conference on grey literature and repositories

proceedings 2019

CONFERENCE ON GREY LITERATURE AND REPOSITORIES

Proceedings 2019

National Library of Technology, 2019

Conference website

(**https://nusl.techlib.cz/en/conference/12th-conference-on-grey-literature-and-repositories**)

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

Editor: Mgr. Hana Vyčítalová

ISSN 2336-5021

ISBN 978-80-86504-41-4

Citation: VYČÍTALOVÁ, Hana (ed.). *Conference on Grey Literature and Repositories: Proceedings 2019* [online]. Prague: National Library of Technology, 2019 [Accessed 10 December 2019]. ISBN 978-80-86504-41-4. ISSN 2336-5021. Available from: **http://www.nusl.cz/ntk/nusl-407854**

Programme Committee:

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

Mgr. Jitka Dobbersteinová, Slovak Centre of Scientific and Technical Information (SCSTI)

Dr. Dominic Farace, GreyNet International

Ing. Martin Lhoták, Czech Academy of Sciences

Ing. Jan Mach, Ph.D., University of Economics, Prague

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

Dr. Dobrica Savić, International Atomic Energy Agency

Organizing Committee:

Mgr. Petra Černohlávková, National Library of Technology

Mgr. Hana Vyčítalová, National Library of Technology

List of Reviewers:

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

Elly Dijk M.Sc., Data Archiving and Networked Services (DANS)

Mgr. Lenka Hrdličková, Ph.D., Czech Technical University in Prague

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

M.S. Lorrie A. Johnson, U.S. Department of Energy, Office of Scientific and Technical Information

Ing. Martin Lhoták, Academy of Sciences Library

JUDr. Pavel Loutocký, Ph.D., BA (Hons), Masaryk University

Ing. Jan Mach, Ph.D., University of Economics, Prague

doc. PhDr. Richard Papík, Ph.D., Silesian University

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

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

Mgr. Małgorzata Rychlik, Adam Mickiewicz University in Poznań

Ing. Jakub Řihák, Charles University

Jan Skůpa, Brno University of Technology

PhDr. Ila Šedo, Museum of West Bohemia in Pilsen

MLIS Marcus Vaska, Alberta Health Services

JUDr. Jan Zibner, Masaryk University

Table of Contents

FOREWORD

The 12th Conference on Grey Literature and Repositories was held at the National Library of Technology in Prague, the Czech Republic, on 17 October 2019. In addition to experts from the Czech Republic, speakers from the U.S., Austria, France and Algeria also contributed to the agenda of this annual conference.

The first conference block, represented in this proceedings by the first three articles, focused on grey literature as such. The contribution of Alistair Reece from GeoScienceWorld in the U.S. deals with the possibility of incorporating grey literature into a portal that usually provides conventional published outputs (e-books, specialized journals etc.). The future of information professionals dealing with grey literature and its processing is considered in the article by Dobrica Savić from the IAEA. Grey literature can also be found in the field of fine art. The author of the paper on the development of the Czech information system for fine art, which also registers grey literature, is Jiří Hůla from the Fine Art Archive.

The second conference block was devoted to different types of digital repositories: a university repository for theses, university and institutional repositories in Algeria, and a repository for research data. In their paper, Jitka Vencláková and Markéta Trykarová present the way of processing and archiving theses and dissertations at the Technical University of Liberec. The contributions by colleagues from University of Constantine 2 in Algeria address making final theses available in digital repositories (Khaled Mettai and Behdja Boumarafi), and Algerian institutional repositories and the process of making grey literature visible in these repositories (Babori Ahcene and Aknouche Nabil). The last-mentioned repository is the Czech LINDAT/CLARIN repository for research data in the field of linguistics and application of FAIR principles described by Pavel Straňák, Ondřej Košarko and Jozef Mišutka from Charles University in Prague.

The Conference on Grey Literature and Repositories focuses on the topics of open science, open access and open data. The article by Joachim Schöpfel and Hélène Prost covers all these areas, in particular the Open Science Monitor service run by the European Commission.

Contributions from the Conference on Grey Literature and Repositories cannot neglect copyright issues. Both this year's legal contributions by Michal Koščík and Jakub Míšek focus on the innovations prepared for libraries and repositories through the new Copyright Directive in the Digital Single Market and the various permissible exceptions.

Hana Vyčítalová

THE CHALLENGES OF INCORPORATING GREY LITERATURE INTO A SCHOLARLY PUBLISHING PLATFORM

Alistair Reece

reece@geoscienceworld.org

GeoScienceWorld, USA

Abstract

GeoScienceWorld are in the process of acquiring, converting, and loading a major content repository with a significant amount of grey literature, to be hosted alongside our existing collection of peer-reviewed journals and books in the geosciences. The following issues will be addressed:

What happens when a traditionally scholarly content provider decides to incorporate grey literature into their online content platform?

What are the challenges of preparing the content for publication and discoverability?

How does the presence of grey literature in the database affect cross-search?

How do differing business models find a common home in a unified content platform?

Keywords
GeoSciences, project management, publishing platform, XML, search, business models

---

Introduction
Within the realm of geosciences there is an increasing demand for access to professionally produced, though not peer-reviewed, literature. Such literature comes in the form of reports, both corporate and governmental, meeting abstracts, presentations, and maps, as well as a range of other content types.

The one unifying feature of this content is that while it maintains a high level of integrity within the geoscience community, it has not gone through the academic peer-review process prior to publication.

As one of the leading providers of scholarly content to the geoscience community, GeoScienceWorld saw both an opportunity, and a responsibility, to bring this valuable content to its diverse subscriber base via a single access point, the GeoScienceWorld website.

Background
Established in 2004 to provide a single online source for some of the world's leading scholarly journals and e-books in geosciences. GeoScienceWorld today hosts 47 journals and more than 2 000 books on our platform, from several of the pre-eminent scholarly societies in the geosciences, including the Geological Society of America, Geological Society of London, and the Mineralogical Society of America.

The aim of GSW's founding societies was to bring together peer-reviewed, society led, research on an online platform that would encourage collaboration among the societies in order to benefit the whole collective. This approach allows smaller societies to benefit from being part of a global network of publishers, bring their content to a broader audience, whilst maintaining their independence as societies within the publishing ecosystem.

GeoScienceWorld actively supports the continued research efforts of our member societies, and has channeled more than $35 million back to the societies since our founding 15 years ago.

From the beginning of the platform, our customer base has included corporations and government bodies for whom the body of valuable geoscience content is not limited to peer-reviewed academic journals.

During the migration to our current platform provider, Silverchair, an opportunity arose to acquire a large set of content, of which more than 30% constituted "grey" literature, mainly in the form of meeting abstracts.

With the migration complete, it was decided that incorporating such content into our offering was a valuable, and strategic way forward. Naturally such an acquisition of new content types, there have been a number of challenges raised in the course of planning for the implementation phase of the project, which is scheduled to be complete in early 2020.

These challenges can be summarized as being:

- how to prepare grey literature for loading and publication
- the impact of grey literature on search functionality
- new business models required to support grey literature

Preparing the Content

The process of bringing a new journal or e-book onto the GeoScienceWorld platform is relatively straight forward. Our platform provider, Silverchair, has a stable XML specification for both journal and e-book content, in both instances using a subset of the JATS and BITS tag suites respectively. GeoScienceWorld provides new publishers with the latest version of the specifications and their content vendors create XML files, with associated assets, to be loading through the content loading tool.

The challenge in bringing grey literature into the mix is that there doesn't exist a single authoritative tag suite for handling non-peer-reviewed content. In this circumstance it is necessary to create a custom DTD and the associated XSLT required to get metadata and content into the database. An additional consideration here is that a custom DTD and XSLT is required for each unique content type within the body of content.

Before being able to get to the stage of creating the XML, DTD, and XSLT it is necessary to identify those content pieces which lack the kind of identifiers that are standard in the scholarly publishing world, such as ISSNs, DOIs, and ISBNs. For much of the grey literature in the body of content being brought into the GeoScienceWorld website, such identifiers are either not contained in the content itself or just do not exist.

Impact on Search

With 47 scholarly journals and more than 2 100 e-books on the GeoScienceWorld platform, search is the single most important feature of the website. As such, how the search engine presents non-peer-reviewed content to our users in a manner that reduces potential user confusion while maintaining the current overall user experience is a key consideration in this project.

The GeoScienceWorld platform uses the open source, enterprise scale search engine SOLR to power discoverability throughout the website. Incorporating grey literature into the site requires custom modifications to the SOLR core as well as to the index, adding new fields for the engine to search on. In order to provide the relevant data points to the search engine, the absence of a formal peer-review process has to be indicated through the XML.

Given that the GeoScienceWorld user community consists of both academic and corporate users, it is necessary to clearly indicate the peer-reviewed status of a particular piece of content. To achieve this aim GSW is implementing two approaches, firstly to introduce a facet

into the left rail of the search results page that will allow the user to filter out non-peer-reviewed content, and secondly by using a graphical indicator on the search result that identifies a piece of content as non-peer-reviewed.

The image below shows the search results page as it currently exists. The facet allowing users to filter out non peer-reviewed content will display in the left rail, immediately above the "Format" facet. The image also shows the information presented to the user for each search result. One of the options for the graphical indication that content is not peer-reviewed is to display it on the same line as the Abstract, PDF, Purchase, and Citation Manager options.

