: **http://dspace.ut.ee/bitstream/handle/10062/44052/RD_questionnaire_eng_muuli_14.pdf?sequence=1&isAllowed=y**

MYŠKA, Matěj, KYNCL, Libor, POLČÁK, Radim a ŠAVELKA, Jaromír. *Veřejné licence v České republice* [online]. Brno: Masarykova univerzita. 2012 [Accessed 26 September 2017]. ISBN: 978-80-263-0344-2. Available from: **https://is.muni.cz/www/102870/Prirucka.pdf**

ROSENTHAL, Colin, BLEKINGE-RASMUSSEN, Asger, HUTAŘ, Jan a kol. *Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER)* [online]. Praha: Národní knihovna ČR, 2009 [Accessed 26 September 2017]. ISBN 978-80-7050-569-4. Available from: **http://www.ndk.cz/platter-cz/Platter.pdf**

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: http://nrgl.techlib.cz/conference/conference-proceedings.

# INSTITUTIONAL RULES AND POLICIES FOR SHARING AND STORING RESEARCH DATA

**Michal Koščík**

**michalkoscik@gmail.com**

**Institute of law and technology, Faculty of Law, Masaryk University**

**Department of public health, Faculty of Medicine, Masaryk University**

## Abstract

The paper aims to provide readers with a practical view on how to adapt the internal policies of research institutions to the upcoming General Data Protection Regulation. Since the Regulation enters force six months after the conference takes place, it can be expected that this issue of readjustment of internal processes to GDPR will be very important for majority of conference participants. With regard to the time and space limit, the paper will focus exclusively on the issues connected with archiving and sharing research data. Emphasis will be put on the rights of research subjects and the public interest in research as an entitlement to process of personal data without consent.

## Key words

Privacy; Compliance; Data Protection; General Data Protection Regulation; GDPR

This article is a result of the research funded by the Czech Science Foundation as the project GA15-20763S Právni rámec sběru, zpracování, uchovávání a užívání výzkumných dat (*Legal Framework* for Collecting, Processing, Storing and Utilizing of Research Data).

# Introduction

There is little point in an extensive introduction and description of the General Data Protection Regulation (hereinafter GDPR) which enters into force on 28 May 2018, as it has already been covered in numerous articles[1]. This article does not intend to provide readers with a detailed checklist of all data protection issues that need to be addressed, but rather provide practical guidance on where to begin and how to proceed in the process of adapting institutional rules and processes for operators that share research data through repositories. We presume that the majority of operators of repositories are research institutions (universities, academies or central authorities) that employ more than 250 employees or process special categories of data and are thus are subjected to the new obligation to keep records on processing activities under Article 30 of the GDPR, which in practice means a significant step towards the strong formalisation of the compliance procedures. Therefore, the main purpose of the article is to outline the steps that need to be taken in order to bring rules of institutions to the appropriate formal level.

The GDPR compliance procedure consists of two major steps. The first requires knowledgeto be gathered about the dataflows within the institution, and the second requires internal policies to be adopted or adjusted. Therefore, this article is divided into two chapters. The first chapter describes the steps that have to be taken in the process of collecting information, and the second chapter suggests ways in which the policies of institutions should be structured.

# Gathering information and identifying personal data in repositories

The first step in compliance procedure is the identification of personal data in repositories and its life cycle. This means both the revision of existing data and its origin (source) as well as the identification of the channels where the data will flow in the future. It is necessary to identify storage spaces, departments or employees who are responsible for the administration of the data and the recipients of the data (i.e. the persons or entities that find the data useful).

### Identification of personal data in all repositories

One has to keep in mind that the definition of "personal data" is very extensive and covers any information that can be directly or indirectly related to an individual. The data does not have to be structured in order to qualify as personal data. Any information in any media format, including photographs, audio and visual records, may meet the definition of personal data and thus make the repository of the institution subject to regulation.

It is important to point out that even pseudonymized information is to be considered personal information. The borderline between pseudonymized and anonymized information might not be exactly clear in many practical situations. Since the definition of personal data is very broad, it can be advised that even anonymized data be handled with great cautionn, if possible under the same standards as if the personal data were involved (see subchapter 1.3. of this article).

[1] See also previous articles of the author of the Article that covered the development in this area in recent years: KOŠČÍK, Michal. Privacy and anonymization in repositories of grey literature. In: Conference on Grey Literature and Repositories. 2015. p. 72. KOŠČÍK, Michal. The Impact of the General Data Protection Regulation on grey literature. Grey Journal (TGJ), 2017, 13. See also: WIPP EKMAN, Leon; BILLGREN, Petter. Compliance Challenges with the General Data Protection Regulation. 2017.

### Identification of the purpose and activities related to data processing

Personal data processing is a daily activity in every public institution or business. The governance of personal data has to be based on the purpose served by the data being processed (i.e. its value to the organization) and on the activities (processes) that involve the particular data. After the personal data has been identified, it is necessary to attribute each set of records to a certain purpose (or purposes) for which they have been collected and processed. To put it simply, the institution has to seriously question each individual database record and answer the question "do we really need to keep this record and why?". Virtually no common purpose of processing is illegitimate per se[2]. After the purpose of processing is identified, it is possible to assess whether the processing is legitimate in this particular case and what steps need to be taken in order to keep the processing legitimate. Keeping personal data without a specific purpose[3] is equivalent to non-compliance with the regulation.

It is necessary to define the purpose of each set of data in order to determine whether or not the institution requires the consent of the data subject. The general regulatory principles of purpose limitation[4], data minimisation[5] and storage limitation[6] are directly related to the purpose of data processing. Hence, if the institution does not define the purpose of each particular set of personal data it processes, it cannot comply with these fundamental principles. The purpose of data processing is also crucial in dealing with requests for data erasure [7] or the right to restriction of processing[8].

Recital 39 of the GDPR states that the purpose needs to be determined at the time when the personal data is collected and that changing the purpose of processing after the data has been collected is limited by the GDPR and restricted to several explicitly defined cases[9]. Operators of repositories will benefit from the provisions of the second paragraph of Art. 9 of the GDPR, which enables so-called "further processing" or secondary use of data for archiving purposes in the public interest, for scientific or historical research purposes or for statistical purposes[10] even in cases where data in special data categories (sensitive data) is being proccesssed.Even if the repository operator intends to rely on the provisions of Art. 9 section 2, the purpose has to be defined. It is advised that the purpose be defined more specifically than by the mere declaration of public or scientific interest so that the proportionality between the public interest and the interest of the subject can be demonstrated.

[2] with the exception of clear excesses, usually well defined in criminal codes

[3] for example storing historical data collected during past activities just because someone failed to delete it or keeping data "just in case"

[4] Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes

[5] Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which it is processed

[6] Personal data shall be kept in a form allowing the identification of data subjects for no longer than is necessary

[7] See also COFONE, Ignacio N. Google v. Spain: A Right To Be Forgotten?. Browser Download This Paper, 2015.; ROSNAY, Melanie Dulong de; GUADAMUZ, Andres. Memory Hole or Right to Delist?. Implications of the Right to be Forgotten for Web Archiving. *RESET. Recherches en sciences sociales sur Internet*, 2016, 6. KOŠČÍK, Michal. The Impact of the General Data Protection Regulation on grey literature. Grey Journal (TGJ), 2017, 13.

[8] The data subject shall have the right to obtain from the controller restriction of processing the controller no longer needs the personal data for the purposes of the processing.

[9] One of the reasons may be protecting the vital interest of a data subject, or vital interest of another natural person or archiving. Here, it has to be noted that even the change in purpose of processing personal data collected for the public interest is limited.

[10] The national law could however specify requirements for link between those purposes and the purposes of the intended further processing - See recital 50 and

After the institution identifies the data and its purpose, it has to identify the activities in which it is necessary to process the particular personal data. Each activity in which the personal data needs to be processed shall have a delegated person who is responsible for compliance with internal rules and policies (see below). These persons are not necessarily (and most likely not) data protection officers, as the data protection officer is more the role of the internal auditor and not the person who will perform all the tasks associated with data protection.

### Identification of the data sources

Every repository needs to identify sources from which it retrieves personal data, mainly for three compliance reasons:

**A. Identifying whether the repository is a controller or processor.** It should be noted that the repository is rarely established as a mere processing service without any interest of its operator in collecting data and determining what goes into the repository. We presume that the repository will be a controller of most of its data. In cases where the repository serves as a data processor, we strongly recommend that the the contractual framework be reviewed with the data contollers[11].

### B. Determining whether the repository processes raw, pseudonymized or anonymized data

The first practical issue with anonymized data is the question of whether a repository storing data obtained from a third party which the repository's operator cannot himself attribute to an individual natural person is to be considered as anonymous or pseudonymous data if the subject that encrypted the data still keeps the key to its decryption. According to Article 4 of the GDPR, pseudonymized data is defined as "personal data that can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures for ensuring that the personal data is not attributed to an identified or identifiable natural person".

The first possible approach is to admit that the anonymity or pseudonymity of data is relative. Two (or more) subjects can process the same set of data, whereas one is unable to encrypt it and the other one is able to encrypt it. If we accept that the concept of data anonymity is relative, it would mean that the first subject can use and share data freely without any significant restrictions, whereas the person that possesses the encryption/ decryption key is restricted in handling the data. The second approach is to presume that the anonymity of the data is absolute. If the encryption key exists anywhere in the world or can be deciphered in any way, such data is not anonymous but only pseudonymous. One of the main problems of pseudonymized personal data is that (if shared) it can potentially be de-anonymized by a third party when merged with other data-sets[12], and basically any anonymized data can be de-anonymized by forensic methods. The CJEU addressed this issue in the judgment in Case C-582/14: Patrick Breyer v Bundesrepublik Deutschland, in which the CJEU ruled that the possibility to combine the data with this additional data must constitute a means of which it can

[11] Detailed formal requirements on a contract between controller and processor are described in the Article 28 GDPR.

[12] HARAŠTA, Jakub a Matěj MYŠKA. Secondary use of research data in the EU: Complex institutional approach. In Erich Schweighofer; Franz Kummer; Walter Hötzendorfer; Christoph Sorge. Trends and Communities of Legal Informatics IRIS 2017 Proceedings of the 20th International Legal Informatics Symposion. Wien: Oesterreichische Computer Gesellschaft, 2017. s. 539-542, 4 s. ISBN 978-3-903035-15-7.

reasonably be assumed that it will likely be used to identify the individual[13]. The interpretation of the Breyer case speaks in favour of the "relative approach". The data is not anonymous for a person that has a legal and material capacity to de-anonymize it. We can add that data may remain anonymous/anonymized for entities that lack legal and material capacities to de-cipher it. This approach is favourable for data repositories, since they can share anonymized research data with a certain degree of legal certainty.

**C. Determining whether the source collects data in accordance with applicable rules.** If the data processing requires the consent of the subject, the repository operator needs to make sure that the copy of the consent can be found at the source.

## Adopting policies and internal rules

Only after the data and its sources and purpose have been defined is it possible to adapt the internal policies and formalize them into internal documents. Below, we have identified the key areas that have to be addressed by internal policies

### General data protection policy addressing privacy by design and default

The repository should adopt a norm which will address risks, responsibilities and measures as they regard the security accessibility, pseudonymization and anonymization of data and the identification of processes, activities and involved employees[14]. If the repository shares data based on health information or other sensitive data (such as the ethnic origin or political stances of research subjects), it will likely be obliged to carry out a privacy impact assessment under Article 35 of the GDPR[15]. Even in institutions where the privacy impact assessment is not required, it is recommended to identify the major risks to the rights of data subjects and identify organizational units that are required to take measures to protect these rights.

The norm should also implement a notification policy for cases of personal data breaches. The GDPR requires response and notification of the data protection authority within the 72 hour time limit. It is, therefore, advisable to have defined responsibilities for notification of data breaches in advance.

We presume that the majority of institutional repositories will also fall under the obligation under Article 30 of the GDPR that obliges each controller to keep a record of processing activities for which he is responsible. The record must contain all of the following information:

[13] NIEMANN, Fabian a Lennart Schüßler CJEU decision on dynamic IP addresses touches fundamental DP law questions. Bird & Bird [online] [vid. 2017-10-09]. Available from: https://www.twobirds.com/en/news/articles/2016/global/cjeu-decision-on-dynamic-ip-addresses-touches-fundamental-dp-law-questions See also EL KHOURY, Alessandro. Dynamic IP Addresses Can be Personal Data, Sometimes. A Story of Binary Relations and Schrödinger's Cat. *European Journal of Risk Regulation*, 2017, 8.1: 191-197., POLČÁK, Radim. Stock Exchange Interconnections and Legal Issues in Data Exchange. *Masaryk University Journal of Law and Technology*, 2017, 11.2: 351-362.

[14] See also: GJERMUNDRØD, Harald; DIONYSIOU, Ioanna; COSTA, Kyriakos. privacyTracker: A Privacy-by-Design GDPR-Compliant Framework with Verifiable Data Traceability Controls. In: International Conference on Web Engineering. Springer International Publishing, 2016. p. 3-15.

[15] Accoriding to Art. 35 GDPR, the privacy impact assessment is required if the processing is „ likely to result in a high risk to the rights and freedoms of natural persons, see also BIEKER, Felix, et al. A process for data protection impact assessment under the European general data protection regulation. In: *Annual Privacy Forum*. Springer International Publishing, 2016. p. 21-37.

- the name and contact details of the controller
- the categories of data subjects and categories of personal data;
- the categories of recipients to whom the personal data has been or will be disclosed, including recipients in third countries or international organisations;
- transfers of personal data to a third country or an international organisation, including the identification of that third country or international organisation and the documentation of suitable safeguards;
- the expected time limits for the erasure of the different data;
- description of the technical and organisational security measures referred to.

We strongly advise that obligations be formulated and record keeping be delegated to the respective departments and employees. Data protection by design and default should become a formal responsibility of every employee of an institution that has access or the right to upload content to a repository[16].