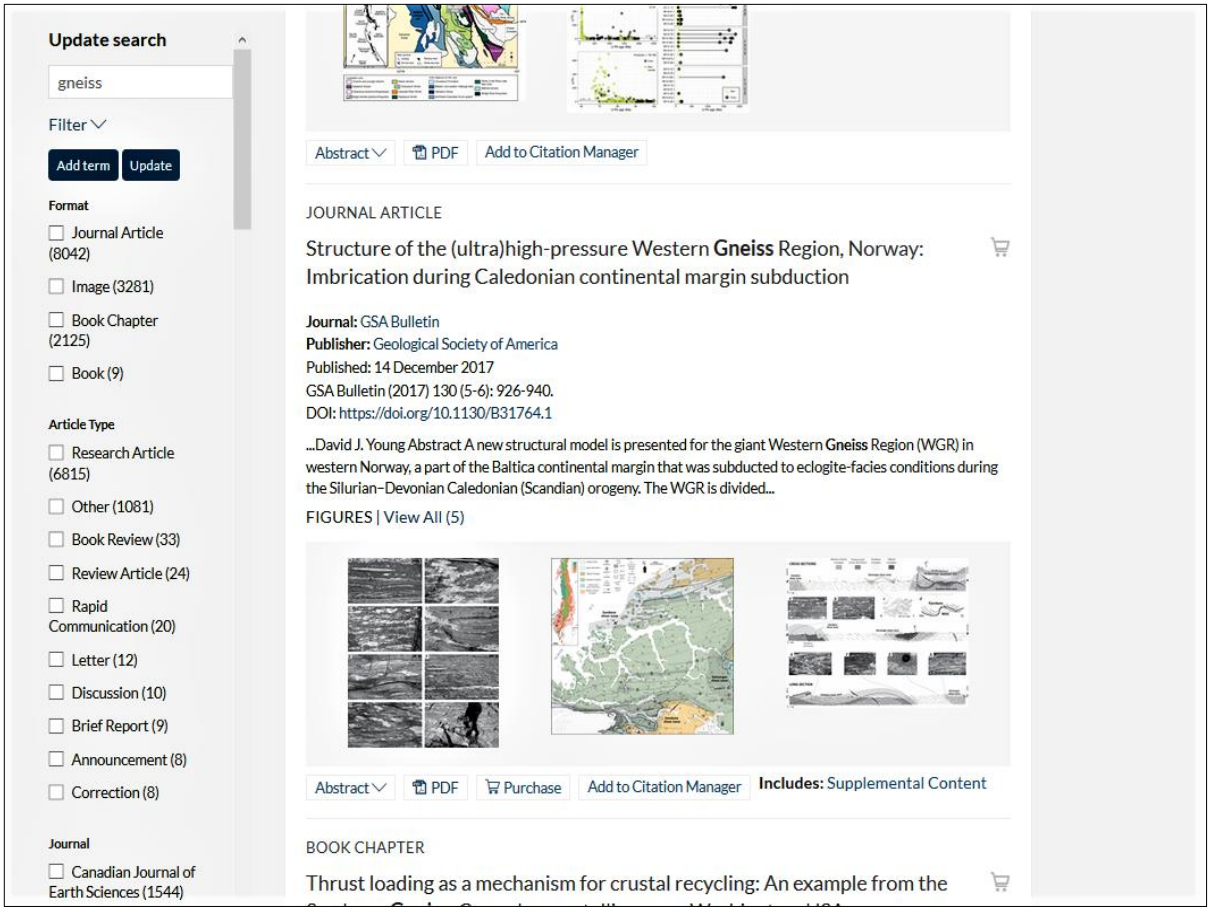

Figure 1: Search results page (GeoScienceWorld)

In reviewing how federated search tools such as EBSCO and ProQuest handle identifying the peer-review status of a piece of content, we noticed that it tended to be hidden as part of a "journal information" drop down. GSW's intention is, however, to make that identification clear to the user without further clicking, thus providing a cleaner user experience. As we move deeper into the migration project, the identification of content as either peer-reviewed or not, and it is attendant impact on user experience will be considered in more detail.

As well as the technical considerations with regards to including non peer-reviewed content into the search experience, it is important that the expectations of the user be accounted for. At present, the majority of users on the GeoScienceWorld platform are students and researchers at academic institutions, whereas the majority of users for the grey literature are coming from the corporate market. The challenge in terms of search here is to make the grey literature accessible to the latter user group while not diminishing the value of peer-reviewed

content in the eyes of the academic market. This consideration is the driving force behind the facet to allow the user to filter out the content that is not relevant to their search.

Our intention is to make the initial search results page contain both peer-reviewed content and grey literature, then allow the user to use the facets in the left rail to further narrow their search as they see fit. We believe this approach has two main benefits, firstly it shows that we trust our users to make their own research decisions, and secondly it potentially brings the grey literature in our corpus to a wider audience. Our aim here is to do nothing that impedes discovery of content, an approach that is widely considered to be best practice for search within a website.

Business Models

In order to support the presence of grey literature on the GeoScienceWorld platform, it is necessary to introduce new business models to the site that support the expected behaviors of targeted customer groups. GSW identified these groups as being corporations, consultancies, government bodies, as well as non-governmental organizations. The current supported business models of subscriptions and pay per view are felt to be restrictive for these target audiences.

Based on GSW's internal research and anecdotal evidence from conversations at conferences and similar events we plan to extend our pay per view functionality to allow bulk purchases of content. Such bulk purchases will be supported through workflow modifications to the classic cart and checkout process through which are currently provided to handle pay per view. In order to cultivate relationships with corporations and consultancies in particular we will also support tokenized purchasing where the customer prepays for a set number of downloads, to be used within a given timeframe.

The current cart and checkout process while technically capable of supporting the purchase of multiple content pieces mitigates against this behavior. When the user places an article or book chapter in the cart, the user is taken to a cart view page that encourages the user to immediately checkout, with no clear method of continuing to browse content. By placing an intermediary step between the functionality to place content in the cart and viewing the cart itself, users will find buying multiple pieces of content less onerous and repetitive. The intermediary step takes the form of a popup that informs the user that content has been placed into the cart and presents them the option to continue browsing or to checkout.

The other form of purchasing and fulfillment that is being investigated to support this content is to allow customers the option to pre-pay for content to be accessed on an ad hoc basis. For example, a customer would opt to buy $1000 of pay per view content and would then have a year to use up that balance. Each time a user associated with that customer is on the site and wants to purchase a piece of content, the price of the content is subtracted from the customer’s remaining balance, with notifications sent to the customer’s administrator informing them of the purchase and new balance.

While this purchase model would not be restricted to only grey literature, it is being considered to support the primary expected users of grey literature, non-academic corporations and consultancies. Such organizations, in general, have moved away from having subscriptions to

journals and ebooks, preferring to cherry pick content and curate their own collections through a knowledge management department. Having such a tokenized offering supports this workflow, as well as streamlining the content expenses process for the customer.

To further support both our academic and corporate customers, especially with the coming of grey literature into the platform, GeoScienceWorld are looking into opening our content to text and data mining tools, either custom built or by implementing an existing tool. This tool would feed into both the extended pay per view and tokenized purchase models.

With these new business models, and purchase methods, we intend to further extend what it means on GeoScienceWorld to purchase content. Traditional access to content, whether through a subscription or pay per view model has given the user the ability to view full text HTML content online or download a PDF version of the same content. Our extended model will allow users the option to download the content's XML files, including metadata, as well as the PDF and any supplementary material.

Conclusion

Any migration of large bodies of content from one platform to another presents a raft of challenges, in the case of peer-reviewed journals and e-books these challenges are largely known and documented as part of a migration process. Migrating grey literature between platforms, especially from a proprietary platform to one of the scholarly platforms such as Silverchair, is very much a case of starting with a few basic assumptions and then discovering the unknowns as the project unfolds.

In pursuing this migration project, GeoScienceWorld have faced challenges related not just to the content itself but how our platform supports, or can be extended to support, the business that surrounds this content. We have been reminded again of the importance of engaging in a thorough discovery process in order to at least have a broad understanding of the major work involved in the project. Such a discovery process though only has lasting value to the project if its findings are accurately represented through requirements documentation, including assumptions, stating work that will not be undertaken, and the acceptance criteria that define the successful fulfillment of the requirement.

While it is important to document the findings of the discovery phase, it is just as important to recognize that requirements can never be fully set in stone, they develop as the project proceeds and more of the unknowns come to light. For this reason, GeoScienceWorld works with our platform partner using an Agile methodology, in this case SCRUM, to constantly be refining the requirements. The ongoing refinement of the requirements allows the software being specifically developed to support the grey literature being incorporated into the site, and for that content to benefit from the features and functionality available, whether that be cross-search, purchase options, or identifying similarly themed content through related content widgets.

As GeoScienceWorld embarks on the next phase of this migration project, actually building out the features needed to support grey literature, we expect most of our assumptions to be challenged, the requirements to need changing many times, and to have a strong partnership

with our platform provider to meet the architectural problems that will likely pop up as we try to make grey literature work in a framework specifically designed for scholarly content.

References

GeoScienceWorld [online]. McLean, VA: GeoScienceWorld, 2019 [Accessed 13 September 2019]. Available from: **https://pubs.geoscienceworld.org/**

Journal Article Tag Suite. *U.S. National Library of Medicine* [online]. Bethesda, MD: National Center for Biotechnology Information, National Institutes of Health, 2019 [Accessed 13 September 2019]. Available from: **https://jats.nlm.nih.gov/**

BITS: Book Interchange Tag Set, 2019. *U.S. National Library of Medicine* [online]. Bethesda, MD: National Center for Biotechnology Information, National Institutes of Health, 2019 [Accessed 13 September 2019]. Available from: **https://jats.nlm.nih.gov/extensions/bits/**

DIGITAL TRANSFORMATION AND GREY LITERATURE PROFESSIONALS

Dobrica Savić

linkedin.com/in/dobricasavic

Austria

Abstract

Digital transformation changes the way we do business and disrupts industries and work processes, while challenging existing management practices and, in some instances, the nature of the work itself. While much attention is devoted to various digital tools such as AI, cloud computing, big data, and mobility, digital transformation is much more than information technology. The brunt of the digital transformation impact will be on the workforce itself, and it is the workforce that will determine its success or failure. Information management professionals, including managers of grey literature, represent a small, but important, part of the workforce that will be impacted by this digital disruption of the way information is managed. This paper looks at some characteristics of digital transformation and their impact on the workforce, particularly on information management professionals working in the field of grey literature. After presenting some basic terminological definitions of digital transformation, grey literature, and grey literature professionals, the major portion will review the changing nature of grey literature work, the changes required on a personal level, the impact on work organizations, and the redefined role of leadership. It is assumed that although challenging, digital transformation also provides an important opportunity for grey literature professionals.

Keywords
Grey literature, digital transformation, information management, information management profession, information technology

---

Introduction
Throughout history, the invention and development of new tools has brought about the need for new professions, while eliminating some of the previous ones. This has been especially evident throughout the past two hundred years, from the first to the most current industrial revolution, from the introduction of steam engines, through the use of powerful and smart information technology.

Digital transformation has changed the way we do business and has disrupted industries and work processes. It has challenged existing management and organizational practices, the nature of work, the workforce itself, and the role of leadership. Although much attention is devoted to various digital tools such as artificial intelligence, robotics, quantum computing, nanotechnologies, cloud computing, big data, and mobility, digital transformation is much more than information technology.

The brunt of the digital transformation impact will be on the workforce itself, and it is the workforce that will determine its success or failure. OECD (2019) estimates that 14% of jobs are at high risk of automation, while an additional 32% of jobs could undergo a radical transformation in the next 15 - 20 years. Combined, 46% of currently existing jobs on the market will undergo some type of change.

Information management professionals, including managers of grey literature, represent a small but important part of the workforce that will be impacted by digital disruption and the way information is being managed. This paper looks at some characteristics of digital transformation and their impact on the workforce, particularly on information management professionals working in the field of grey literature. After presenting some basic terminological definitions of digital transformation, grey literature, and grey literature professionals, the major portion will review the changing nature of grey literature work, the changes required on a personal level, the impact on work organizations, and the redefined role of leadership.

Brief conclusions will be offered on the increased dependency on IT tools, the changing nature of grey literature, new grey literature requirements and ways to strengthen the grey literature profession through training, and personal and professional development.

It is assumed that although challenging, digital transformation provides an important opportunity for grey literature professionals.

Digital Transformation

There are many ways to understand, define, and implement digital transformation within organizations. The main characteristic of digital transformation is that it brings about major change and introduces new ways of running a business. “Customers and employees expect a paradigm shift in their respective experiences” (Solis & Littleton 2017). Digital transformation, according to many practitioners, is a major paradigm shift where we start doing things differently.

This business change is generally based on the smart use of newly available information and technologies. They include maximized use of mobile applications, artificial intelligence (AI), machine learning (ML), cloud computing, the existence of large data sets, powerful analytics, chatbots, the internet of things (IoT), virtual and augmented reality, and many other new digital tools and services.

However, the existence and use of modern and powerful IT tools is not enough. Organizations need solid vision and forward-looking leadership. New business models need to be created using available IT solutions, leveraging existing knowledge and profoundly changing the essence of organizations - their culture, management strategies, technological mixes, and operational setups. All this is geared towards pursuing new revenue streams, creating new products and inventing new services.

Another major characteristic of digital transformation is its focus on customers (von Leipzig et al. 2017). It is regarded as a customer-centric approach, where the customer is in the centre of all decisions and actions. Focus is on customers’ needs and their overall satisfaction. With such an approach, business, and particularly financial benefits follow. By implementing some form of digital transformation, businesses manage their processes and procedures better through streamlining, profitability is increased, and new business opportunities created.

Grey Literature Professionals

In order to review grey literature professionals as a category, we need to first define the field of grey literature, and then review the terms ‘profession’ and ‘professionals’.

One of the more popular and more comprehensive definitions states that “grey literature represents any recorded, referable and sustainable data or information resource of current or future value, made publicly available without a traditional peer-review process” (Savić 2017).

At the same time, a professional is a person formally certified by a professional body or belonging to a specific profession by virtue of having completed a required course of studies and/or practice, and whose competence can be measured against an established set of standards. In other words, a professional is a person who has achieved an acclaimed level of proficiency in a calling or trade[1]. The Merriam-Webster dictionary defines a professional as a person or calling requiring specialized knowledge and often long and intensive academic preparation.

[1] www.businessdictionary.com

Taking all this into account, a grey literature professional can be defined as someone who has completed study or certification in this area, possesses specialized knowledge and skills, follows established standards, possesses the required work competencies, and regularly maintains and further develops his or her professional expertise.

In order to empirically check the above statements and do a reality check, ten major job search engines were queried using the term 'grey literature'.

Around 100 job postings were found mentioning grey literature in the job description. The largest number was found at Linkedin.com, a social networking site designed specifically for a wide spectrum of professionals (see fig. 1[2]).

Job openings were found in health, research, academia, and intelligence, with the following job titles:

Job search engines

Search engine	Hits
Indeed.com	7
CareerBuilder.com	0
Dice.com	0
Glassdoor.com	10
Jobisjob.com	10
Idealist.com	2
LinkedIn.com	40
LinkUp.com	6
Monster.com	22
US.jobs	0

Figure 1: Job search engine results

- Analyst
- Researcher
- Librarian
- Consultant
- Linguist

Analysis of the roles listed in the descriptions of the job openings indicated an interesting set of functions expected from the incumbents (see fig. 2., author is the creator of the picture). They included general knowledge of grey literature, work experience in the field, and general skills such as critical evaluation, analysis, and interpretation. However, the most interesting finding was that almost all of the jobs required some kind of search or retrieval skill, and relevant knowledge. This is an alarming finding, since according to the World Economic Forum (2018), by 2022 augmentation of existing jobs through technology may ***free up*** workers from the majority of data processing and information search tasks! In other words, there is a direct threat to all these jobs that are predominantly oriented towards information retrieval, searching and providing information.

Roles

- identify, collect and interpret
- critically evaluate
- search for
- review
- experience locating sources
- create records
- knowledge of GL

Figure 2: Job roles

[2] The author is the creator of all figures.

Impact of Digital Transformation on Grey Literature

The impact of digital transformation on grey literature is determined by the existing operational IT infrastructure, as well the decisions made by the organization's leadership on future development and investments made in new IT tools and services. Different stakeholders need to play a role in the process of digital transformation, but the main factor is ultimately the customers and users. During the decision-making process, and especially during implementation, a multitude of targets will undergo some level of change. As Figure 3 illustrates, the expected changes will be on the nature of grey literature work, on the role of leadership, the workforce itself, and on the work organization.

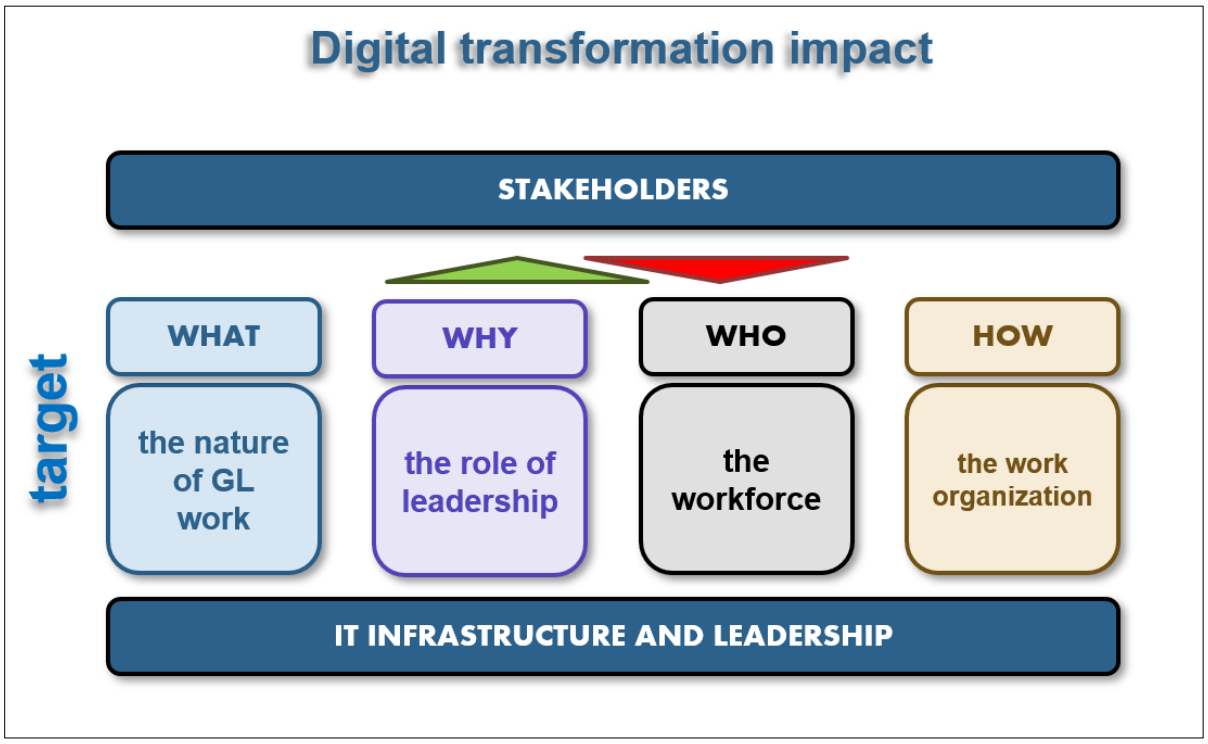

Figure 3: Digital transformation impact

A. Impact on the nature of grey literature work

There are different ways of looking at the nature of information or data management work. As Figure 4 demonstrates, one way of looking at it is through the 5 V's. Namely, the variety, volume, veracity, velocity, and value of information.

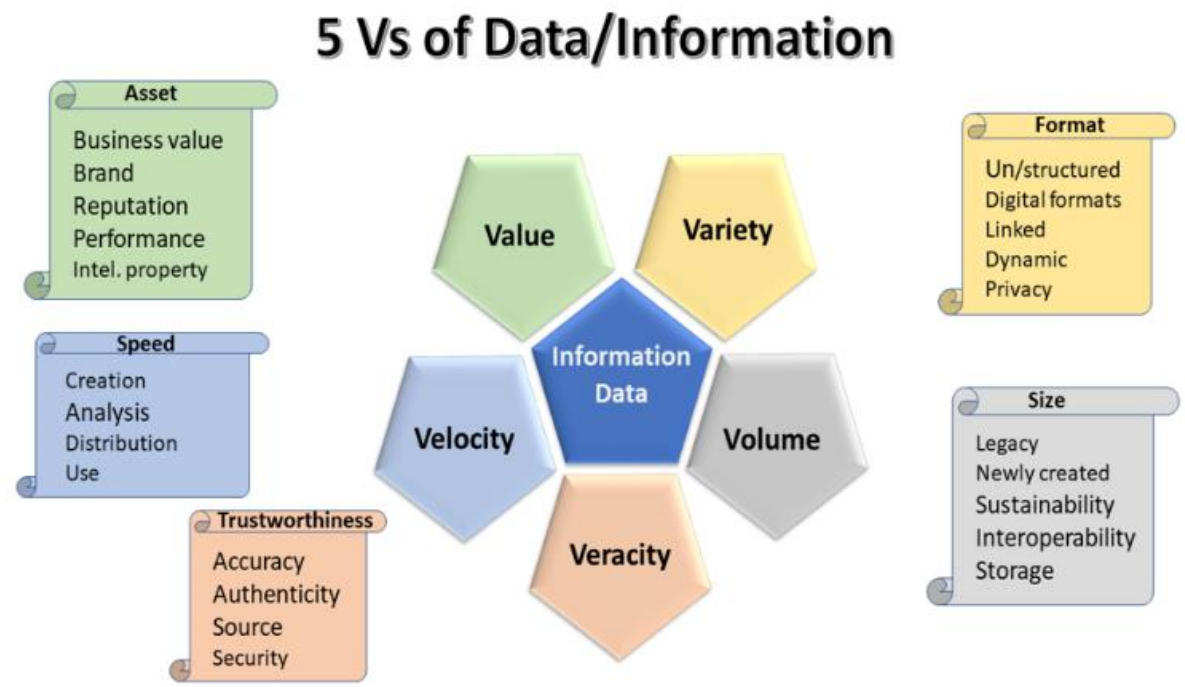

Figure 4: 5 V's of data/information management

Variety - The GreyNet website lists over 150 document types specific to grey literature, including data management. Multiple sources of grey literature include the IoT, AI, Machine to Machine communication (M2M), self-driven cars, robots, sensors, security systems, and surveillance cameras. There are also billions of connected devices creating specific formats, all within the scope of grey literature. It is obvious that a common approach to handling such a great variety of formats, resources, and types of information is not feasible through the existing grey literature scope of methods and principles.