### Privacy (transparency) policy

Section 2 of the GDPR, which deals with information and access to personal data, sets forth requirements for the information that has to be provided to the subject. It is recommended that a document containing the basic information that has to be provided to data subjects under Art. 13 to 15 of the GDPR be drafted and published. Among others, this information includes:

- the contact details of the controller and the controller’s representative;
- the contact details of the data protection officer;
- the purposes for which the personal data is processed as well as the legal basis for the processing;
- the legitimate interests pursued by the controller or by a third party and the recipients or categories of recipients of the personal data;
- the period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period;
- the existence of the right to request from the controller access to and rectification or erasure of personal data or a restriction of processing concerning
- information regarding the existence of the right to withdraw consent[17]

It is evident that the recipients of the general policy (under Art. 2.1.) are the employees of the institution, whereas the privacy policy is a non-binding informative document for data subjects. The language and scope of detail should be adjusted accordingly. The privacy policy is supposed to be relatively short and approachable, whereas the general data protection policy will be more detailed and could be further specified by technical norms applicable to the respective divisions of an institution. Most academic institutions will likely have a general policy covering all the major data protection issues and might also adopt a separate policy governing data protection issues in a respective repository. This approach is recommended for larger institutions that operate several repositories which store data from distinctive research fields and serve different purposes.

[16] See also KOŠČÍK, Michal. Sharing Liability for a Repository Between Employer and Employee. In: *CONFERENCE ON GREY LITERATURE AND REPOSITORIES*. 2016. p. 69.

[17] The list is not exhaustive, the author has selected information that is most likely to be relevant for a repository

## Conclusion

The article outlined two phases of the procedure for compliance with the GDPR at public institutions that operate a repository. We suggest that the institution needs to identify its data and processes and link the data to the processes (and thus define their purpose) before drafting new rules and documents. The institution needs to make it clear which set of data is processed in the role of "data controller" and which set of data is processed in the role of "data processor" (in cases where the institution is the data processor, it is also necessary to review the contractual framework with the controllers).

We presume (and also recommend) that most institutions will aim to draft at least two documents - one internal policy that will address most data involving processes in order to comply with the objective of "data protection by design and default" and one publicly available policy document that will provide information about the privacy standards of the institution operating the repository.

## References

BIEKER, Felix, et al. A process for data protection impact assessment under the European general data protection regulation. In: *Annual Privacy Forum*. Springer International Publishing, 2016. p. 21-37.

COFONE, Ignacio N. Google v. Spain: A Right To Be Forgotten?. *Chicago-Kent Journal of International and Comparative Law* [online]. 2015, **15**(1). [Accessed 9 October 2017]. Available from: **https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2548954**

GJERMUNDRØD, Harald, Ioanna DIONYSIOU a Kyriakos COSTA. PrivacyTracker: A Privacy-by-Design GDPR-Compliant Framework with Verifiable Data Traceability Controls. In: *Current Trends in Web Engineering*. Springer International Publishing, 2016, 3 - 15.

HARAŠTA, Jakub a Matěj MYŠKA. Secondary use of research data in the EU: Complex institutional approach. In: SCHWEIGHOFER, Erich, Franz KUMMER, Walter HÖTZENDORFER a Christoph SORGE. *Trends und Communities der Rechtsinformatik / Trends and Communities of Legal Informatics: Tagungsband des 20 Internationalen Rechtsinformatik Symposions IRIS 2017*. Wien: Oesterreichische Computer Gesellschaft, 2017, 539 - 542. ISBN 978-3-903035-15-7.

KOSCIK, Michal. Privacy Issues in Online Service Users' Details Disclosure in the Recent Case-Law: Analysis of Cases Youtube v. Viacom and Promusicae vs. Telefonica. *Masaryk University Journal of Law and Technology* [online]. 2009, 3, p. 259. [Accessed 9 October 2017].

KOŠČÍK, Michal. Privacy and anonymization in repositories of grey literature. The Grey Journal (TGJ). 2015, **11**(special issue).

NIEMANN, Fabian and Lennart SCHÜßLER. CJEU decision on dynamic IP addresses touches fundamental DP law questions. In: *Bird & Bird* [online]. [Accessed 9 October 2017]. Available from: **https://www.twobirds.com/en/news/articles/2016/global/cjeu-decision-on-dynamic-ip-addresses-touches-fundamental-dp-law-questions**

POLČAK, Radim. Getting European data protection off the ground. *International Data Privacy Law* [online]. 2014, **4**(4), 282-289 [Accessed 9 October 2017]. DOI: 10.1093/idpl/ipu019. ISSN 2044-3994. Available from: **https://academic.oup.com/idpl/article-lookup/doi/10.1093/idpl/ipu019**

POLČÁK, Radim. Stock Exchange Interconnections and Legal Issues in Data Exchange. *Masaryk University Journal of Law and Technology* [online]. 2017, **11**(2), 351-362 [Accessed 9 October 2017]. DOI: 10.5817/MUJLT2017-2-7. ISSN 18025943. Available from: **https://journals.muni.cz/mujlt/article/view/6681**

ROSNAY, Melanie Dulong de and Andres GUADAMUZ. Memory Hole or Right to Delist?: Implications of the Right to be Forgotten for Web Archiving. In: *RESET: Recherches en sciences sociales sur Internet* [online]. 2017(6). [Accessed 9 October 2017]. ISSN 2264-6221. Available from: **https://reset.revues.org/807**

WIPP EKMAN, Leon and Petter BILLGREN. *Compliance Challenges with the General Data Protection Regula- tion* [online]. Lund, 2017 [Accessed 9 October 2017]. Available from: **http://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8911983&fileOId=8911995. Master thesis. Lund University**

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

# ORPHAN AND OUT-OF-COMMERCE WORKS AFTER THE AMENDMENT OF THE CZECH COPYRIGHT ACT

**Matěj Myška**

**matej.myska@law.muni.cz**

**Masaryk University, School of Law, Institue of Law and Technology, Czech Republic**

## Abstract

This paper focuses on the changes introduced by the major amendment of the Czech Copyright Act in 2017 in the area of orphan and out-of-commerce works licensing. It offers a brief description and basic black letter law analysis of the new legal regulation, discusses the newly introduced regulatory regimes in the context of the judgment of the Court of Justice of the European Union *Soulier and Doke* and offers some tips on practical functioning thereof.

## Keywords

Orphan Works; Out-of-Commerce Works; Memory and Educational Institutions; Copyright; Extended Collective Licensing.

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 main legislative aim[1] of the last major amendment[2] of the Czech Copyright Act[3] was to implement the Collective Rights Management Directive (32014L0026). However, as the initiator of the amendment, the Czech Ministry of Culture added further important points to the agenda. Apart from other interesting issues, such as the implementation of the exception for parody (Sec. 38g of the CA), the Amendment also broadened the possibilities of orphan works licensing (further referred to as "OW") and introduced a legal framework for licensing of out-of-commerce works (further referred to as "OCW"). The aim of this short paper is to present and critically analyse the new legislation. Next, it aims to discuss and evaluate it within the context of the recent Court of Justice of the European Union (further referred to as the "CJEU") judgment in the case *Soulier and Doke*.[4] Finally, the paper aims to formulate some practical operational principles that should ensure the legal certainty of the subjects involved. As regards the scope of the paper, it must be noted that it focuses mainly on the newly adopted regulation introduced by the Amendment.[5]

## Orphan works

The Orphan Works Directive (32012L0028) was implemented in Czech law already in 2014.[6] The system for the use of OW[7] that was established at the time could be described as minimal – only memory and educational institutions[8] and public service broadcasters could make use of the statutory exception, for specific uses and only for non-profit purposes based on the statutory exception (Sec. 37a of the CA) and for achieving aims related to their public-interest missions. As such, the use was gratuitous and payment of remuneration was due only when the OW status was ended by the author and should be paid retrospectively. Pursuant to Sec. 37a of the CA, the amount of the remuneration was to be determined *"based on the purpose for which and circumstances in which the work is used, as well as the extent of the damage incurred on the author by such use".* Other subjects or other types of use or use for

[1] See the Explanatory Memorandum to the Act No. 102/2017 Coll., Amending Act No. 121/2000 Coll., on copyright, on rights related to copyright and on the amendment of certain acts (Copyright Act), as amended – in the Parliamentary press no. 724/00, p. 46 (further referred to as "ER").

[2] The Act No. 102/2017 Coll., Amending Act No. 121/2000 Coll., on copyright, on rights related to copyright and on the amendment of certain acts (Copyright Act), as amended (further referred to as the "Amendment").

[3] Act No. 121/2000 Coll., on copyright, on rights related to copyright and on the amendment of certain acts (Copyright Act), as amended (further referred to as the "CA").

[4] C-301/15, EU:C:2016:878.

[5] The functioning of the previous regulation in detail as well as the basic underlying concepts of the issue of orphan works has already been thoroughly described and analysed by others (See e.g. (Bertoni, Guerrieri, Montagnani 2017, pp. 41–51; Mackovičová 2016; Prchal 2013).

[6] Act No. 228/2014 Coll., amending Act No. 121/2000 Coll., on copyright, on rights related to copyright and on the amendment of certain acts (Copyright Act), as amended, and Act No. 151/1997 Coll., on Property Valuation and on the amendment of certain acts (Act on Property Valuation), as amended.

[7] The term "orphan work" includes pursuant to the Sec. 37a of the CA not only works published in the form of books, journals, newspapers, magazines or other writings, but also cinematographic or audiovisual works. Due to the referral provisions (Sec. 74, 78 and 82 of the CA), the exception for orphan works shall apply, by analogy, to the performer and his performances (Sec. 74 CA), to the phonogram producer and his phonogram (Sec. 78 of the CA) and to the producer of the audiovisual fixation and to his fixation (Sec. 80 of the CA).

[8] The term "memory and educational institution" is used as a general term for library, archive, museum, gallery, school, university and other non-profit school-related and educational establishment as used in Sec. 37 of the CA.

commercial purposes were not exempted, and therefore, the license to use OW could be granted only within the system of mandatory or extended collective licensing (hereinafter referred to as “ECL”).

The main purpose for updating the OW regulation was to allow their broader use – namely by means and subjects other than those already stipulated under the exception or in the regimes of the mandatory or already-existing extended collective licensing. The Czech legislator made use of and relied on recital 24 and Art. 1 para. 5 of the Orphan Works Directive and introduced the possibility for anyone to use the OW (i.e. not only memory and educational institutions) for any purpose, and in particular, for any type of use. The Hungarian, Canadian and U.K. regulations are mentioned as an inspiration.[9]

Therefore, the user (or potential licensee) does not have to rely only on the institute of mandatory and/or extended collective licensing of rights and can acquire a license also for uses that are not covered by these institutes. The respective CMO that is entitled to collectively manage rights to certain types of works shall therefore newly serve as a one-stop-shop[10] for any licensee who is interested in using OW of that type (Sec. 103 of the CA). However, the user must first perform a diligent search[11] before negotiating the license with the CMO, who acts as the legal representative of the author and must pay a license fee for the use. The scope of the license is however limited by law to the territory of Czech Republic and can be granted only for the duration of five years (Sec. 103 para. 3 of the CA).[12] The respective license fee is kept by the CMO for three years and must be paid to the right holder if the status of OW is ended. If this does not happen, the CMO is obligated under the Sec. 103 para. 5 of the CA to transfer the fees to the State Fund for Culture, or the State Cinematography Fund, in the case of orphaned audiovisual works and works used in audiovisual works. However, the end of the OW status does not affect the validity of the license granted by the CMO (Sec. 103 para. 6 of the CA).

Lastly, the Amendment introduced a change regarding the end of the status of OW.[13] Newly, Sec. 27a of the CA entitles the author to end the status of OW directly by informing the user thereof. This provision aims at the situation when the memory and educational institution relies on the statutory exception regulated under Sec. 37a of the CA (i.e. the standard and “old” regulation of OW). After receiving such a notice from the author, which could be submitted also in electronic form,[14] the user is obligated to inform the respective CMO. Furthermore, the author may end the status of OW at any time by notifying the respective CMO, even if the OW is not currently being used pursuant Sec. 37a of the CA. In practice, this last option means that the author could at any time ask the CMO to remove the Work from the Registry of OW[15] that these CMOs are obligated to keep. In order to fulfil the transparency principle, the CMOs are required to keep the Registry updated and available online on their webpages (Sec. 99f para. 1 let. l) of the CA).

[9] ER, p. 132.

[10] The CMO acts in his own name on the account of the rightholder, i.e. as mandatary.

[11] Sec. 27b of the CA referring to Annex 2 of the CA, which contains the source that must be searched.

[12] However, the user may conclude the license agreement repeatedly.

[13] ER, p. 96.

[14] Sec. 562 para. 1 of the Civil Code (Act 89/2012 Coll., Civil Code, as amended).

[15] Every CMO operates the Registry for the type of works to which it is entitled to exercise the collective management and which are known to the CMO from its own activity (Sec. 27a para. 4 and Sec. 97c para. 2 let. b) of the CA).

In other, specifically unregulated issued, the standard rules (i.e. the Head IV of the CA) relating to CMOs should apply mutatis mutandis in the case of OW licensing in the new ECL regime (pursuant to Sec. 103 para. 7). However, this does not include the handling of incomes from the use of OW under the "standard" (i.e. other ECL uses or uses under mandatory collective licensing), as the regulation in Sec. 103 para. 4 of the CA is to be regarded as *lex specialis* to the Sec. 99c of the CA. In practice, this means that the CMOs have to carefully separate the fees on the basis of the legal ground on which the OW have been used.[16]

## Out-of-commerce works

The complex regulation of the use of OCW is a novelty previously unknown to Czech copyright law. However, such regulation is not revolutionary, as similar licensing systems are already functioning in the Scandinavian countries, the UK and Germany.[17] Ironically, the ER[18] also mentions the French regulation that was later evaluated as non-compliant with the Information Society Directive (32001L0029) as one of the up-and-running systems of OCW ECL.[19]

Before the Amendment, the simple fact that the work is not available in traditional distribution channels was not reflected in any way, and such works had to be treated as any other work. Consequently, such works could be used only on the basis of a copyright exception or limitation (Sec. 30–39 of the CA), or the user could rely on the institute of mandatory[20] or extended collective licensing.[21] The memory and educational institutions relied especially on the "library" exception (Sec. 37 of the CA) that allowed limited copying, lending and making works available "on-site". Such a stringent regulation did not allow the full use of digitalized work that is not normally available on the market.