Volume - The huge amount of data generated, and its speed of growth are detrimental factors. It is estimated that there are 38 Zettabytes of data today, out of which 90% has been generated over the last two years. 2.5 exabytes of data are produced every day, which is equivalent to 250,000 Libraries of Congress. It's a number very difficult to comprehend and even more difficult to manage. There are over 5.1 billion unique mobile users, 4.4 billion Internet users, and 3.5 billion active social media users in the world, and it is expected that these numbers will grow. There were almost 200 billion apps downloaded in 2018 and 3 billion eCommerce users, with numbers expected to grow. In addition, there are 130 million published books worldwide, with over 800,000 new titles added annually. In other words, a very large volume of data and information to be properly and efficiently managed.

Veracity - Trustworthiness and reliability of information is another huge problem which is expected to increase even further with digital transformation. Examples are numerous. They include spam email, fake news, computer bots, botnets, Web spiders, crawlers, viruses, trojans, disinformation, misinformation and many others that make veracity a real problem.

Velocity - Currently it takes 13 minutes to download the content of a DVD (4.7 GB) over a DSL line with a bandwidth of 50 Mbit/s. A 5G-enabled smartphone or laptop could download the content of an entire DVD in just 4 seconds. 5G technology involves more than just the transfer

speed. Availability and reliability are other decisive factors that make the role of grey literature professionals hard to carry out in an acceptable time frame with the required quality.

Value - Data is being widely commercialized, sold and resold, bringing a whole new spectrum of issues, required skills and organizational changes. With the change of the originally intended purpose in information and data comes the change in the role grey literature professionals should play. The notion that the value of information and data is not being depleted after consumption requires a strategic approach to long-term preservation, interoperability, and reusability.

B. Impact on the workforce

Digital transformation is more than just implementation of a new technology. It requires the adoption of a "digital workforce mindset". A digital mindset requires a deep understanding that the power of technology can democratize, scale and speed up every form of interaction and action. The main characteristics of a digital mindset are: abundance, growth, agility, comfort with ambiguity, explorer's mind, collaboration, and embracing diversity (Chattopadhyay 2016).

The impact on the workforce is expected to be multiple, including:

- Digital literacy, technical knowledge
- Lifelong micro learning and personal development
- Engagement
- Mobile force and remote work
- Generation gap
- Digital ethics

The World Economic Forum (2018) estimates that by 2022 over 50% of all employees will require significant reskilling and upskilling. This will be a huge task for HR and other mangers, especially since 85% of 2030 jobs don't exist yet (Dell Technologies 2018).

C. Impact on the workplace

The major challenge brought by digital transformation regarding the workplace, is that the advance of technology almost always outpaces existing workplace structures. Still, there are some useful approaches that can mitigate this organizational challenge. They include:

- Use new IT tools - to enhance communication, collaboration and knowledge-sharing across disparate teams. Create strong IT infrastructure and IT literacy.
- Insure digital dexterity - to fluidly and dynamically reconfigure and deploy both human and digital resources at the speed of rapidly changing technological and market conditions.
- Foster digital culture - to move away from 'paper culture' to digitally born, user-generated content collaboratively created. Increase use of social media, virtual and augmented reality tools.
- Remove information silos - to create open access data lakes, warehouses, and repositories as the basis for new intelligence, idea generation, and more effective decisions.

- Implement agile, fluid, and flexible teams - to deliver quicker and higher quality results, decrease waste of time and effort, better use resources, make staff more involved.
- Introduce remote work - to offer communication, collaboration, and learning at any time and any place.

D. Impact on the role of leadership

It is already known that digital transformation needs leaders. “Transformation leadership skills are essential and require the active involvement of the different stakeholders affected by the transformation” (Matt, Hess, & Benlian 2015). It is a people issue, not a technology one. Leadership widely differs from management. It is the art of influencing others to achieve their maximum performance and to accomplish any task. Leadership needs to offer a clear answer to WHY there is a need for change, and what the purpose or problem is that needs to be resolved.

The goal of implementing digital transformation should not be to add new technology for its own sake, but to improve competitiveness and productivity, and achieve better results and high-quality services.

Skills required from a new breed of leaders should encompass:

- Forward thinking/Visionaries/Strategists
- Customer focus
- Open communication, partnerships and collaboration
- Data-driven decision making (KPIs, value measurement, analytics)
- Tech savvy/Agile/Risk taker
- Employee empowerment/Talent promotion
- Support for creativity, innovation, experimentation
- Continuous improvement/Quick learner
- Leading by example/Role modelling

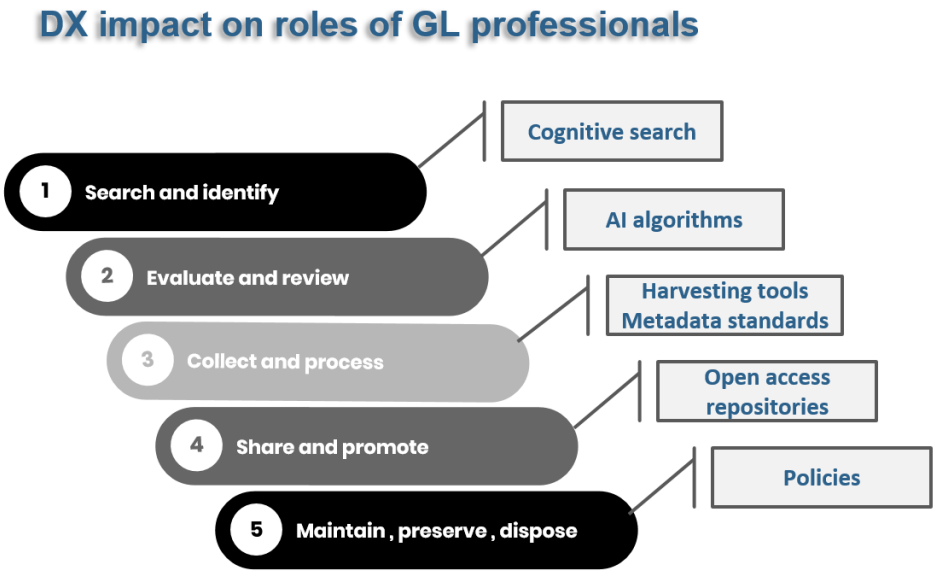

Figure 5: Roles of grey literature professionals

D. Impact on roles of grey literature professionals

Figure 5 lists five specific roles grey literature professionals currently perform and some of the major IT tools that might be beneficial in making their work more efficient and more relevant. The roles include information searching, evaluation and review, processing and sharing, maintenance, preservation, and disposal. Some of the existing IT tools already offer huge benefits in more quality and precisely performing these functions. However, continued automation efforts might completely replace human involvement and intervention. As long as there is no fear of using these tools, their existence and operational deployment can be beneficial for grey literature professionals, as well as for end-users. Open mindedness, quality education and well-planned training can make this transition less painful and more useful for everyone involved.

Conclusion

In the last few decades, developments in information technology have had an immense impact on the way we manage information in general, and particularly on the way we create, disseminate, use, and preserve grey literature. Many things have already been substantially changed and even bigger changes are imminent. As wisely stated by Charles Darwin long ago - *It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change*. Grey literature professionals will be faced with the following major changes, making it necessary to change and adapt to new realities.

Dependence on IT tools is already considerable, but it will continue to grow with new developments, the implementation of new solutions, and new, sometimes competing, requirements. The impact of digital transformation will be felt by industries and all types of work, including information and grey literature management.

Changing nature of grey literature can be easily seen by the increase in grey literature types, the volume, the speed of its creation, the trustworthiness and its value. Grey literature professional ***need to develop a new digital mindset*** so that they can survive as valuable and respected staff members of future organizations. Besides directly impacting the workforce, ***the role of leadership*** needs to change, as well as the adaptability to the increased complexity of the workplace.

And finally, major work needs to be done on strengthening the grey literature profession through organized ***training***, acquisition of new skills, professional certification, standardization, cooperation with related disciplines, and hard work of professional associations.

References

CHATTOPADHYAY, Sahana, 2016. 7 Characteristics of a Digital Mindset. *People Matters* [online]. [Accessed 8 September 2019]. Available from: **https://bit.ly/2mzNplZ**

DELL TECHNOLOGIES, 2018. *Realizing 2030: A Divided Vision of the Future* [online]. [Accessed 8 September 2019]. Available from: **https://bit.ly/2FvF1yi**

MATT, Christian, Thomas HESS, and Alexander BENLIAN, 2015. Digital Transformation Strategies. *Business & Information Systems Engineering* [online], **57**(5), 339-343 [Accessed 8 September 2019]. Available from: **https://doi.org/10.1007/s12599-015-0401-5**

OECD, 2019. *OECD Employment Outlook 2019* [online]. [Accessed 8 September 2019]. ISBN 9789264727151. Available from: **https://doi.org/10.1787/9ee00155-en**

SAVIĆ, Dobrica, 2017. Rethinking the Role of Grey Literature in the Fourth Industrial Revolution. *10th Conference on Grey Literature and Repositories: proceedings* 2017 [online]. Prague: National Library of Technology [Accessed 8 September 2019]. ISSN 2336-5021. Available from: **http://www.nusl.cz/ntk/nusl-370664**. Also published by TGJ (The Grey Journal) Special Winter Issue, Volume 14, 2018.

SOLIS, Brian and Aubrey LITTLETON, 2017. *The 2017 State of Digital Transformation: research report* [online]. Altimeter [Accessed 8 September 2019]. Available from: **https://bit.ly/2mzyXAL**

VON LEIPZIG, T., et al., 2017. Initialising Customer-orientated Digital Transformation in Enterprises. *Procedia Manufacturing* [online]. **8**(2017), 517-524 [Accessed 8 September 2019]. ISSN 2351-9789. Available from: **https://doi.org/10.1016/j.promfg.2017.02.066**

WORLD ECONOMIC FORUM, 2018. *The Future of Jobs Report 2018* [online]. Geneva: Centre for the New Economy and Society, World Economic Forum [Accessed 8 September 2019]. ISBN 978-1-944835-18-7. Available from: **https://www.weforum.org/reports/the-future-of-jobs-report-2018**

abART, NATIONAL LIBRARY OF THE CZECH REPUBLIC, VIAF AND EARTHQUAKE

Jiří Hůla

hula@artarchiv.cz

The Fine Art Archive, Czech Republic

Abstract

The Fine Art Archive develops and operates its own abART information system, a database of Czech and Slovak art based on the atomization and interconnection of the data input. There are six categories - persons, documents, events, groups, institutions and locations and terms. The system distinguishes 250 types of document, many of which (press release, diploma thesis, work collection, hand-out etc.) are grey literature. In addition to documents and art or cultural events (concert, auction, book launch), abART can also define other events, such as Sokol gatherings, fires etc. To process a remote field such as an earthquake, it would be necessary to create a broad information base and identify the entered elements (e.g. persons) with records in the National Library of the Czech Republic (NL). Yet what about the 50,000 people who have been filed in abART but who are not listed by the NL?

Keywords

Database, information system, openness, non-selectiveness, link, fine art

Introduction

The Fine Art Archive (hereinafter referred to as the “Archive”) was established in Kostelec nad Černými lesy in 1984 as part of the activities of the private Gallery H (Hůla et al. 2016). The archive collects, processes and makes available all documents about contemporary – especially Czech and Slovak – fine art. It is non-selective and versatile, and open in terms of time, territory and field. Today it is probably the largest such specialized collection, with approximately a hundred thousand archival units. The Archive acquires new copies through inter-gallery and interlibrary loans, purchases and donations. The Archive also stores documents that are not systematically collected elsewhere yet are important information sources, such as invitations to exhibitions. The types of archived documents (photographs, slides, posters, clippings, songs, calendars, obituaries,…) and the way they are processed are presented by the Archive through the ‘View into the Archive’ cycle of exhibitions in the Small Tower[1] of the DOX Centre for Contemporary Art in Prague - Holešovice.

Figure 1: The Fine Art Archive (artarchiv.cz) collects, processes and makes available all documents about fine art.

Archive’s workplaces

The Archive has three workplaces in Prague, a selective library in the DOX Centre for Contemporary Art, a cellar space close to Jiřího z Poděbrad Square (clippings and duplicates), and study and storage facilities near the Vyšehrad metro station (Pod Terebkou 15) leased from Prague 4. Last spring, the Archive had to leave the storage facilities of the National Library of Technology in Prague - Písnice. A large part of the documents (typographic collection,

[1] Record of the Small Tower in abART, available from: https://cs.isabart.org/institution/24518/events

additions, journals etc.) remains sorted but provisionally stored in banana boxes. All the document types (catalogue, invitation, poster, additions etc.) are sorted in the same way in the Archive – authorial works by surname and time of performance or edition, collective and group works by title, and gallery (institution) profiles by time of performance. The disadvantage of this sorting method is the necessity to release archival materials from time to time; the advantage is that the documents for one author or gallery are kept together and it is possible to find even those not yet processed in the archive database for study purposes.

Archived documents and processed information are used for writing doctoral or diploma theses, preparing exhibitions, bibliographic entries, catalogues, dictionaries etc. Archive users include not only historians and students of art history but also collectors, gallery operators, municipalities, schools, information centres etc.

System abART

Since 2003, the Archive has been developing and implementing its own abART encyclopaedic system to process documents and make information available. abART, like the Archive, is non-selective and open, for example enabling access to the work of historians dealing with old or foreign art, exiles and personalities working in several different or outlying fields.

The bibliography of painter Alén Diviš (1900-1956), who lived in Paris in the 1920s and 1930s like composer Bohuslav Martinů, and in New York during World War II, contains among other things a crucial text by Bohuslav Martinů. In order to process his testimony of Diviš's personality and work in abART and link it to both the painter (person-document link) and the composer (person-text author link[2]), it was necessary to supplement the code list of persons with Bohuslav Martinů[3] (1890-1969).

abART is based on the atomization and interconnection of the entered data - links. Atomization is understood as the decomposition of data into further indivisible units, which can serve as the contents of search filters. Thus, for example, the decomposition of the date and place of birth or death - day, month, year, municipality (parent municipality), district - allows the creation of anniversaries, lists of natives, and regional (local, district) personalities, e.g. natives of Pilsen[4]. Listings of personalities are sorted by frequency of links created in abART.

The database structure is based on code lists and linking tables. The translation into English is program-generated from continuously supplemented Czech-English code lists. A new modification of the browsing version based on the full-text Elasticsearch search engine will offer users a more user-friendly environment in early 2020 (Elasticsearch 2001).

[2] **https://cs.isabart.org/document/3402**

[3] Record of person Bohuslav Martinů in abART, avalaible from: **https://cs.isabart.org/person/9767**

[4] **http://bit.ly/abARTosoby**

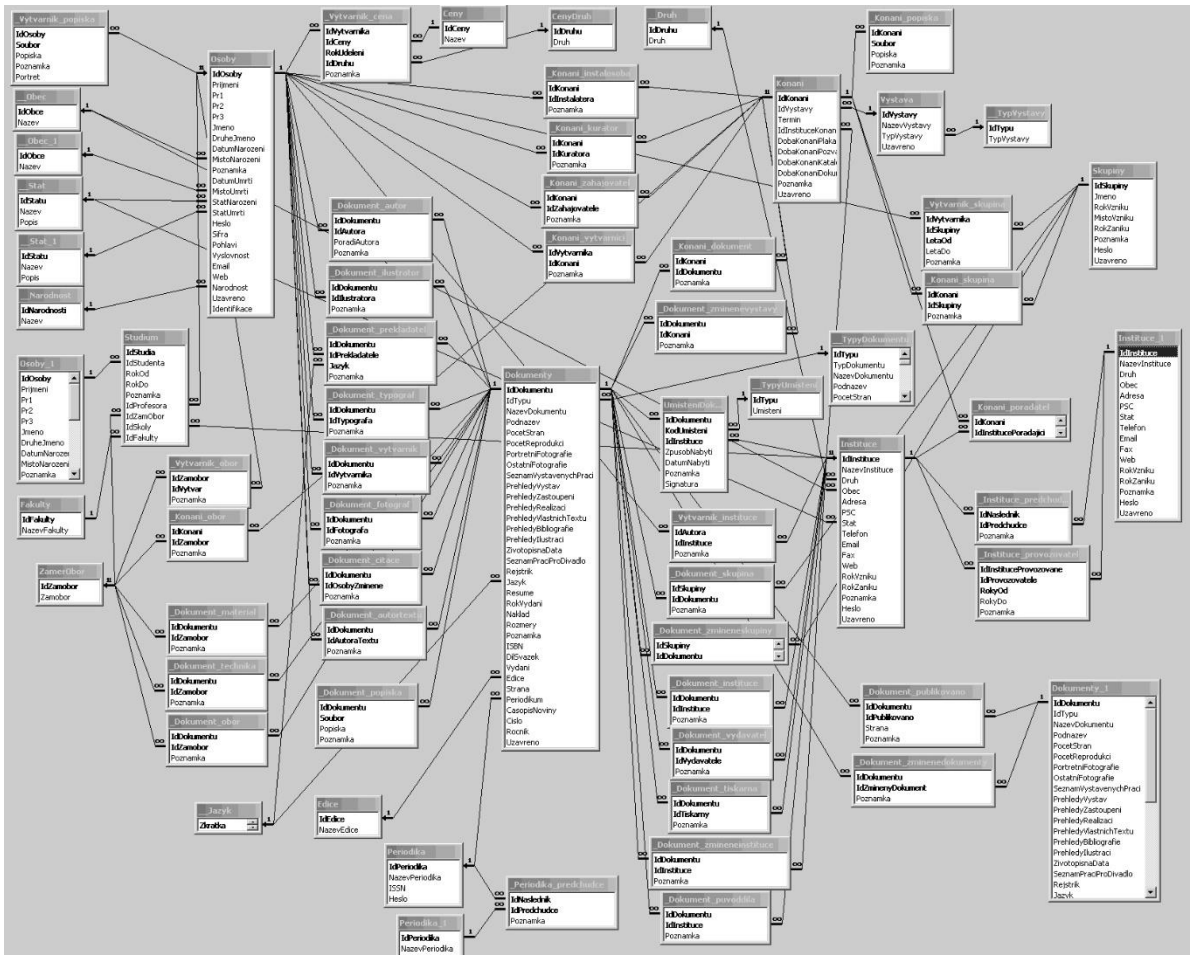

Figure 2: Structure of the abART database system (isabart.org). This image can be displayed separately in the attachment (**http://repozitar.techlib.cz/record/1427/files/Hula_diagram_of_sys_abART.jpg**).

After the discontinuation of the Ateliér magazine, the Archive acquired its specialized library. In order to preserve the idea of this gift's uniqueness in its entirety, a now non-existent institution was established in abART as another document placement possibility - Ateliér, a bi-weekly of contemporary fine art[5], a virtual library, to which documents from the now non-existent library were added.

Besides the preparation of exhibitions, publishing works and the daily storage of new data, the Archive is currently involved in two major projects: Mikuláš Medek's monograph, to be published by Academia in 2020, and a project supported by the Czech Science Foundation - Hypnotist of Modern Painting. Bohumil Kubišta and the Unrest of the Early European Avant-garde[6]. In addition to catalogues of exhibitions and literature, all the well-known works of both artists (paintings, prints, drawings and sculptures by Bohumil Kubišta, paintings by Mikuláš Medek) are filed in abART, including various titles and information on where and when the work was exhibited and reproduced.[7] Using the two-way document/work exhibited link, lists of exhibited works are created in abART together with the representation of the work at exhibitions.

[5] Ateliér available from: **https://cs.isabart.org/institution/39140/placed**

[6] Record of the project in STARFOS: **https://starfos.tacr.cz/cs/project/GA16-06181S?query_code=u4aiaacdirvq**

[7] **https://cs.isabart.org/document/82855/mentionedexhibitions**

Properly created links also enable abART to create lists of exhibitions, literature, members of art groups and associations, students and professors, and to export selected data to other databases or websites. As an example, two birth and two death anniversaries randomly selected in abART are posted on the Archive's home page (artarchiv.cz) daily. This export is conditional on the existence of a portrait photograph. (The Fine Art Archive 2019)

Grey literature

In addition to the basic sources (for the Archive, these are catalogues, books, proceedings, journals, magazines, articles, invitations and posters), abART distinguishes another two hundred and fifty types of document (including photographs, letters, New Year cards, business cards, wedding announcements, telegrams and obituaries). Many of these are grey literature, e.g. press releases, Bachelor's, Master's, rigorous, dissertation and habilitation theses, proceedings, lists of works, lists of exhibitions, lists of exhibited works, hand-outs etc. New types of documents can be added to abART as needed or according to requirements.