Inspired by the Memorandum of Understanding "Key Principles on the Digitisation and Making Available of Out-of-Commerce Works" (further referred to as the "Memorandum") and the Commission Recommendation of 2006,[22] the Amendment subjected the copying of OCW and its communication to the public to the regime of extended collective licensing (Sec. 97e para. 4 let. i) of the CA). Consequently, an OCW could be reproduced and made available to the respective individual also via a computer network[23] by a library[24] upon the payment of the license fee for the maximum term of five years.[25]

[16] The CMOs also have the obligation to keep the incomes obtained from "standard" ECL and mandatory collective licensing separate (Sec. 99c para. 3 of the CA).

[17] These countries are explicitly mentioned by the Czech legislator as a source of inspiration on p. 120 of the ER. For a detailed overview of the respective national regimes, see the respective references cited in footnotes 13–24 in (Gera 2017, p. 262).

[18] ER, p. 120.

[19] The proposal of the CA was sent to the Members of the Parliament on 17. 2. 2016. The *Soulier and Doke* case was decided on 16. 11. 2016.

[20] Sec. 95 of the CA before the Amendment.

[21] Sec. 101 para. 9 of the CA before the Amendment.

[22] European Commission, Commission Recommendation of 24 August 2006, on the digitisation and online accessibility of cultural material and digital preservation (2006/585/EC) (OJ L 236/28, 31 August 2006).

[23] The making available must therefore be controlled to the extent to which the library is able to ascertain the number of views. The licensing fee should also take into account the amount of views of the OCW (Richter 2017).

[24] As defined in the Library Act (Act No. 257/2001 Coll., on Libraries and of Conditions for the Operation of Public Library and Information Services (Library Act), as amended). According to Sec. 2, a "library" is a *"facility in which all public library and information services are provided for everyone on a non-discriminatory basis and which is registered in the register of libraries".*

[25] The duration of the license might be prolonged repeatedly.

Contrary to the Memorandum,[26] the CA does not contain an exact and explicit definition of OCW. Its constituting elements must therefore be deduced from the required conditions that must be fulfilled so that the work is eligible for inclusion into the List of OCW (further referred to as the "List"). Firstly, the OCW might be only works expressed in words ("verbal" works). It, therefore, seems that the precise classification as regards the category of the work, i.e. whether the work is literal, scientific or a work of art (Sec. 2 of the CA), is not important. Only the expressive means used are important. Consequently, OCW might include also scientific works or dramatic works, but not sole photographic works, audiovisual works, computer programs or sculptural works, for example. However, works embedded or incorporated in the "verbal" works that are an integral part of it might also be included on the List and consequently used in accordance with the law. Next, the work must be (logically) "out-of-commerce", i.e. not obtainable through general business channels. According to Sec. 97f para 3. let. a) of the CA, this condition is fulfilled when the work in the same or a similar format is not obtainable with reasonable efforts and under normal conditions and upon payment in the normal business network within 6 months from the date of receipt of the proposal for inclusion in the List. Furthermore, under Sec. 97f para. 3 let. b) of the CA, the OCW must also not be subject to licensing conditions or terms and conditions of sale excluding inclusion in the List. As a result, periodicals available from licensed databases as well as e-books with prohibitive terms and conditions are not considered as OCW.

The work is listed only based on proposal from the right holder, the library or the respective CMO that is to be made publicly available on the web pages of the National Library (Sec. 97f para. 2 of the CA). In accordance with the second principle of the Memorandum,[27] the right holder (i.e. not only the author but also, for example, her heir or the publisher if he is entitled to do so) may opt out from the regime of OCW and withdraw her work from the List during this timeframe. However, the realization of such a right by the right holder does not affect the validity of the license granted to the library by the respective CMO (Sec. 97f para. 5 of the CA). Periodicals have a specific regime (Sec. 97f para. 4 of the CA), as they might be included in the List by the National Library provided that they had been published ten and more years ago and they are not subject to the terms and conditions excluding such inclusion.

## Soulier and Doke case and the Czech regulation

The new Czech regulation set up a relatively sophisticated ECL system for OW and OCW. However, during the legislative process, the CJEU issued its decision in the seminal case *Soulier and Doke*. To put it very simply, this judgment imposed quite strict conditions on the prior implicit consent of the author to use of a work, including use under an extended collective license.[28] The CJEU explicitly stated that *"every author must actually be informed of the future use of his work by a third party and the means at his disposal to prohibit it if he so wishes."*[29] Furthermore, it concluded that a mere lack of opposition on the part of the authors cannot be interpreted as an implicit consent to use the work under extended collective licensing.[30] Therefore, an implicit consent might be possible, but only if the author has effective knowledge

[26] According to the Memorandum (p. 2), a work is out of commerce *"when the whole work in all its versions and manifestations is no longer commercially available in customary channels of commerce, regardless of whether tangible copies of the work* exist *in libraries and among the public (including through second hand bookshops or antiquarian bookshops)."*

[27] Para. 5 of Principle No. 2 of the Memorandum.

[28] Para. 37–38 *Soulier and Doke* (C-301/15, EU:C:2016:878).

[29] Para. 38 *Soulier and Doke* (C-301/15, EU:C:2016:878).

[30] Para. 43 *Soulier and Doke* (C-301/15, EU:C:2016:878).

of the future potential uses of the work and has means to stop it, *"without having to depend […] on the concurrent will of persons".*[31] The crucial conclusions of the CJEU could be found in para. 45 of the decision, where it acknowledged the worthiness of pursuing the *"cultural interest of consumers and of society as a whole"*.[32] In the same paragraph, the CJEU stated that the public interest in making the OCW available to the broad public cannot be justified if the derogation from the author's rights is not provided for by the EU legislature (i.e. the appropriate exception as in the case of OW).[33] Sganga (2017, p. 330) rightly argues that the *"requirement to inform individually each and every author about the existence and functioning of the scheme"* directly opposes the "catch-all" nature of ECL and essentially renders it futile. The contemporary jurisprudence further expresses valid doubts about the proper meaning and implications of the *Soulier and Doke* case. Suthersanen, for example, claims that the ECL licensing mechanism may not be possible (2017, p. 382). Borghi, Erickson and Favale (2016, p. 147) are more cautious and express doubts about the full impact of this decision on ECL systems throughout Europe. Sganga (2017, p. 330) criticizes the CJEU for creating *"further uncertainties"* and opening *"the gate for a potential flow of complaints against national collective management schemes"*. The main reason being the unclear conceptualization of the term "derogation". Namely, in the *Vereniging Openbare Bibliothekencase*,[34] the derogation from public lending right was treated as an "exception" (Gera 2017, n. 40) and therefore basically compliant.

In the light of the above-mentioned, the Czech regulation of both OW and OCW ECL must be evaluated. In general, the OW ECL has a higher chance of passing the test, as it shares its basic features (and therefore protective elements) with the exception-based system of OW (Sec. 27a and 27b of the CA). In order to be able to make use of the ECL, the potential licensee must perform the diligent search before requesting a license from the CMO. On the other hand, this ECL is not based on an EU-legislature exception. The OCW licensing system is problematic especially in the case of periodicals, where the simplified procedure does not involve the prior provision of information about the proposal to list the work. For both of the ECL systems, the protection of *iura questia*, i.e. that the granted license is still valid even after the OW/OCW status has been ended, might be in conflict with the conclusions of the CJEU presented above.

Even though the Czech legislator characterized the system of OCW licensing as a *"carefully balanced compromise for balancing the interests"*[35] of right holders and the public, the CJEU would probably not be of the same opinion. The main reason is that the system lacks the *"guarantees ensuring that authors are actually informed as to the envisaged use of their works and the means at their disposal to prohibit it"*,[36] especially in the case of OCW ECL for periodicals.

[31] Para. 49 *Soulier and Doke* (C-301/15, EU:C:2016:878).

[32] During the consultations preceding the legislative process, the National Library mentioned the enabling of *"broad public access to cultural heritage in digital form to support education and science and personal development"* as the main purpose for the introduction of the OCW licensing (ER, p. 78). This reasoning was consequently also adopted by the legislator (ER, p. 120).

[33] Para. 45 *Soulier and Doke* (C-301/15, EU:C:2016:878).

[34] Para. 50–51*, Vereniging Openbare Bibliotheken* (C-174/15, EU:C:2016:856).

[35] ER, p. 120.

[36] Para. 40 *Soulier and Doke* (C-301/15, EU:C:2016:878).

## Practical implementation

In the light of the above-mentioned, it seems almost impossible to reconcile an ECL system not founded on an EU legislature-based derogation such as the Czech one with the *Soulier and Doke* decision. On the other hand, as was already noted, the question of whether the needed derogation from the author's rights must be understood only as a specific and explicit exception from the exclusive rights still remains (Sganga 2017, p. 330). However, the ultimate goal of a compliant ECL system should be the actual and individual informing of the respective author whose work is to be subjected to the ECL regime, or as the CJEU put it, to set up a mechanism that ensures that *"authors are actually and individually informed".*[37]

As of November 2017, the issue of OW ECL does not seem to be reflected sufficiently by the CMOs – the respective CMOs offer only scant information,[38] even though the standard transparency rules should apply to the CMOs mutatis mutandis pursuant to Sec. 103 para. 7 of the CA (i.e. the Head IV CA), especially as regards to the rate of license fees. Furthermore, the actual realization of a diligent search and proving that it has been carried out could be regarded as an issue lacking detailed guidelines and the best practices. It must be noted that these problems only seem to mirror the problematic issues of exception-based uses of OW.[39] However, the best practices should include a rigid framework that would enable adequate handing over of verifiable results to the CMOs. The practical problem in the case of diligent search is that the result desirable for the potential licensee is actually a negative one. Proving a negative fact is not generally practical. Ideally, all of the needed sources to be searched (Annex 2 of the CA) shall be able to produce an electronically signed log of results that should be presented to the CMO. Furthermore, the CMOs shall (Sec. 98f of the CA) offer their tariff fees – which is not the case for OW. The verifiable methodology for setting the licensing fee should be presented publicly as well.

As regards the OCW ECL, the actual details of its practical functioning still remain unknown in November 2017. However, the List shall be set up and operated by the National Library[40], and the scheduled date for the commencement of its operation is set for 2018 (Richter 2017). At least, in accordance with Sec. 98f of the CA, the CMO DILIA (CMO for literary works) made public the proposal of tariff fees for the use of OCW.[41] However, the licensing agreement with the CMO DILIA and the National Library (which acts as the representative of the libraries) is yet to be concluded (Richter, 2017). In this case, the OWC ECL system should be implemented in a more stringent way. As an example, the proposal to list the work should be made public not only via the website of the National Library, but also in a more "traditional" offline media in order to target authors in the broadest possible way and thus fulfil the requirement of "actually and individually" informing the author. The same should be done for periodicals as well.

[37] Para. 43 *Soulier and Doke* (C-301/15, EU:C:2016:878) a contrario.

[38] Such info consists mainly of providing the index of OW (i.e. orphaned subject-matter). This information is provided by OSA – Ochranný svaz autorský pro práva k dílům hudebním, z.s., (CMO representing authors of music and lyrics); INTERGRAM – nezávislá společnost výkonných umělců a výrobců zvukových a zvukově obrazových záznamů (CMO representing performers and producers of phonograms and audiovisual fixations).

[39] These problems were addressed in Bertoni, Guerrieri, Montagnani 2017, pp. 41–51.

[40] The initial investments costs are estimated at two and a half million CZK. The same amount is estimated as operating costs per year (ER, p. 80–81).

[41] Available from: **http://www.dilia.cz/index.php/component/k2/item/download/567_cf881a146082cef0a6b64920cedde1b4**.

## Conclusion

The main aim of the Amendment is to alleviate legal uncertainty by simplifying the licensing of technologically determined new uses of the majority of the works included in library funds of the memory and educational institutions. The newly introduced national regulation of OW and OWC ECL systems is ahead of the current developments in the European regulatory framework (i.e. the proposed Digital Single Market Directive). It allows for the licensing of OW to subjects other than memory and educational institutions for commercial purposes and introduces a specific ECL regime for OWC. The regulation lays down a perfect theoretical blueprint of a system that should offer the needed legal certainty as well as flexibility to all the subjects involved. However, this intended objective is significantly challenged by the recent decision of the CJEU in the *Soulier and Douke* case*.* The Czech system does not seem to fulfill the requirements set therein, and prima facie, it is not compatible with EU law (both the non-exception ECL of OW, as well as the ECL of OCW). As a proposed *de facto* solution, the practical implementation should strive to achieve the actual and individual informing of the author. Paradoxically, the contested ECL is also proposed as a solution for the treatment of OCW the Digital Single Market Directive.[42] In order to achieve full compliance with the *Soulier and Douke,* the solution might lie in introducing a specific exception of the rights of reproduction and communication to the public for the OCW.[43] Therefore, a further amendment of the Czech CA might be necessary.

## References

BERTONI, Aurora, GUERRIERI, Flavia and MONTAGNANI, Maria Lillà, 2017. *Report 2 Requirements for Diligent Search in 20 European Countries* [online]. [Accessed 29 September 2017]. Available from: **http://diligentsearch.eu/wp-content/uploads/2017/06/REPORT-2.pdf**

BORGHI, Maurizio, ERICKSON, Kris and FAVALE, Marcella, 2016. With Enough Eyeballs All Searches Are Diligent: Mobilizing the Crowd in Copyright Clearance for Mass Digitization. *Chicago-Kent Journal of Intellectual Property*. 2016. **16**(1), 135.

GERA, Matej, 2017. A tectonic shift in the European system of collective management of copyright? Possible effects of the Soulier & Doke decision. *European intellectual property review*. **39**(5), 261-264.

KELLER, Paul, 2016. CJEU ruling in Doke & Soulier case emphasizes the need for a real solution to the out-of-commerce problem. *International Communia Association* [online]. 23 November 2016 [Accessed 11 October 2017]. Available from: **https://www.communia-association.org/2016/11/23/cjeu-ruling-doke-soulier-case-emphasizes-need-real-solution-commerce-problem/**

MACKOVIČOVÁ, Michaela, 2016. Osiřelá díla v autorském zákoně. *Bulletin advokacie*. **60**(3), 29–32. ISSN 1210-6348

[42] Art. 7 of the Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market, COM (14 September 2016) 593.

[43] As suggested, for example, by Keller (2016).

*Memorandum of Understanding: Key Principles on the Digitisation and Making Available of Out-of-Commerce* [online]. European Comission [Accessed 2 November 2017]. **Available from: http://ec.europa.eu/internal_market/copyright/out-of-commerce/index_en.htm**

PRCHAL, Petr, 2013. *Osiřelá díla*. Praha: Linde. ISBN 978-80-7201-919-9.

RICHTER, Vít, 2017. Novela autorského zákona a možnosti nových služeb knihoven v digitálním prostředí. *Knihovna plus* [online]. Praha: Národní knihovna, [Accessed 7 September 2017]. **13**(mimořádné číslo). Available from: **http://knihovnaplus.nkp.cz/archiv/mimoradne-cislo-2017/informace-a-konference/novela-autorskeho-zakona-a-moznosti-novych-sluzeb-knihoven-v-digitalnim-prostredi**

SGANGA, Caterina, 2017. The eloquent silence of Soulier and Doke and its critical implications for EU copyright law. *Journal of Intellectual Property Law & Practice* [online]. [Accessed 27 November 2017]. **12**(4), 321–330. DOI 10.1093/jiplp/jpx001. Available from: **https://academic.oup.com/jiplp/article-abstract/12/4/321/3056743?redirectedFrom=fulltext**

SUTHERSANEN, Ema, 2017. Who owns the orphans? Property in digital cultural heritage assets. TORREMANS, Paul. *Research Handbook on Copyright Law*. Second edition. Cheltenham, U.K. ; Northampon, MA: Edward Elgar Publishing. p. 359–390. Research Handbooks in Intellectual Property series. ISBN 978-1-78536-143-2.

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

# CHANGES IN THE AREA OF EXTENDED COLLECTIVE MANAGEMENT IN RELATION TO MEMORY AND EDUCATIONAL INSTITUTIONS IN THE LIGHT OF THE CZECH AMENDED COPYRIGHT ACT

**Lucie Straková**

lucie@lucie-strakova.cz

**Masaryk University, Czech Republic**

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings**.

## Abstract

The paper deals with changes brought by the amendment of the Czech Copyright Act (Act No. 102/2017 Coll.)[1] within the relation of the extended collective management to the memory and educational institutions. Especially will be focused on a preservation exception and an exception for teaching purposes. The paper presents the current legislation, compares it to the previous one, analyses it impacts and effects especially on the area of memory institutions and educational institutions.

## Keywords

Collective Management; Extended Collective Management; Memory and Educational Institutions; Copyright; Preservation Exception; Teaching Purposes Exception

This paper was created within the project: Proměny institutu kolektivní správy ve světle směrnice č. 2014/26/EU [Transformation of the Institute of Collective Management in the light of Directive 2014/26/EU] (MUNI/A/1393/2016).

## Introduction

The latest major amendment to the Czech Copyright Act[2] was motivated primarily by the need to transpose the new EU Directive 2014/26/EU, on Collective Management of Copyright and Related rights and Multi-territorial Licensing of Rights in Musical Works for Online Use in the Internal Market[3], but it does not concern only these areas. The changes also affected memory and educational institutions (libraries, museums, schools, etc.)[4], especially their digitization efforts and the possibility to copy copyrighted works for the purpose of education.[5]

According to the explanatory report on the Amended CA, the purpose of the changes in the legislation is to facilitate the licensing of uses such as accessing digitized content on the Internet and accessing content digitized by another library in the Czech Republic at its terminals for memory and educational institutions.[6] The solution might be an extended collective management for such licensing. Thus, based on a collective agreement, libraries could share digitized content among themselves and schools and make copies of works available to the public more widely than ever before.

The second part of this paper introduced issues in the collective management of rights, with the emphasis in particular on the *extended* collective management of rights. In the following

[1] Consolidated version of Act No. 121/2000, on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (the Copyright Act), effective from 20 April 20 2017, (further referred as "the Amended CA"). Available from: **https://www.mkcr.cz/doc/cms_library/autorsky-zakon-uplne-zneni-zakona-c-121-2000-sb-k-1-7-2017-7586.docx**

[2] Ibid.

[3] Further referred as "the Directive". Available from: **http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0026**.

[4] Further referred as "memory" and/or "educational" institutions.

[5] *DŮVODOVÁ ZPRÁVA: Návrh novely autorského zákona (zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, ve znění pozdějších předpisů)*. Praha: Ministerstvo kultury. p. 22-23.

[6] Ibid.

parts, it focuses on specific novelties in the field of extended collective management which relate to the subject institutions of this paper.

## Extended collective management of rights

Collective management of rights serves as a means of protecting the legitimate rights and claims of the authors. For them, it is (usually) more useful and economical to associate and manage these matters collectively.[7] Its purpose is to apply and protect copyright law and copyright-related rights and to make the copyrighted works available to the public.

In the Czech legal environment and in many jurisdictions of EU member states, three categories of collective management can be distinguished. Namely:

- Compulsory collective management,[8] which includes the rights that cannot be effectively managed by the authors (or by the rightholders) themselves. These rights include (regardless of whether or not the rightholders are represented by the collective management organisation[9]): the right to remuneration for the specific use of the protected work,[10] the right to remuneration for the rental of the original or reproduction of the protected work,[11] and the right to use the protected work by cable transmission[12] and the right to an additional remuneration for granting of an exclusive license to use a recorded performance of a performer.[13] As regards the nature of these rights, they cannot be managed individually. The only obstacle to receiving a remuneration from this category of collective management is the need to register with the relevant CMO.[14] Otherwise, the remuneration will not be received.
- Extended collective management,[15] created as a fiction of legal representation,[16] which is based on the right of the rightholder to reserve rights that fall into this category for himself. A passive author, who is not represented by CMO, will be represented anyway in these cases.[17] A collective or cumulative agreement[18] can be concluded and entitle the potential users to use protected works. This license may be granted for rights such as performing artistic performances from a commercial sound record,[19] broadcasting a certain type of work[20] or lending the original or a reproduction of a work.[21] An active

[7] HARTMANOVÁ, Dagmar. *Kolektivní správa autorských práv a práv souvisejících s právem autorským: historicko-teoretická studie: výklad platné právní úpravy včetně porovnání s právní úpravou vybraných evropských zemí, soutěžně právní aspekty kolektivní správy*. Praha: Linde, 2000. ISBN 978-807-2012-183. pp. 19-32.

[8] Sec. 97d of the Amended CA.

[9] Further referred as "CMO" or "CMOs".

[10] Sec. 97d Art. (1) a) of the Amended CA.

[11] Ibid. Sec. 97d Art. (1) b).

[12] Ibid. Sec. 97d Art. (1) c).

[13] Ibid. Sec. 97d Art. (1) c).

[14] Ibid. Sec. 99c Art. 1.

[15] Ibid. Sec. 97d.

[16] ŠALOMOUN, Michal. Kolektivní správa – formace a deformace autorské vůle. Beck-online: Právní rozhledy [online]. 2004, č. 6 [cit. 2017-09-12]. Available from: **https://www.beck-online.cz/bo/document-view.seamtype=html&documentId=nrptembqgrpxa4s7gzpxg5dsl4zdaoa&groupIndex=6&rowIndex=0&conversationId=18992**

[17] Ibid.

[18] Collective agreements have an extensive effect, because they also affect rightholders who are not contractually represented - this is the fictitious representation mentioned above.

[19] Sec. 97e Art. (4) a) of the Amended CA.

[20] Ibid. Sec. 97e Art. (4) c), d).

[21] Ibid. Sec. 97e Art. (4) e).

rightholder who is not represented by CMO can exclude the extended collective management using a unilateral act towards the CMO and towards the user. [22] Therefore, her works could be excluded from this regime, and the management of these rights will be reserved for her.23

- Voluntary collective management, where the CMO represents the author to the extent agreed upon with the author in the contract.

The most complicated regime is extended collective management, which is still developing (even faster than the other two).[24] Its purpose is to facilitate the exercise of rights whose management might be facilitated even more in the system of collective management, but in the event that rightholders want to manage their right individually, they can exclude this regime. The rights focused on in this paper are therefore included in this regime. It is necessary to emphasize that the author can exclude this type of collective management, and it is then necessary to negotiate directly with the author (or the rightholder), not the CMO.

## Archival and preservation exception

According to the previous legal regulation, memory and educational institutions could (i) digitalize content for archival and preservation purposes and (ii) allow access (also) to the digitized content of the works in the institution's premises through dedicated technical equipment and only for purposes of study or research.[25] This previous legislation contained only relatively limited legal licenses enabling memory and educational institutions to exploit the potential of digital technologies. They could not access digitized content on the Internet or access content on the terminals digitized by another library in the Czech Republic.

The only possible way was to use one of the exceptions and limitations of the copyright, a so-called library license.[26] Within this library license, memory and educational institutions could allow the public to access both copies and original works. However, it was never possible to use such materials for direct or indirect profit.[27] As regards collective management, no remuneration was paid to the authors in these cases because of the nature of on-site lending.[28]

In cases where there is an interest in using the work beyond the copyright exceptions and limitations (especially beyond the described library licence), memory and educational institutions must conclude a standard licence agreement with the rightholder.[29] Despite the growing digitization efforts (National Digital Library project)[30], these institutions have had a relatively very limited legal scope for disseminating digitized works to the public.

[22] Ibid. Sec. 97e Art. (2).

[23] This unilateral act cannot be performed (in Czech law) in the case of radio and television broadcasting. MYŠKA, Matěj. Vybrané právní problémy licencí Creative Commons a kolektivní správy práv. In: *COFOLA 2014: The Conference Proceedings* . Brno: Masaryk university, 2014. ISBN 978-80-210-7211-4. p. 6.

[24] Some EU member states have just launched this regime in recent years (Slovakia).

[25] Sec. 37 Art. 1 of the previous Copyright Act (effective until 19. 4. 2017).

[26] Ibid.

[27] Ibid.

[28] As in the case of "standard" book lending in libraries. Sec. 37 Art. (2) of the Previous Copyright Act.

[29] *DŮVODOVÁ ZPRÁVA: Návrh novely autorského zákona (zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, ve znění pozdějších předpisů, „AZ“)*. Praha: Ministerstvo kultury. p. 22.

[30] **http://www.ndk.cz/**

The role of the libraries has been recognized as a very important in the Czech Republic[31] and in the EU itself, but the law has to find the right balance between encouraging creativity and allowing access to the works.[32] The regulation on the preservation of cultural heritage is also addressed in the forthcoming Directive on Copyright in the Digital Single Market.[33] This Directive states that cultural heritage institutions[34] shall be permitted to make copies of any works but solely for the purpose of their preservation.[35] However, the Directive on Copyright in the Digital Single Market does not present any wider exception for the communication of digitized works to the public in libraries (with the exception of out-of-commerce works).[36]

Based on this analysis, it has been proposed that the licensing of the above-mentioned uses be facilitated, or that memory and educational institutions be ensure that extended collective management for such licensing can be allowed for. New ways of communicating the work to the public (but only in libraries)[37] are covered in the Amended CA through Sec. 97e Article 4.[38] On demand, the library may make the published work and a copy of it available in an intangible form, which in this case means in particular through a computer network. Computer programs and sheet music (as in other copyright restrictions on reproductions), as well as sound or audio image recordings (i.e. music recordings or films), are excluded from this regime.

In practice, this means that libraries will be able to share all the digitized works among themselves, make these works available to the public on site in the libraries (Sec. 97e Art. (4) f) of the Amended CA), and allow access to the all digitized content they have in their possession via the Internet network (Sec. 97e Art. (4) h) of the Amended CA). In both cases, libraries have to conclude a collective agreement with the CMO. The key subject should be the National Library, which already concluded a collective agreement for electronic document delivery with the CMO Dilia.[39] This agreement applies only to the on-site communication of digitized works to the public (Sec. 97e Art. (4) f) of the Amended CA). As regards Sec. 97e Art. (4) h) of the Amended CA, which states that anyone can access the work at a time and place of their own choosing, in particular on a computer or similar network, no such contracts have yet been concluded. Libraries are also given a number of additional duties (registration of authors who have excluded extended collective management; or sent CMO Dilia information on the use of works communicated through the electronic communication.[40]

[31] The Czech Republic is the country with the densest network of public libraries in the EU. Their number is close to the number of municipalities.

[32] Cf. CLAGGETT, Karyn Temple a Chris WESTON. Preserving the Viability of Specific Exceptions for Libraries and Archives in the Digital Age. A JOURNAL OF LAW AND POLICY FOR THE INFORMATION SOCIETY, 67-80. Also available from: **http://moritzlaw.osu.edu/students/groups/is/files/2017/08/Claggett_Weston.pdf**

[33] Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on copyright in the Digital Single Market **http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016PC0593&from=EN** (further referred to as the "Directive on Copyright in the Digital Single Market").