All types of document, including grey literature, are processed by abART in the same way. abART files and describes them in a similar way as e.g. books (document type, title, subtitle, number of pages, number of images, dimensions, year of publication or creation, number of books published, edition etc.), creating relevant links (document language, person in the document, text author, publication author, photographer, illustrator, typographer, translator, publisher, printer, storage etc.). The grey literature processed in this way can be found in abART not only by name but also by document type, document author, the person in the document, the publisher, the typographer etc.[8]

As a trusted source, information stored in abART is increasingly referred to by the Czech National Authority Database operated by the National Library of the Czech Republic. Elements filed in abART (persons, events, documents, groups, institutions, municipalities and concepts) are gradually being identified with records in the databases of the National Library of the Czech Republic[9]. In the future, the use of personal identification numbers from the National Library is intended to facilitate the connection of the Archive to other information systems that also work with the primary keys listed in the National Library, such as Wikipedia or the Virtual International Authority File – VIAF (VIAF 2019).

Events

In abART, it is possible to define not only other types of art or cultural events (exhibition opening, lecture, performance, discussion, concert and theatre performance) but also any other events, such as Sokol meetings, fires, volcano eruptions and earthquakes. One newly defined type of event (earthquake) made it possible to create a record of a natural disaster in abART, the Great Lisbon Earthquake of 1755[10].

[8] Example of searching the document type "hand out" in abART **http://bit.ly/abART_typedocument**

[9] Example of the record in the Authority database of theNational Library of the Czech Republic **https://aleph.nkp.cz/F/?func=direct&doc_number=000576054&local_base=AUT**

[10] Record of the Great Lisbon Earthquake in abART, available from: **https://cs.isabart.org/exhibition/70783**

Figure 3: Any documents and events can be processed in abART, e.g. the Great Lisbon Earthquake of 1755.

The earthquake in Lisbon is considered one of the most devastating earthquakes in the history of Europe; most of the 60,000 victims died in the tsunami and the fires following the first quakes. This, the first scientifically examined catastrophe of its kind, marked the beginning of modern seismology. Like all exhibitions and events entered in abART, the Lisbon earthquake can be linked to the relevant institutions, persons or documents, websites, drawings, graphics, manuscripts, articles, texts, fair songs etc.[11]

Kárník Archive

The archive of geophysicist and seismologist Vít Kárník (1926-1994), stored at the Institute of Geophysics of the Czech Academy of Sciences in Prague, is a collection of diverse earthquake-related documents from the oldest records and testimonies to the present. In addition to books, journals and proceedings, the archive includes notes, extracts, quotes, seismic questionnaires etc. All the documents in the Kárník Archive could be linked to the relevant seismic events in abART.

[11] Part of the record of the Great Lisbon Earthquake in abART – exhibited documents, available from: **https://cs.isabart.org/exhibition/70783/exhibited**

The catalogue of earthquakes in the Czech Republic published in the Geophysical Collection 1957 (Czechoslovak Academy of Sciences Publishing House, Prague 1958) was compiled - according to the bibliography entries - by three authors: V. Kárník, E. Michal and A. Molnár. The first two is listed in the Czech National Authority Database and there have been assigned identifiers to them by the database. Those identification numbers were collected from there by abART.

Ing. Vít Kárník, DrSc., born on 5 October 1926 in Prague, died on 31 January 1994 in Prague, geophysicist and seismologist, has an identification number jk01053078. PhDr. Emanuel Michal[12], born on 14 July 1894 in Starý Plzenec (Pilzeň-city), geologist, teacher, seismologist, zoologist, has an identification number jk01081413. A. Molnár is not listed in the Czech National Authority Database.

The identification of people is usually relatively simple. What is more difficult is the unambiguous identification of other elements such as institutions, groups, documents, exhibitions/events, places of birth or death, or works. New and so far insufficiently defined elements are created in abART with at least minimal identification. In the case of persons, this can be e.g. profession, place of birth and place of work. If the person is not yet listed at the time of the search in the Czech National Authority Database, this is abbreviated in a note - NL no, 2019/09.

Geophysicist and seismologist Molnár[13], first name Alexander, worked at the Institute of Geophysics in Prague in the 1950s. He is not yet listed in the Czech National Authority Database or the VIAF. If he was registered in the Czech National Authority Database, he would automatically be assigned a number in the VIAF. Dozens of national libraries participate in the VIAF, however linking in the opposite direction, meaning from the VIAF to the National Library, does not work. The identification number in the Czech database is unique, but in the VIAF the same person may be listed with several different personal numbers, e.g. painter Anton Perko (1833-1905) (ID1 - 53644535966399551900005, ID2 - 305601156, ID3 – 259797213)[14].

European Art Net (EAN)

In the summer of 2019, the Archive became part of the European Art Net (european-art.net) project linking twelve European institutions (archives, libraries and galleries) and their digital databases focused on information about contemporary fine art. (European Art Net 2019)

As of 26 October 2019, 76,928 exhibitions and 197,301 documents had been processed, 2,545,498 basic links and 159,676 records of persons had been created in abART, of which the Czech National Authority Database does not list 50,000 persons. These are often persons specified in abART through multiple links in addition to the basic characteristics. For example, graphic artist and sculptor Lenka Janušková, born on 22 April 1986 in Fryšták in the district of Zlín, took part in six exhibitions according to abART and is listed in five catalogues. The identifier assigned by the Czech National Authority Database would open the sculptor's access

[12] Record of the person Emanuel Michal in abART, available from: **https://cs.isabart.org/person/154180**
[13] Record of the person Alexander Molnár in abART, available from: **https://cs.isabart.org/person/153735**
[14] Record of the person Anton Perko in VIAF, available from: **http://bit.ly/VIAF_Perko**

to the VIAF. However, Lenka Janušková can easily be found in the EAN database using the full-text search engine.[15]

Figure 4: International Conference of the European Network of Art Archives - European Art Net, Fine Art Archive, 6 September 2019.

abART and earthquake

abART is primarily focused on fine art. To be able to process and make available an area as remote as an earthquake in as much detail as possible, including new types of grey literature, it would have to expand the existing code lists with thousands of new elements - seismic events, geophysicists and seismologists, geographers, scientific institutions, localities etc. and create a broad background similar to that systematically being built by the Fine Art Archive since 2003.

[15] Search of person Lenka Janušková in the EAN database, available from: **www.european-art.net/eingang_besucher/index.cfm?**

Conclusion

All documents stored in the Archive and processed in abART are easily accessible. Processing is performed from the most famous personalities, through current events and orders, less well-known authors, the profiles of all groups and exhibition halls, to regional and local personalities. Thanks to the openness and links to other cultural areas (theatre, film, literature, music, philosophy, history etc.), the information entered in abART can provide new, often unexpected, context.

The strengths of the Archive include its own information resources, a unique encyclopaedia system, the size of the database, the number of links, non-selectiveness and openness. Its weaknesses include poor promotion and marketing, organization status and poor cooperation with experts and institutions.

The information stored in abART is licensed under Creative Commons Attribution-NonCommercial-Share Alike 3.0. abART is used by many people and institutions for their work and business. The archive database recorded 10,487 visitors in September 2019.

Depending on the structure of the records taken from abART (e.g. an author's exhibition always has the title and full event duration in abART, if it can be found), it is often obvious that the information comes from abART yet users usually do not refer to the source.[16] Cooperation is not a one-way street, cooperation is reciprocity.

It is often difficult to establish the right link between insufficiently specified elements – there are still many incorrect links and deficiencies in abART. Corrections and additions, splitting two different elements merged into a single record or unifying duplicates is done in one place and will be reflected in all links. Although all errors can be corrected in abART, it is not always possible to recognize them without the help of collaborators, art historians, gallery operators, librarians, collectors, students and other people interested in fine art. Clarifying the place of birth or death is often a major problem. The browsing version will include an attempt to encourage feedback through an electronic error form. However, the greatest weakness of the Archive is insufficient funding, even though the Archive's activities have been aided for a long time by the Ministry of Culture of the Czech Republic and the City of Prague. One possible solution to the most pressing problem could be the involvement of the Archive in wider projects, for example in the building and updating of Czech National Authority Database, cooperation with similarly oriented European institutions, or the processing and export of regional personalities to the websites of regions, districts, cities, libraries, galleries, information centres and other institutions.

References

Elasticsearch, 2001. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation [Accessed 27 October 2019]. Available from: **https://cs.wikipedia.org/wiki/Elasticsearch**

[16] Example of person Jiří Balcar in the Artlist, available from: **https://www.artlist.cz/jiri-balcar-5253/**

European Art Net [online], 2019. Wien: Archiv und Dokumentationszentrum für moderne und zeitgenössische Kunst [Accessed 27 October 2019]. Available from: **http://bit.ly/EuropeanArtNet**

HŮLA, Jiří a Barbora ŠPIČÁKOVÁ, 2016. *Galerie H*. Kostelec nad Černými lesy: Archiv výtvarného umění. ISBN 978-80-905744-8-9.

The Fine Art Archive [online], 2019. The Fine Art Archive: Archiv výtvarného umění [Accessed 27 October 2019]. Available from: **https://www.artarchiv.cz/en/**

VIAF Virtual International Authority File [online], 2019. OCLC, 2019 [Accessed 27 October 2019]. Available from: **http://viaf.org/**

INTEGRATION OF UNIVERSITY QUALIFICATION THESES INTO THE TUL REPOSITORY

Jitka Vencláková

jitka.venclakova@tul.cz

Technical University of Liberec

Markéta Trykarová

marketa.trykarova@tul.cz

Technical University of Liberec

Abstract

The paper maps the process of building and putting into operation the university repository of the Technical University of Liberec (TUL). It also briefly discusses the situation that preceded the commissioning of the repository and the original options for storing and making qualification theses available. The creation of the repository is seen from the point of view of the direct participants in the process, and this paper is therefore a summary of practical experience and examples of good practice, yet does not avoid demonstrations of failed attempts. The paper also describes the current state of the active processes of the repository and a brief plan for its further development and improvement.

Keywords

University repository, digital repository, DSpace, archiving, electronic resources, university qualification theses

Introduction

The TUL institutional repository currently serves more than 8,000 students and 3,000 teaching staff. It is also open to the general public, either without registration or using a Shibboleth login.

The original idea of an internal repository originated at the TUL university library in 2004 and was motivated by the necessity to digitise and subsequently safely archive university qualification theses (hereinafter referred to as “theses”). The need to archive the publishing activities of TUL authors and for a reliable repository in connection with the entering of DOI for selected publications and articles turned out to be a secondary reason.

Creating the Repository

As early as 2012, there was an intention to build an institutional repository under the new Concept of Management and Development of the University Library of the Technical University of Liberec in the medium term of 3 years. The EPrints and DSpace systems were considered. The DSpace system was chosen because of its greater presence and support in the Czech Republic, and was first installed as version 2.0 in 2013.

Data testing was performed for one year. Seminar papers from the former Faculty of Textile Engineering of TUL in Prostějov were entered individually. Theses had been registered in the Verbis library system (including full text) since 2006. Their data were tested during bulk imports. The content structure was originally divided among the individual faculties, which were offered their own space in the repository. Later, for practical reasons, the structure was reversed and the basic division is now by document type.