[34] E.g. libraries, museums or galleries, or as referred in this article "memory institutions".

[35] Article 5 of the Directive on Copyright in the Digital Single Market.

[36] Ibid. Article 7.

[37] Libraries according to Sec. 2 a) of Act No. 257/2001 Coll., on libraries and the conditions for the operation of public library and information services (Library Act).

[38] Sec. 97e Art. 4, in particular letter h), i), j) of the Amended CA.

[39] BARTÁK, Jan. Elektronické dodávání dokumentů. DILIA: Divadelní, literární, audiovizuální agentura, z. s. [online]. Praha, 2017. Available from: **http://www.dilia.cz/index.php/component/k2/item/8418-elektronicke-dodavani-dokumentu-edd**.

[40] Ibid.

## Exception for teaching purposes

Under previous copyright law, there was no possibility for schools, universities and other educational institutions to legally allow a printed copy to be made. The only way was (respecting the conditions of the Three-step test)[41] to use the standard copyright exception for the internal needs of the educational institution (Sec. 30a Art. (1) of the Amended CA), which did not involve the distribution of copies of these works. Therefore, making copies of educational materials (typically parts of poorly accessible[42] textbooks or books) and providing them to students was considered as a breach of law. The following provision were, therefore added to the list of authorizations to exercise the rights covered by extended collective management

*"Making a printed copy of a work beyond the scope laid down in Sections 29 and 30a (1) and the distribution of such a reproduction by a school, educational institution or higher education institution for the* sole *purpose of education and not for the gain of a direct or indirect economic or business advantage."*[43]

It is important to note that the fundamental restrictive conditions for such use are contained in the text. This provision is intended solely for printed copies (it, therefore, cannot be a digital copy), while the condition of use solely for educational purposes must be respected, as must the condition of non-commercial use. Here, we arrive at a difficult situation with regard to the interpretation of the terms *economic* and *business*. The law does not interpret these terms (similar to the term "commercial"), and we can find a solution in the case law of neither national nor European courts.

We could assume that the legislature intended to express to the fullest extent possible what can be considered as income, i.e. profitable activity. The Czech Copyright Act Commentary states that the important aspect for distinguishing commercial or non-commercial use is the purpose for which the work is used.[44] In a case where the user of the work intents to gain some profit by using the work, the commercial intension exists – such use will be always commercial. If these copies are used for educational purposes, does that automatically mean that they are not used for gaining any direct or indirect economic or business advantage? If the education is provided for a fee, will the distribution of copies be considered an act carried for the purpose (direct or indirect) of gaining profit?

Regarding the other requirements, it should be emphasized that there is no (compared to the previous part about exceptions for libraries) possibility of accessibility via a computer network or the Internet. Therefore, digital systems with educational materials (educational resources) or courses such as MOOC (massive open online courses) are not affected by this novelty.

[41] "The *exceptions and limitations of copyright may be exercised only in the specific cases provided for in this Act and only if such use of the work is not inconsistent with the normal use of the work and does not unduly prejudice the legitimate interests of the author*." Sec. 29 Art. 1. of the Amended CA.

[42] In particular, so-called out-of-commerce works are newly managed, both in the Amended CA (Sec. 97f , Sec. 97e Art. (4) i) of the Amended CA) and in the Directive on copyright on the Digital Single Market (Title III Chapter I), but they will not be further discussed in this article because of the lack of space.

[43] Sec. 97e Art. (4) k) of the Amended CA.

[44] TELEC, Ivo and Pavel TŮMA. *Autorský zákon: komentář*. 1. vyd. Praha: C.H. Beck, 2007, xviii, 971 s. ISBN 978-80-7179-608-4. pp. 201-206, section 3.

Educational institutions can thus copy material that they themselves have legally available for their students (but it cannot be, for example, reproduced on paper by a legal person for internal use - Sec. 30a Art. (1) b). It has to be emphasized that to make this possible, this is again subject to the conclusion of a collective agreement with the CMO.

## Conclusion

We can now find more cases of extended collective management in the Amended CA. We are talking about situations in which the CMO grants a collective licence agreement to use protected work in a determined manner not only within the represented rightholders for whom CMOs perform collective management on the basis of an agreement, but also copyrighted works of unrepresented rightholders.

Those exceptions for archival and educational purposes are undoubtedly a good step in the direction towards the 'user-friendly' use of works under copyright law. Memory and educational institutions will have to conclude contracts with the CMOs to pay the authors their remunerations. However, it is not yet clear whether these fees will be payed by users,these institutions, or by the state as in the case of "standard" lending in libraries. Unfortunately, some other (more theoretical) questions are not resolved for now, whether we are talking about the complex issue of defining the term"non-commercial" or the definition or effectiveness of this collective management with regard to the redistribution of remunerations (what amounts will the authors actually receive, won't the remunerations exceed the administrative costs?).

The purpose of this article was to briefly outline novelties in the field of collective management in relation to memory and educational institutions. Currently, it is not yet possible to say how these new uses under extended collective management will work and be used. Some issues that need be addressed in this area, as well as in the rest of the amended Copyright Act in relation to collective management, still remain.

## References

BARTÁK, Jan. Elektronické dodávání dokumentů. In: *DILIA: Divadelní, literární, audiovizuální agentura, z. s.* [online]. Praha: DILIA, divadelní, literární, audiovizuální agentura [Accessed 1 October 2017]. Available from: **http://www.dilia.cz/index.php/component/k2/item/8418-elektronicke-dodavani-dokumentu-edd**

CLAGGETT, Karyn Temple and Chris WESTON. Preserving the Viability of Specific Exceptions for Libraries and Archives in the Digital Age. *A Journal of Law and Policy for the Information Society*. **13**(1), p.67-80. Available also from: **http://moritzlaw.osu.edu/students/groups/is/files/2017/08/Claggett_Weston.pdf**

Consolidated version of Act No. 121/2000 on Copyright and Rights Related to Copyright and on Amendment to Certain Acts (the Copyright Act), effective from April 20, 2017. [Accessed 11 November 2017]. Available from: **https://www.mkcr.cz/doc/cms_library/autorsky-zakon-uplne-zneni-zakona-c-121-2000-sb-k-1-7-2017-7586.docx**

*Directive 2014/26/EU, on Collective Management of Copyright and Related rights and Multi-territorial Licensing of Rights in Musical Works for Online Use in the Internal Market*. [Accessed 11 November 2017]. Available from: **http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014L0026**.

*DŮVODOVÁ ZPRÁVA: Návrh novely autorského zákona (zákon č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů, ve znění pozdějších předpisů, „AZ“)*. Praha: Ministerstvo kultury.

HARTMANOVÁ, Dagmar. *Kolektivní správa autorských práv a práv souvisejících s právem autorským: historicko-teoretická studie : výklad platné právní úpravy včetně porovnání s právní úpravou vybraných evropských zemí, soutěžně právní aspekty kolektivní správy*. Praha: Linde, 2000. ISBN 978-807-2012-183.

MYŠKA, Matěj. Vybrané právní problémy licencí Creative Commons a kolektivní správy práv. In: *COFOLA 2014: The Conference Proceedings* [online]. Brno: Masaryk university, 2014 [Accessed 11 November 2017]. ISBN 978-80-210-7211-4. Available from: **http://cofola.law.muni.cz/dokumenty/29104**

*Proposal for a DIRECTIVE OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL: on copyright in the Digital Single Market* [online]. Brusel: European Commission, 2016 [Accessed 27 September 2017]. Available from: **http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016PC0593&from=EN**

ŠALOMOUN, Michal. Kolektivní správa - formace a deformace autorské vůle. *Právní rozhledy: časopis pro všechna právní odvětví* [online] Praha: C. H. Beck, **2004**(6), p. 208-215 [Accessed 12 September 2017]. ISSN 1210-6410. Available from: **https://www.beck-online.cz/bo/document-view.seamtype=html&documentId=nrptembqgrpxa4s7gzpxg5dsl4zdaoa&groupIndex=6&rowIndex=0&conversationId=18992**

TELEC, Ivo and Pavel TŮMA. *Autorský zákon: komentář*. 1. vyd. Praha: C.H. Beck, 2007, xviii, 971 s. ISBN 978-80-7179-608-4.

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: http://nrgl.techlib.cz/conference/conference-proceedings.

# GREY abART

**Jiří Hůla**

**hula@artarchiv.cz**

**The Fine Art Archive**

## Abstract

The Fine Art Archive collects, processes, and provides a large collection of documents pertaining mainly to Czech and Slovak art of the 20th and 21st centuries. We distinguish 200 various types of documents; some of them belong to the category of “grey literature” (e.g., press releases, theses, bibliographies, exhibition catalogues, and annual reports). In our paper, we introduce our collection of grey literature and its processing and archiving in the abART Information System, provided by the Fine Art Archive since 1993. We’ll focus on grey literature, searching in the abART system, records linking with all other related information, and availability of documents.

## Keywords

Grey Literature; Fine Arts; Fine Art Archive; Databases

The Ministry of Culture of the Czech Republic and the Capital City of Prague have long supported the activities of the Fine Art Archive.

*10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, 2017. ISSN 2336-5021. Available from: http://nrgl.techlib.cz/conference/conference-proceedings.

## Introduction

The Fine Art Archive gathers, processes and makes available documents and information about contemporary fine art, predominantly Czech and Slovak. The history of the Archive stretches back to 1984, when it was established as part of the work done by the private Galerie H in Kostelec nad Černými lesy, since when it has become the largest specialised collection of this type in the country, holding hundreds of thousands of documents. The Archive is non-selective, in that it maintains and processes documents in this specific area that are generally not collected (for example, invitations to exhibitions, exhibition catalogues, death notices, documentary photographs and so on). The Archive currently holds hundreds of thousands of documents, making it the largest collection of documents specialising in this area in the country.

We store documents within the storage space of the National Library of Technology in Písnice, Prague 4. In addition to storage space, we also run a library that concentrates on fine arts at the DOX Centre of Contemporary Art in Prague's Holešovice district. The library at DOX also functions as a documentary and research centre, providing people with the opportunity to study documents from our archive, print them out, scan them and work in our abART database. The Archive has become essential to researchers, students, institutions and the general public.

The archived documents and processed information are/is used to prepare dissertations, biography entries, dictionaries and so on. The Archive also organises exhibitions, forums and talks, publishes catalogues, books (for example, *Edice Divadlo 1961–1970*, 2014; *Sídliště Solidarita*, 2014; *Zdenek Seydl a knihy*, 2015) and anthologies (*Výtvarné umění 1950–1971, 1990–1996*, 2008; *Rovnoběžky a průsečíky*, 2010; *Cesty mohou býti rozličné*, 2015), and it cooperates with a number of galleries, museums and publishing houses (Academia, Arbor vitae, Gallery, Kant).

We obtain catalogues, invitations, books and other types of documents, including grey literature, by exchange or purchase. In 2016, we were also able to take possession of the uncommonly large and specialised library of the now-defunct *Ateliér* magazine.

### abART

The Archive has been developing the abART Information System, which it fully owns since 2003 in order to process documents and information. abART was created as a result of the need to provide access to information found in collected documents that might be hard to find otherwise. It is a relational, universal, open database (in terms of time, geographical area and specialisation) that is based on atomisation of the data entered (breaking down into elements that cannot be further broken down) and its mutual interconnection.

**Example:**

Milan Kundera. *Majitelé klíčů. Hra o jednom dějství se čtyřmi vizemi* (The Owner of Keys. A Play of One Act with Four Visions).

Description of document:
number of pages – 90, [2], Director’s note added – 11, [1], cover
dimensions – 190 × 120 mm
year of publication – 1962
edition – 1st
number of copies – 4,200
volume number – 35
résumé – yes

Identifiers:
national bibliography number – cnb000623503
edition – Theatre
language – Czech
author – Milan Kundera (Epilogue)
Otomar Krejča (Director’s note)
author of publication – Milan Kundera
typographer – Jaroslav Fišer (cover)
publisher – Orbis, Prague
printer – Knihtisk, Prague
location – Archive, typography, Theatre, volume 35
National Library, I 154373

A total of 143,313 persons were entered in abART at the end of September 2017, a total of 166,121 documents and 63,391 exhibitions and events processed and a total of 2,115,924 basic identifiers created. The target group of users is broad indeed – galleries, art historians, gallery owners, collectors, artists, journalists, state and municipal administration, Czech centres abroad, schools, libraries, tourist information centres and so on.

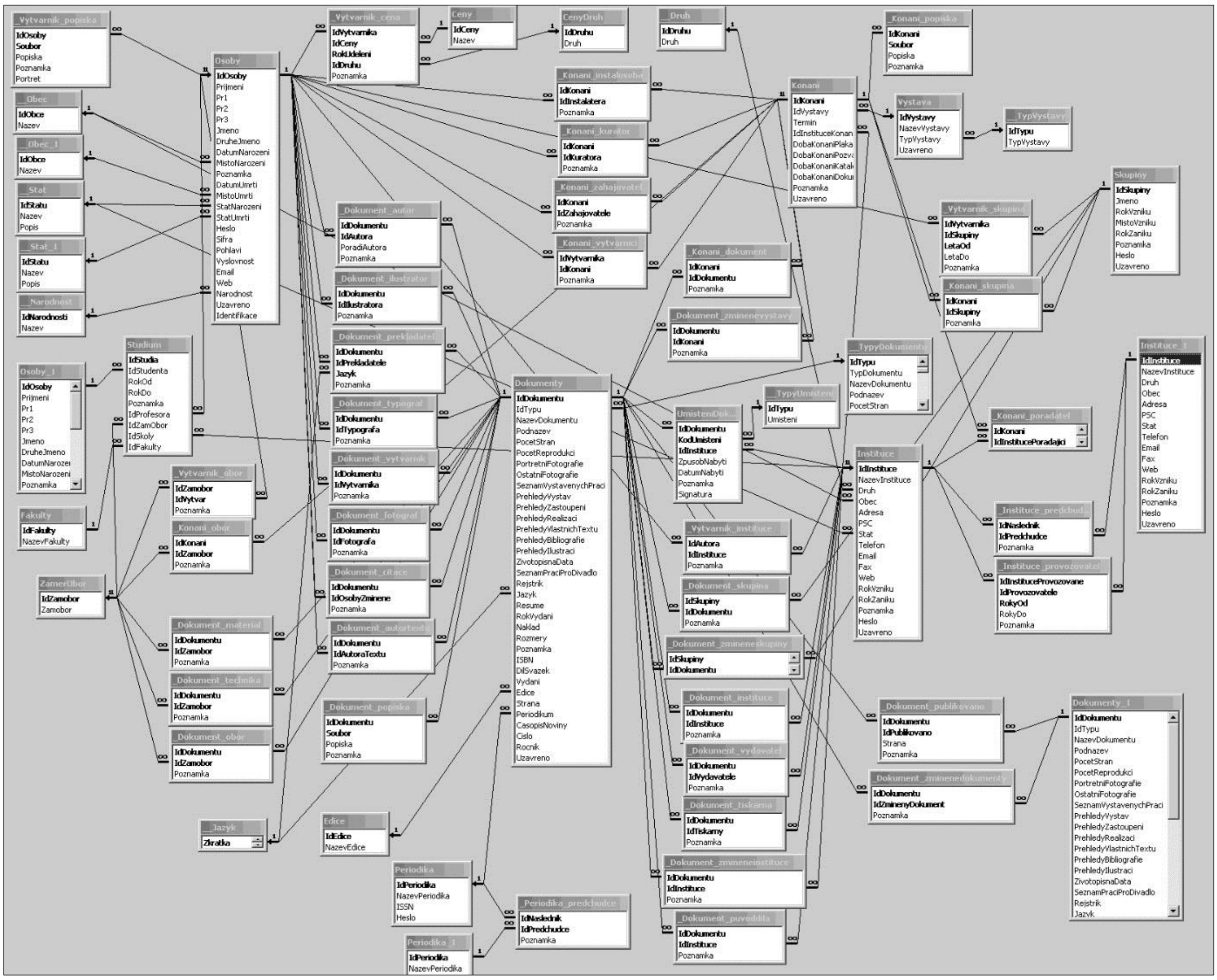
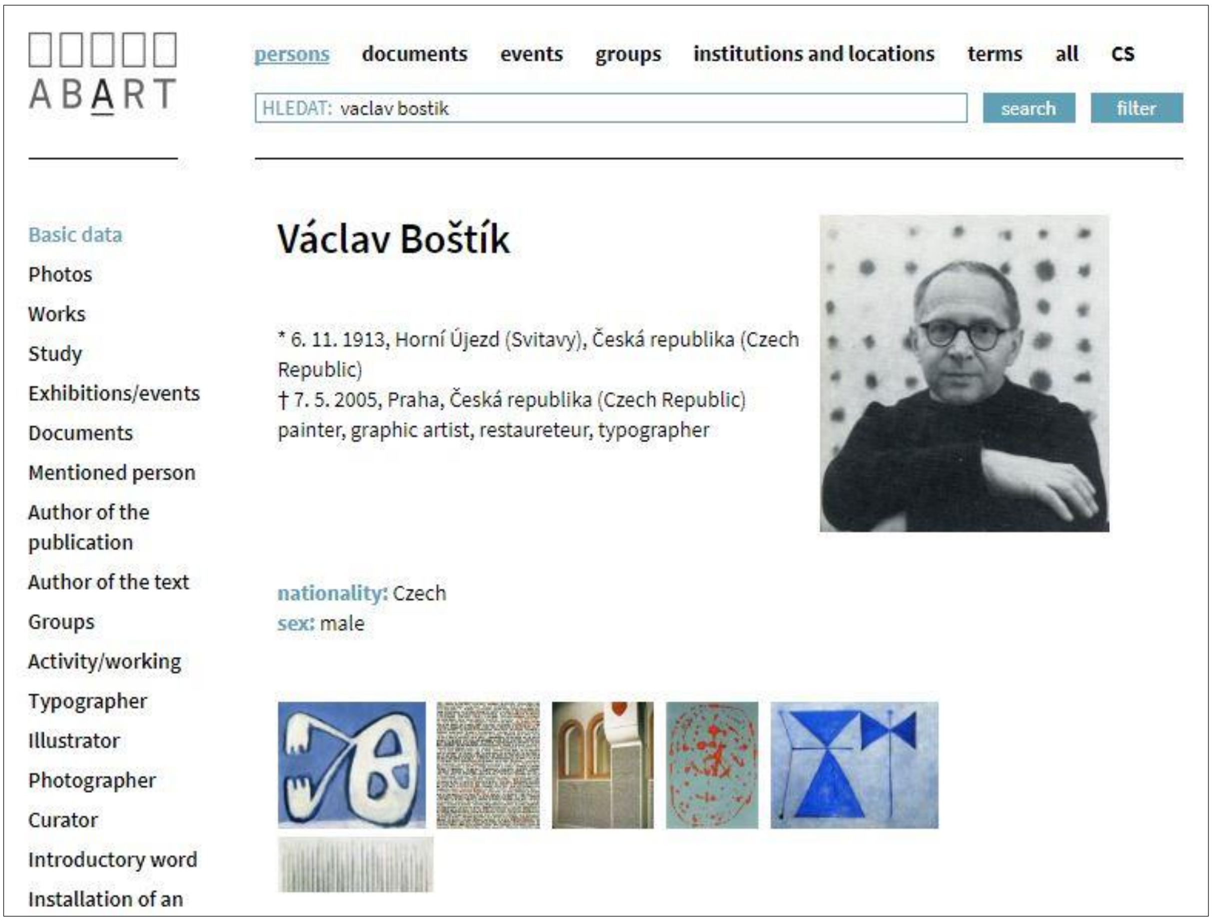


Figure 1: The structure of abART

Code lists and identifier tables form the basis of the database structure of abART. Code lists of professions, languages, types of documents, types of exhibitions/events, key words, etc. are created in Czech and English and elements of the code lists of basic categories (person, group, institution, exhibition/event, document) are specified, if we are able to trace the original name, in the relevant language, including accents and special symbols. For example, Françoise Sagan, Ivo Andrić, Sławomir Mrożek, Poul Ørum, Akademia Sztuk Pięknych we Wrocławiu, Accademia Albertina delle Bele Arti Torino, Akademie der bildenden Künste, La transfiguration de l´art tchèque.

Translation into English is generated from the code lists based on a programme. It is also possible to broaden the browser version to include other languages.


ABART

persons documents events groups institutions and locations terms all CS

HLEDAT: vaclav bostik

search filter

Basic data
Photos
Works
Study
Exhibitions/events
Documents
Mentioned person
Author of the publication
Author of the text
Groups
Activity/working
Typographer
Illustrator
Photographer
Curator
Introductory word
Installation of an

Václav Boštík

* 6. 11. 1913, Horní Újezd (Svitavy), Česká republika (Czech Republic)
† 7. 5. 2005, Praha, Česká republika (Czech Republic)
painter, graphic artist, restaureteur, typographer

nationality: Czech
sex: male

Figure 2: English version

Elements of all categories are stated only once in the code lists, apart from the required duplicate records – elements cannot always be reliably determined. It is often difficult to properly connect elements from different categories. Indeed it is currently impossible at this time. In such cases, we establish a new element with at least minimal identification (for persons, e.g. writer, painter, author of a text, curator) and a link to the source. In 2011 we organised an exhibition related to this issue entitled “Velká jména” (Big Names) (Lhoták, Kolář, Kubíček, Novák, Sýkora) at the Lesser Tower of the DOX Centre of Contemporary Art. This was about how to distinguish and assign the right artist when there are 11 Lhotáks, 48 Sýkoras, 51 Kubíčeks, 93 Kolářs and 332 Nováks in abART. Corrections and additions, dividing two different elements or unifying duplicate records is done in abART at one place and is immediately projected in all identifiers created. The newly-entered information is ordered in the browser version in the relevant place according to the relevant regulation (for example, attendance at an exhibition in the list of exhibitors alphabetically according to the surname of the artist, an exhibition/event according to its time, the authorship of a book according to the year of publication, etc.). abART does not collect data and information mechanically – it always refers to the source and primarily to its own documents – only a small part of the archive has been processed thus far. Text, image and other files can also be attached to all elements in abART (persons, institutions, documents, groups, exhibitions/events, terms).

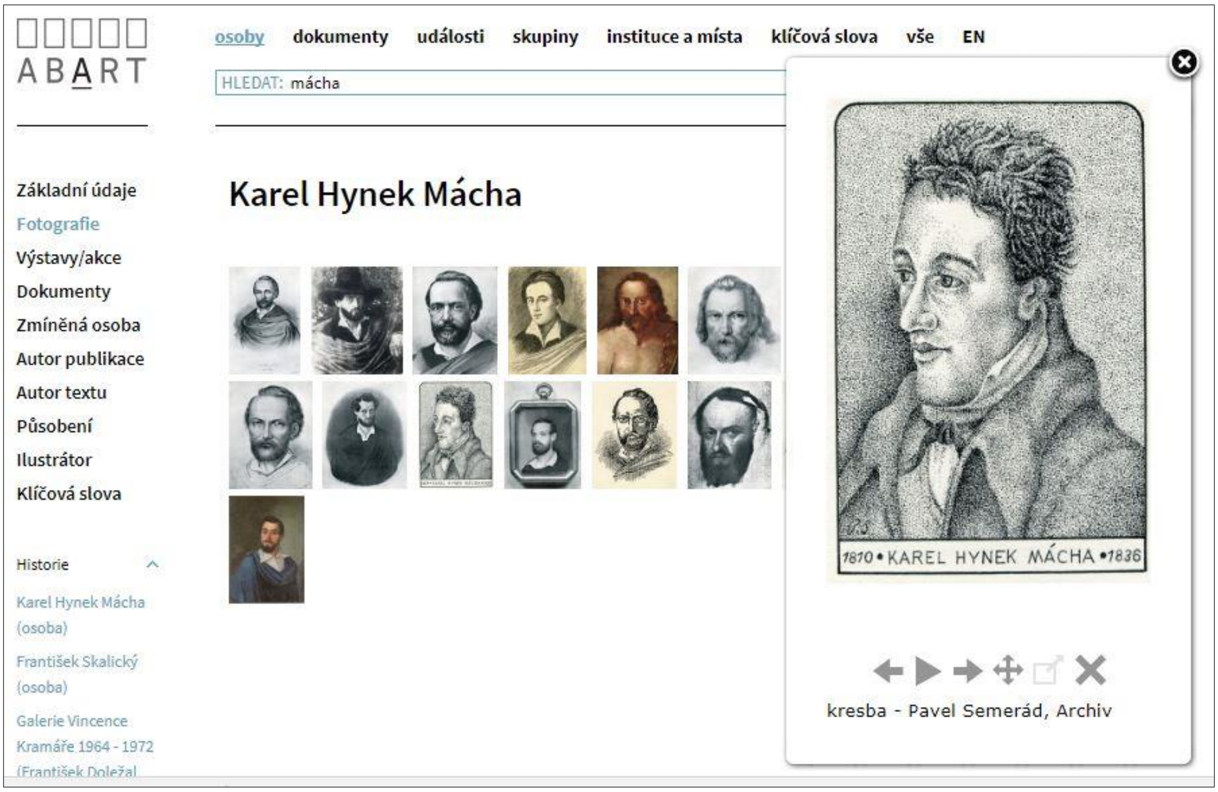
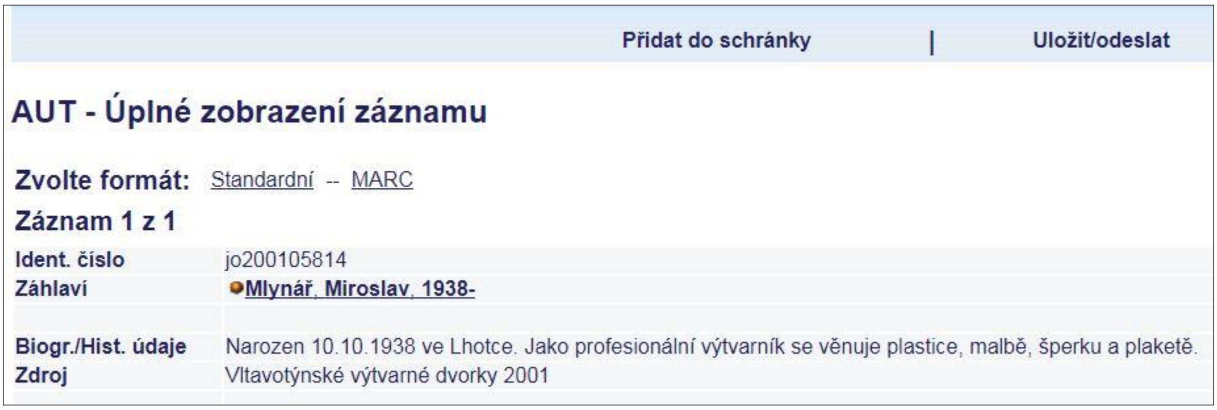


Figure 3: Karel Hynek Mácha – photograph

We are progressively matching persons, documents, institutions, communities etc. with the records in the databases of the National Library of the Czech Republic. Assuming the identification numbers of persons and the national bibliography numbers of documents makes it possible to automatically link abART to the National Library, or to other databases that also work with primary keys taken from the National Library, for example Wikipedia. However, some 35,149 persons that national authority databases do not specify are currently entered in abART. These are principally persons specified in abART by way of several identifiers (date and place of birth, date and place of death, profession, attendance at exhibition, authorship of book, authorship of text, work, etc.). The National Library now refers to abART as a reliable source more and more often in authoritative records and could therefore take persons sufficiently defined by identifiers into national authority databases and assign them with identification numbers. The identification of persons is generally relatively simple. What is more complicated, however, is the unambiguous determination of other elements – institutions, groups, documents, exhibitions/events, places of birth or death, work. Places of birth (death) are often unclear in national authority databases and are moreover specified in Czech in a case other than the nominative. By taking the places of birth (death) from the National Library, we carried a great many inaccuracies over to abART.