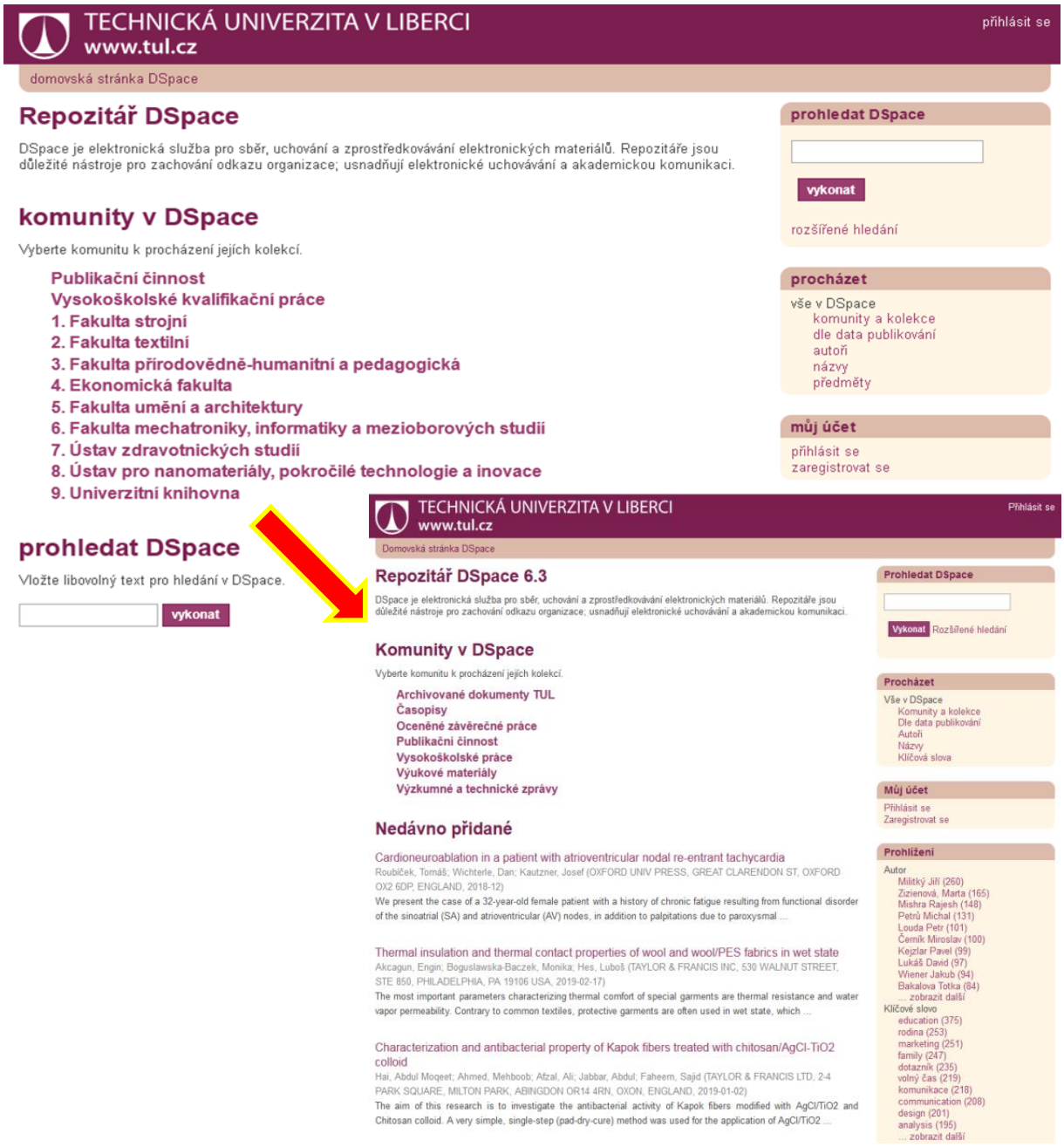

Figure 1: Change of content structure

Another opportunity to test DSpace came with a contract to digitize the Speciální pedagogika (Special Education) journal and its earlier mutations, including the installation of a DSpace repository: http://dspace.specpeda.cz/. Here again, the university library staff tested the metadata as selection data.

In 2014, the TUL repository acquired a public domain: https://dspace.tul.cz/ and the last update to version 6.3 took place in 2018. With each new release, developers have improved the functionality and modifiability of the repository, so our goal is to have the latest supported version as it continues to evolve.

History of the Digitization of and Access to Theses

Until 2006, the TUL university library provided access to a thesis study room, where users were able to view and study physical specimens from the previous five years (other theses were provided from storage upon request).

The first wave of their digitization began in 2005 and was carried out in the library digitization centre by scanning. The most recent theses were digitized first, and the process continued back to the early 1990s. Those theses were the most popular among users, and the physical specimens were in hardcover form that allowed scanning on Elsyst scanners. The images were edited in the Atlas software into the resulting PDF format.

The second wave of digitization took place between 2008 and 2011. Theses from the entire history of the university were processed, i.e. from 1956 to the 1990s. The bindings of these theses were predominantly screwed together, and this complicated imaging. Hence the bound specimens had to be carefully cut open before scanning to ensure the subsequent scanning would be completely free of defects. For faster processing, imaging took place on a Canon scanner and images were processed in Capture Perfect - again into the resulting PDF files. Almost every thesis required individual settings depending on paper type and print strength. At the end of this period, the library could declare all the theses from 1956 to 2005 digitized. Since 2006, theses texts have also been submitted on compact discs in addition to print copies, and so the full texts were entered into the library system using those CDs.

Figure 2: Digitization workplace

The gradually digitized theses were uploaded to the server, which was connected to six computers through which users could view the full texts. The computers were not connected to the Internet, had no keyboards (mouse inputs only) and had covered USB ports. The theses were thus protected against dissemination.

In 2010, the library switched from the Daimon library system to KPWinSQL (now Verbis). From 2011, full texts and metadata were downloaded from the Study Information System (STAG) to Verbis via web services.

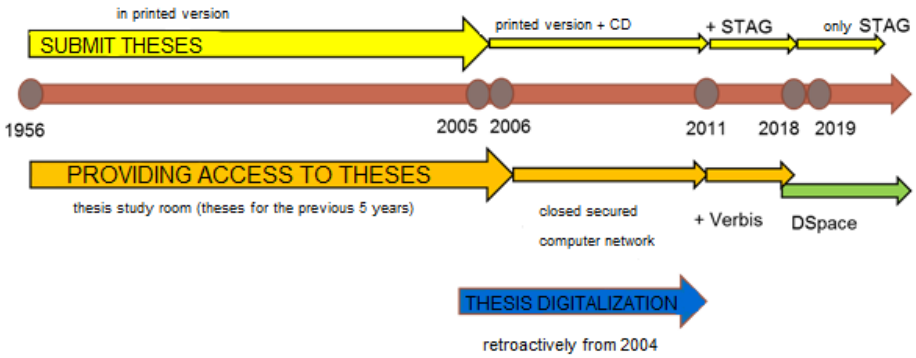

Figure 3: Submission, accessing and digitization of theses over time

After DSpace was installed, a batch import of the theses data from the Verbis library system was performed. Only metadata were imported into DSpace, and digitized full texts of the theses had to be entered manually. All the library employees participated in this work.

In 2015, the first-ever shredding of theses more than twenty years old took place (according to the valid TUL Document Rules). This was preceded by manual checking of all the PDF files, their conversion to PDF/A, their saving onto external disks and their backup onto UDO disks. 11,773 theses from the 1956-1995 period were shredded in total.

The Digitization of and Access to Theses – Present

Last year, i.e. 2018, the submission of theses to the TUL university library in paper form was cancelled and the DSpace repository is now linked to STAG. The thesis texts, their attachments, opinions and defence records are exported via an API and web services directly to DSpace. The individual metadata from STAG are mapped into the Dublin Core format in DSpace, where they remain unchanged. The full texts are imported into DSpace and only their links are available in STAG. The metadata are subsequently transferred, again via web services, from DSpace to Verbis (converted to MARC21 from the Dublin Core data format), where the metadata are modified in line with library standards. All the theses are searchable both in DSpace and in the online Portaro catalogue, where there are links to the full texts in DSpace.

CITATION | DETAIL | MARC

Field name	Field content
Title statement	Korálková nanovlákna = Beaded nanofibers /
Main entry--personal name	Havlíčková, Veronika
Added entry--personal name	Chaloupek, Jiří
Added entry--corporate name	Technická univerzita v Liberci. Textilní fakulta
Subject added entry--topical term	elektrostatické zvlákňování
Subject added entry--topical term	nanomateriály
Subject added entry--topical term	nanotechnologie
Subject added entry--topical term	textilní materiály
Třídění (anglická kl. slova)	textile materials
Třídění (anglická kl. slova)	nanotechnology
Třídění (anglická kl. slova)	electrospinning
Třídění (anglická kl. slova)	nanomaterials
Index term--genre/form	bakalářské práce
Publication, distribution, etc.	Liberec : Technická univerzita, 2011
Physical description	53 s., 46 s., příl. : obr., tab. grafy + CD ROM
Summary, etc.	Bakalářská práce se zabývá problematikou výroby korálkových nanovláken a seznamuje čtenáře s procesem výroby a klíčovými faktory, které ovlivňují jejich tvorbu. Zaměřuje se na vliv změn tří parametrů, a to: koncentrace roztoku, elektrické napětí a vzdálenost kolektoru a kapiláry. Naměřené hodnoty porovnává s teoretickými předpoklady a obsahuje návrh nejvhodnějšího roztoku ze zkoumaných vzorků pro přípravu korálkových vláken.
Katedra	KTM
Electronic location and access	http://knihovna-opac.tul.cz/diplomovaPrace.php?id_dipl=20899 5 MB pdf Plný text práce
Posudek oponenta	http://knihovna-opac.tul.cz/diplomovaPrace.php?id_dipl=20899&typ=1 1 MB jpg Posudek oponenta
Posudek vedoucího práce	http://knihovna-opac.tul.cz/diplomovaPrace.php?id_dipl=20899&typ=2 1 MB jpg Posudek vedoucího práce
Výsledek obhajoby	http://knihovna-opac.tul.cz/diplomovaPrace.php?id_dipl=20899&typ=3 41 kB pdf Obhajoba práce

Figure 4: Links to files stored in DSpace in the thesis record in the online Portaro catalogue

All the theses are now freely accessible, with the exceptions approved pursuant to Section 47b(4) of the Higher Education Act, amending and supplementing other Acts.

Conclusion and Challenges

There are a number of ways the repository can be improved, and we have selected the following areas for the coming period.