**AUT - Úplné zobrazení záznamu**

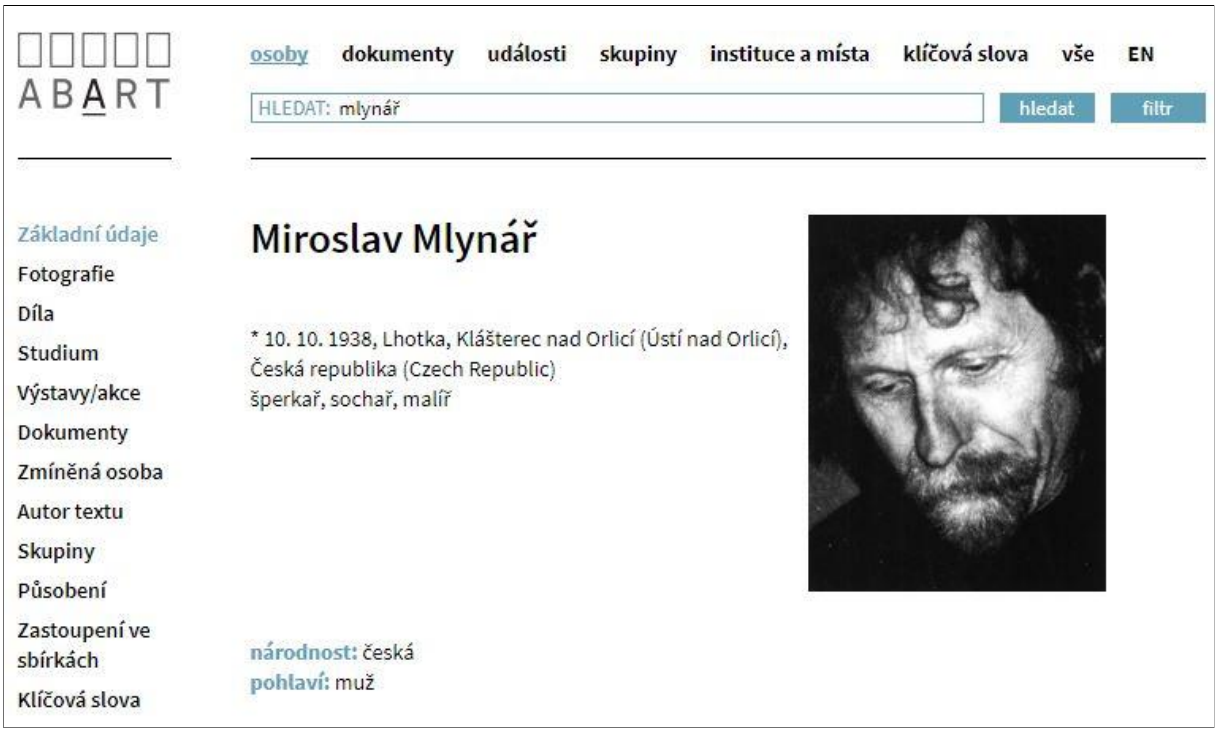
**Zvolte formát:** Standardní -- MARC

**Záznam 1 z 1**

| | |
|---|---|
| **Ident. číslo** | jo200105814 |
| **Záhlaví** | Mlynář, Miroslav, 1938- |
| | |
| **Biogr./Hist. údaje** | Narozen 10.10.1938 ve Lhotce. Jako profesionální výtvarník se věnuje plastice, malbě, šperku a plaketě. |
| **Zdroj** | Vltavotýnské výtvarné dvorky 2001 |

Figure 4: AUT – full display of record

There are more than thirty communities taking the name of Lhotka in the Czech Republic. For this reason, abART always specifies the district or the superior municipality as part of the name of the community. The record * Lhotka, Klášterec nad Orlicí (Ústí nad Orlicí) means that the relevant person was born in the community of Lhotka that is part of the municipality of Klášterec nad Orlicí in the district of Ústí nad Orlicí.



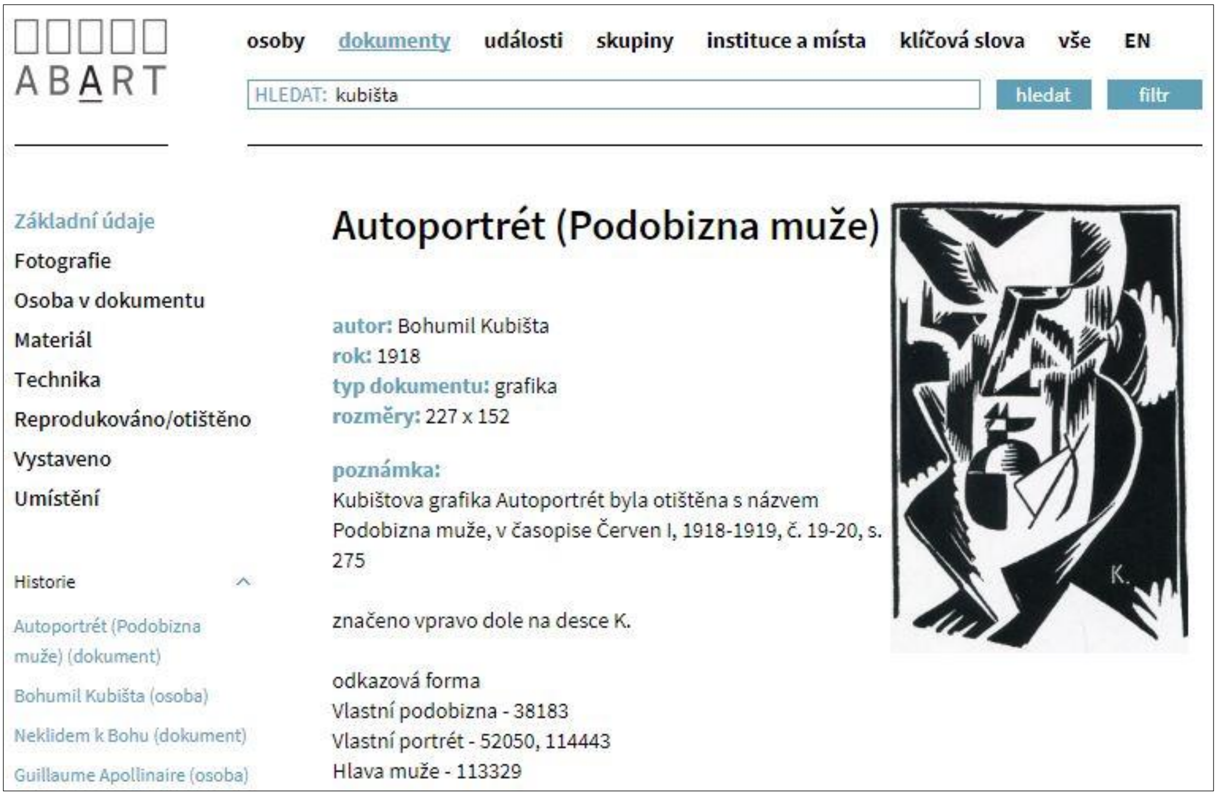
**Miroslav Mlynář**

* 10. 10. 1938, Lhotka, Klášterec nad Orlicí (Ústí nad Orlicí),
Česká republika (Czech Republic)
šperkař, sochař, malíř

národnost: česká
pohlaví: muž

Figure 5: Miroslav Mynář - basic data

We term identifiers between elements of categories as “roles”. With regard to a document, a person might be, for example, the author of a work or publication, the author of a text (an article in an anthology, magazine), a typographer, an illustrator, a translator, a photographer, he/she might be the subject-matter of a book or might feature prominently in the document (in a dictionary, encyclopaedia, catalogue). In such cases we talk of a strong person–person identifier in the document. A weak identifier (person–person mentioned in the document) allows us to create registers of books, catalogues, magazines, articles, etc. In abART. The roles of institutions in relation to a document are, for example, publisher or printer. The location of a document (book, magazine, cutting, invitation, poster, work, etc.) is a special multiple identifier (document-institution).



Autoportrét (Podobizna muže)

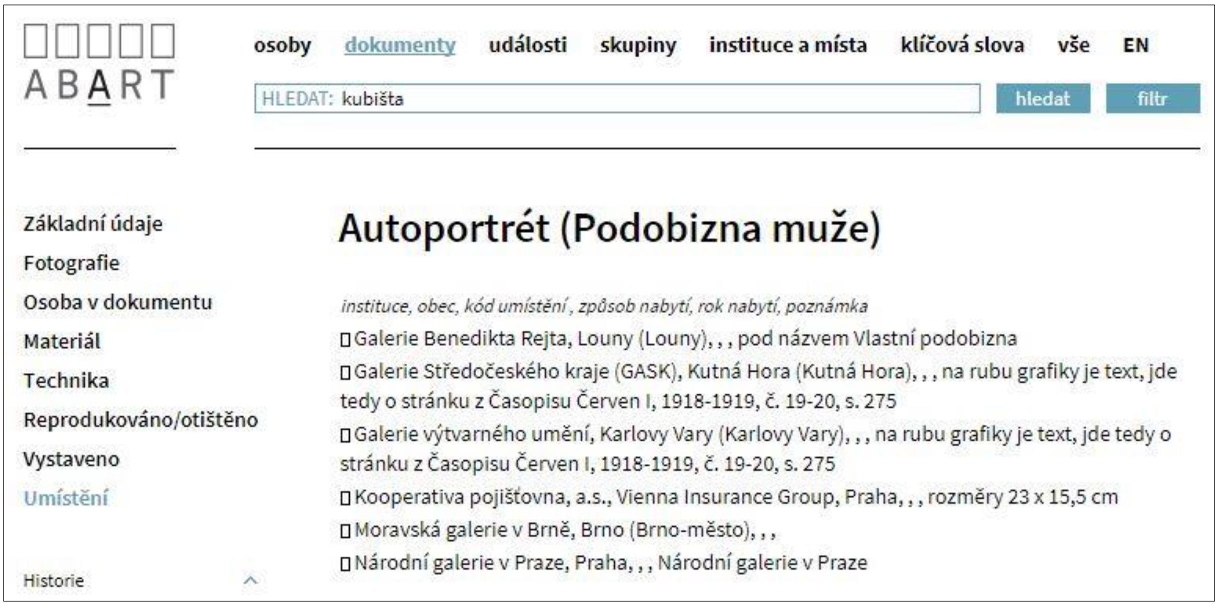
autor: Bohumil Kubišta
rok: 1918
typ dokumentu: grafika
rozměry: 227 x 152

poznámka:
Kubištova grafika Autoportrét byla otištěna s názvem Podobizna muže, v časopise Červen I, 1918-1919, č. 19-20, s. 275

značeno vpravo dole na desce K.

odkazová forma
Vlastní podobizna - 38183
Vlastní portrét - 52050, 114443
Hlava muže - 113329

Figure 6a: Bohumil Kubišta. Self-portrait – location



Autoportrét (Podobizna muže)

*instituce, obec, kód umístění, způsob nabytí, rok nabytí, poznámka*
□ Galerie Benedikta Rejta, Louny (Louny), , , pod názvem Vlastní podobizna
□ Galerie Středočeského kraje (GASK), Kutná Hora (Kutná Hora), , , na rubu grafiky je text, jde tedy o stránku z Časopisu Červen I, 1918-1919, č. 19-20, s. 275
□ Galerie výtvarného umění, Karlovy Vary (Karlovy Vary), , , na rubu grafiky je text, jde tedy o stránku z Časopisu Červen I, 1918-1919, č. 19-20, s. 275
□ Kooperativa pojišťovna, a.s., Vienna Insurance Group, Praha, , , rozměry 23 x 15,5 cm
□ Moravská galerie v Brně, Brno (Brno-město), , ,
□ Národní galerie v Praze, Praha, , , Národní galerie v Praze

Figure 6b: Bohumil Kubišta. Self-portrait – location

## abART is a product, a tool and a process.

### As a product

abART generates overviews of exhibitions, awards, representations in collections, memberships of groups, lists of exhibitors, authors of texts, lists of literature, content of books, anthologies and magazines and registers.

### As a tool

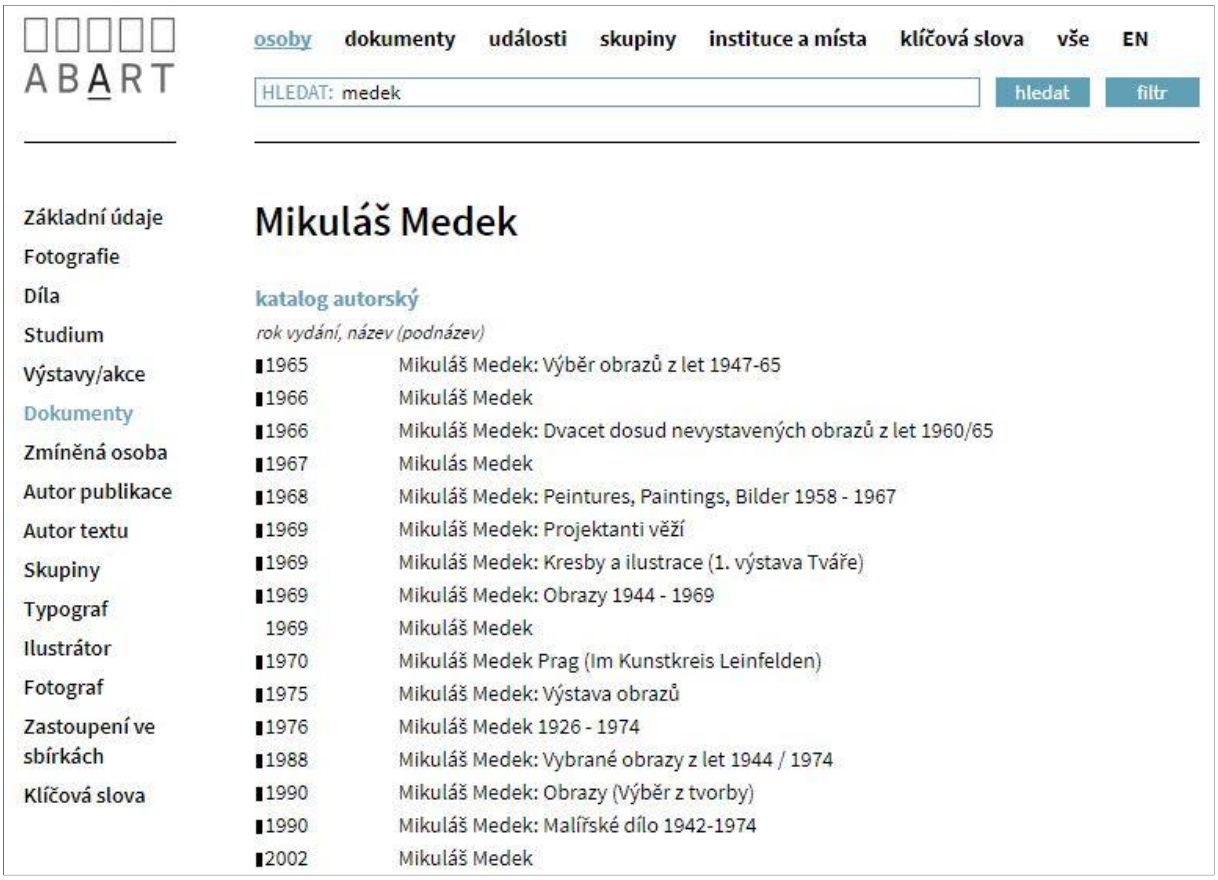
abART generates anniversaries, jubilees, natives and regional personalities, enables the export of selected data to other databases and to websites. As an example, two randomly chosen anniversaries of birthdays and two anniversaries of death from abART are shown on the archive homepage every day (artarchiv.cz). In addition to a date, such export is conditional on the existence of a portrait photograph.

### As a process

abART shows errors and shortcomings, enables collective corrections and is updated daily.

### Grey literature

In addition to basic sources (for the Archive these are catalogues, books, anthologies, magazines, articles, invitations, posters), abART also distinguishes another two hundred types of documents (including photographs, letters, New Year cards, calling cards, wedding announcements, death notices, handouts, texts, etc.). A number of these are categorised as grey literature: for example, press releases, university papers and dissertations, anthologies, annual reports, lists of work, overviews of exhibitions and lists of exhibited work. Other types of documents can be added to abART as required. abART processes all types of documents, including grey literature, in the same way. The grey literature processed in abART is shown in the browser version in the same way as other types of documents. In other words, it can be searched using full-text, identifiers and filters.



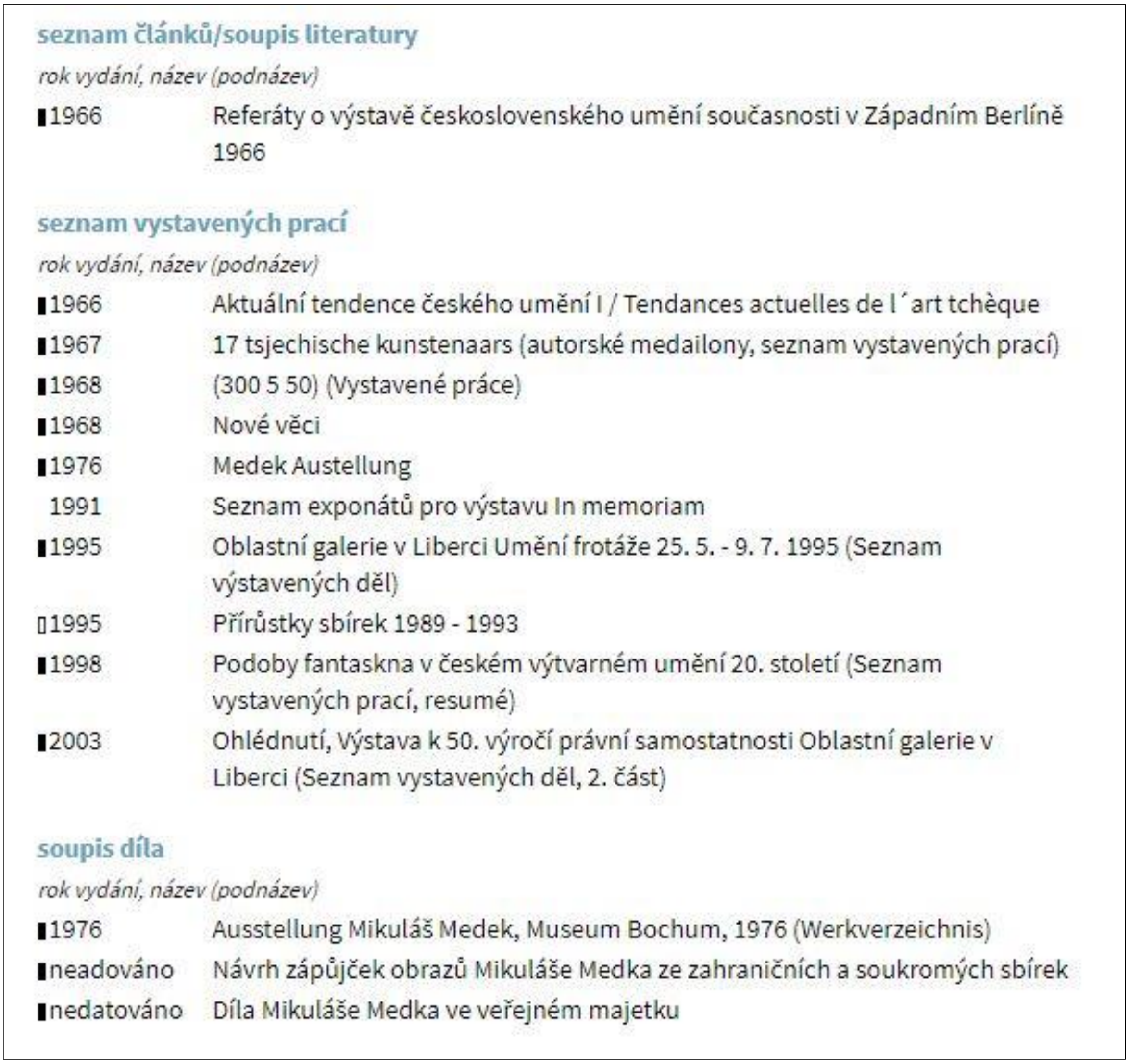
Mikuláš Medek

katalog autorský

*rok vydání, název (podnázev)*

| | |
|---|---|
| ▮1965 | Mikuláš Medek: Výběr obrazů z let 1947-65 |
| ▮1966 | Mikuláš Medek |
| ▮1966 | Mikuláš Medek: Dvacet dosud nevystavených obrazů z let 1960/65 |
| ▮1967 | Mikulás Medek |
| ▮1968 | Mikuláš Medek: Peintures, Paintings, Bilder 1958 - 1967 |
| ▮1969 | Mikuláš Medek: Projektanti věží |
| ▮1969 | Mikuláš Medek: Kresby a ilustrace (1. výstava Tváře) |
| ▮1969 | Mikuláš Medek: Obrazy 1944 - 1969 |
| 1969 | Mikuláš Medek |
| ▮1970 | Mikuláš Medek Prag (Im Kunstkreis Leinfelden) |
| ▮1975 | Mikuláš Medek: Výstava obrazů |
| ▮1976 | Mikuláš Medek 1926 - 1974 |
| ▮1988 | Mikuláš Medek: Vybrané obrazy z let 1944 / 1974 |
| ▮1990 | Mikuláš Medek: Obrazy (Výběr z tvorby) |
| ▮1990 | Mikuláš Medek: Malířské dílo 1942-1974 |
| ▮2002 | Mikuláš Medek |

Figure 7a: Mikuláš Medek – documents



seznam článků/soupis literatury

*rok vydání, název (podnázev)*

| | |
|---|---|
| ▮1966 | Referáty o výstavě československého umění současnosti v Západním Berlíně 1966 |

seznam vystavených prací

*rok vydání, název (podnázev)*

| | |
|---|---|
| ▮1966 | Aktuální tendence českého umění I / Tendances actuelles de l´art tchèque |
| ▮1967 | 17 tsjechische kunstenaars (autorské medailony, seznam vystavených prací) |
| ▮1968 | (300 5 50) (Vystavené práce) |
| ▮1968 | Nové věci |
| ▮1976 | Medek Austellung |
| 1991 | Seznam exponátů pro výstavu In memoriam |
| ▮1995 | Oblastní galerie v Liberci Umění frotáže 25. 5. - 9. 7. 1995 (Seznam vystavených děl) |
| ▯1995 | Přírůstky sbírek 1989 - 1993 |
| ▮1998 | Podoby fantaskna v českém výtvarném umění 20. století (Seznam vystavených prací, resumé) |
| ▮2003 | Ohlédnutí, Výstava k 50. výročí právní samostatnosti Oblastní galerie v Liberci (Seznam vystavených děl, 2. část) |

soupis díla

*rok vydání, název (podnázev)*

| | |
|---|---|
| ▮1976 | Ausstellung Mikuláš Medek, Museum Bochum, 1976 (Werkverzeichnis) |
| ▮neadováno | Návrh zápůjček obrazů Mikuláše Medka ze zahraničních a soukromých sbírek |
| ▮nedatováno | Díla Mikuláše Medka ve veřejném majetku |

Figure 7b: Mikuláš Medek – documents

## Conclusion

The Archive maintains a large number of documents that are not traceable in other libraries and archives, or which can only be traced with much difficulty. All types of documents processed in abART are easily accessible, including grey literature. The best-known artists, art historians are processed, as are current events and orders, less-frequented authors, profiles of all groups and exhibition halls, regional and local personalities. The information stored in abART finds new connections thanks to its identifiers to non-artistic spheres (theatre, film, literature, music, history, etc.).

The Archive fills and improves a freely-accessible database, makes archived documents available and lends them for exhibitions, runs a study room, provides libraries and galleries duplicate books, catalogues and magazines, organises exhibitions and forums about art, and so on. The strengths of the Archive are having its own sources of information, its unique information system, the size of its database, the number of identifiers, its non-selective policy and its openness. Weaknesses? The status of the organisation, insufficient cooperation with experts and institutions, unsuitable capacity, poor publicity and marketing, insufficient financing. One possible solution would be to involve the Archive in broader projects – for example, in the creation and filling of the national authority databases of the National Library, in the special subject gateway ART – Art and Architecture, in processing and exporting regional personalities to the sites of regions, districts, towns, libraries and galleries or in the creation of an international documentation and research centre.

## References

*Archiv výtvarného umění* [online]. Kostelec nad Černými lesy: Archiv výtvarného umění [Accessed 11 September 2017]. Available from: **http://artarchiv.cz/**

Archiv výtvarného umění. *DOX: Centrum současného umění* [online]. Praha: DOX, 2017 [Accessed 11 September 2017]. Available from: **http://www.dox.cz/cs/prostory-a-obchody/archiv-vytvarneho-umeni**

Ateliér: čtrnáctideník současného výtvarného umění. In: *AbART* [online]. Kostelec nad Černými lesy: Archiv výtvarného umění [Accessed 11 September 2017]. Available from: **http://www.isabart.org/institution/39140**

Autoportrét (Podobizna muže). In: *AbART* [online]. Kostelec nad Černými lesy: Archiv výtvarného umění [Accessed 11 September 2017]. Available from: **http://cs.isabart.org/document/113422/placedin**

BYTELOVÁ, Denisa, Jiří HŮLA, Irena LEHKOŽIVOVÁ, et al. *Zdenek Seydl a knihy*. Praha: Archiv výtvarného umění, 2015. ISBN 978-80-905744-5-8.

DVOŘÁK, Jan a Jiří HŮLA. *Edice Divadlo 1961–1970*. Praha: Archiv výtvarného umění and Pražská scéna, 2014. ISBN 978-80-905744-1-0.

Karel Hynek Mácha. In: *AbART* [online]. Kostelec nad Černými lesy: Archiv výtvarného umění, [Accessed 11 September 2017]. Available from: **http://cs.isabart.org/person/14516/portraits**

Mikuláš Medek. In: *AbART* [online]. Kostelec nad Černými lesy: Archiv výtvarného umění [Accessed 11 September 2017]. Available from: **http://cs.isabart.org/person/118/artist**

ŠPIČÁKOVÁ, Barbora a Michaela JANEČKOVÁ. *Sídliště Solidarita*. Kostelec nad Černými lesy: Archiv výtvarného umění, 2014. ISBN 978-80-905744-2-7.

TŘEŠTÍK, Michael. Galerie H. *Tvorba*. 1990, **1990**(3), 11 - 13.

Václav Boštík. In: *AbART* [online]. Kostelec nad Černými lesy: Archiv výtvarného umění [Accessed 11 September 2017]. Available from: **http://en.isabart.org/person/7**

Velká jména – Lhoták, Kolář, Kubíček, Novák, Sýkora. In: *AbART* [online]. Kostelec nad Černými lesy: Archiv výtvarného umění [Accessed 11 September 2017]. Available from: **http://cs.isabart.org/exhibition/39345/portraits**