Unique Personal Identifiers

To date, there is no system of unique author identities in the university system. Some authors (an estimated 20 percent of publishing authors) have at least an ORCID or ResearcherID. However, these identifiers are only the personal choice of the authors, i.e. they are dependent purely on their willingness to set up and use them.

At TUL, there are a number of information systems that work with personal identifiers: the already mentioned STAG, the personnel system and the accounting system. Unfortunately, not all authority types (students, employees, external staff, doctoral students) are included in any of them. Recognizing the need for unique personal identifiers, the library itself has an ID collection system that allows it to distinguish between individual authors. Only time will tell

whether this system will only be used for library purposes or will be adopted and used throughout the university.

Synchronization of Subject Word Entries

Another area we want to improve is the currently neglected field of subject word entries. So far, only the subject word entries that the author of the work chose and entered into the system have been recorded for theses. Now we would like to align these subject word entries with those of the Verbis library system and implement these enriched entries back into the repository.

We expect this to improve document retrieval and the user-friendliness of the system.

Statistics

Another of our goals is to implement Google Analytics and Altmetric statistics. However, we are currently hampered by repository configuration that does not allow the rendering of the altmetric badge HTML code.

This is also related to the role of the external system developer, who must be contacted and tasked in such cases, thus slowing down the entire innovation process. We are therefore excited about DSpace 7.0 which, according to the first reports, should provide us with more room for personalized system modifications we could perform on our own and in a shorter timeframe.

Data Checking in STAG

There are occasionally unpleasant situations when a file downloaded from STAG to DSpace cannot be opened or is corrupt. Incorporating automated data checking would certainly be beneficial.

Publishing Activities

Since 2015, the reporting of publishing activities to the Information Register of R&D Results (RIV) is through the system **https://publikace.tul.cz/**, developed at TUL. In 2018, a one-time import of data from this system took place through the institutional “Automation of archiving scientific and research data” project. Community updates to the publication activity in DSpace are currently manual, based on alerts from the Web of Science and SCOPUS citation databases. TUL is not interested in archiving the full texts of publishing activities. Nevertheless, DSpace is ready.

Figure 5: The "publikace.tul.cz system" interface (the interface is not available in English).

At present, the repository is still primarily a part of the TUL university library. It is only used by individuals or some departments and institutes, but does not yet serve as generally accepted support for science and research at TUL. Rather, it is perceived as a mere repository, a sort of organized virtual catacomb of interest only to individuals searching for full texts of theses. The question of why this is the case is not easy to answer. The library itself is responsible for actively distributing the repository content to a wider audience.

We need to choose a functional strategy to attract more authors for publishing and subsequent archiving. We believe that the integration of the DSpace repository with other university information systems will take place, see Figure 6.

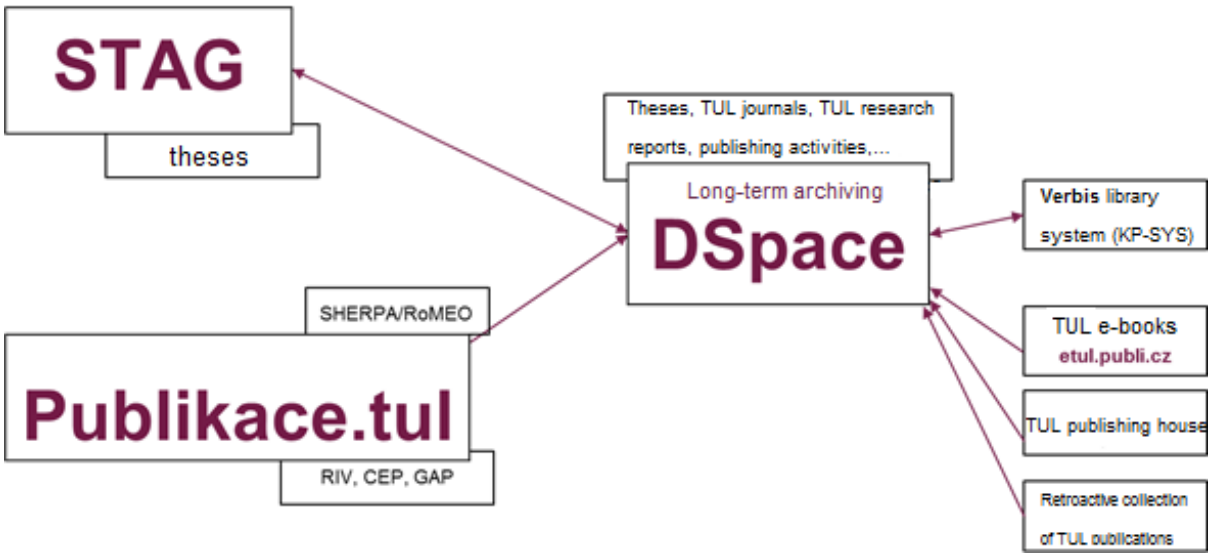

Figure 6: Integration of the DSpace repository with other TUL information systems

References

VERBIS library system. *KP-SYS* [online]. Pardubice: KP-SYS spol., 2017 [Accessed 16 September 2019]. Available from: **https://kpsys.cz/en/verbis-library-system**

DSpace 6.3 repository [online]. Liberec: Technical University of Liberec, 2018 [Accessed 16 September 2019]. Available from: **https://dspace.tul.cz/**

VENCLÁKOVÁ, Jitka. *Koncepce řízení a rozvoje Univerzitní knihovny Technické univerzity v Liberci ve střednědobém horizontu 3 let* [Management and Development Concept of the Technical University of Liberec Library in the medium term 3 years]. Liberec, 2012.

Registry of scientific and research activities [online]. Liberec: Technical University of Liberec, 2015 [Accessed 16 September 2019]. Available from: **https://publikace.tul.cz/**

TECHNICAL UNIVERSITY OF LIBEREC. *Směrnice rektora TUL č. 2/2012: Spisový řád TUL* [TUL Rector's Directive No. 2/2012: TUL Document Rules] [online]. Liberec: Technical University of Liberec, 2012 [Accessed 16 September 2019]. Also available from: **https://www.tul.cz/document/7396**

Czech Republic. Zákon č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách) [Act No. 111/1998 Coll., On Higher Education and on Amendments to Other Acts (Higher Education Act)]. *Sbírka zákonů České republiky*. 1999, částka 39, p. 5388 - 5419. Available also from: **https://www.zakonyprolidi.cz/cs/1998-111**.

MEASURING THE VALUE OF OPEN ACCESS ETDS IN ALGERIAN DIGITAL REPOSITORIES: AN EVALUATIVE STUDY

Khaled Mettai

khaled.mettai@univ-constantine2.dz

LERIST Lab, The Institute of Library Science and Documentation, Abdelhamid Mehri University Constantine 2, Algeria

Behdja Boumarafi

behdja.boumarafi@univ-constantine2.dz

Abdelhamid Mehri University Constantine 2, Algeria

Abstract

Over the past years, grey literature in general and electronic digital theses (EDTs) in particular have been becoming more and more digital. Algeria is ranked first among Arab countries with fifteen (15) digital repositories. EDTs represent a large percentage of the repositories' content. Fourteen (14) of these repositories have policies for the collection of theses and dissertations, as well as other types of documents such as articles and reports. The usage of ETDs by undergraduates has been increasing exponentially. This study aims to highlight and evaluate the tools used to measure the usage of ETDs in the digital repositories, their availability and the evaluation methodology.

Keywords
Digital repositories, usage statistics, ETDs, open access, grey literature, evaluation of ETDs, University of Constantine 2.

Introduction
"We cannot call a digital library or electronic publishing system a success if we cannot measure and interpret its use". (Bishop 1998).

Digital repositories have made significant changes as regards the publishing industry and scholarly communication over nearly two decades, nonetheless institutional repositories (IRs) have a conflicted history in terms of purpose despite being always closely associated with the open access movement and the publishing of research produced by universities through academic community members.

IRs have a significant role in providing visibility for the research outputs of the academic community, both locally and internationally. Algeria, with fifteen (15) digital repositories registered in the OpenDOAR directory, is ranked first in the Arab countries for its high number of institutional repositories. 46% of these repositories are using metrics such as Google Analytics. However, the fact remains that some repositories have already been launched and exist yet are not registered in international directories like OpenDOAR or ROAR. These projects have been launched since 2013 with the aim of increasing the ranking of the universities by providing visibility to their publications at an international level.

Many digital repositories have been undergoing technical issues affecting their visibility and web presence. A study conducted by Hachani (2017) about the web presence of three countries in the Maghreb, namely Algeria, Tunisia and Morocco, concluded that these open repositories do not seem to implement a clear open access policy as most of them restrict access to registered users, which contravenes the essence of open access philosophy allowing access to scientific literature, and this has negatively impacted the performance of those repositories and the ratio of their open access literature. This seems to be the reason why users are not using or discovering these tools.

Bouderbane et al. (2018) indicated that 80% of the sample respondents, consisting of university teachers in Algeria, mentioned that they “rarely” use grey literature in their information searches because of the tremendous obstacles they face when attempting to use these resources, while 15% of them affirmed they “never” used grey literature documents in their research because of the harsh and complex environment surrounding these specific resources. However, the different understandings of the term 'use' represent a challenge to clarifying what the term can be considered to mean. At institutional level, “popular” metrics (tweets, Facebook shares, mainstream media mentions, etc.) can be appreciated in terms of public relations. It may be difficult for library administrators to see the value in popular metrics beyond marketing, but item level metrics are more and more important today for scholars and, to some extent, academic departments because scholarly exchanges are increasingly taking place via online networks, including social media platforms (Lavoie et al. 2014).

This emphasizes the importance of metrics in giving different understandings of use through tweets, Facebook shares and mentions), which can be a useful reference for information seekers in the social web. Repository managers can take advantage of these new metrics to meet the different needs of stakeholders and communicate with them. Different metrics serve different needs of stakeholders in the academic community, leading to continued use of repository services and benefits from their content in terms of access, sharing, use and reuse. This whole activity can also be tracked, assessed and counted to increase the value of the content - especially grey literature such as ETDs - and faculty research, as well as other types of content. Altmetrics - new tools to measure impact in the social web - can add more usefulness to how the value of ETDs and other IR can be increased and assessed. As of 2014, only 9% of repositories collected and displayed altmetrics or citation metrics for their content. These repositories tend to source their altmetrics data from Altmetric and PlumX (Rehemtula et al. 2014). Many IR platforms track usage statistics “out of the box” (Konkiel and Scherer 2013). Pageviews and downloads are commonly reported on public-facing pages, while systems track and privately report other useful information (top search queries, unique visitors, etc.) to system administrators or individual authors. This might be a good option for IRs that do not issue permanent identifiers like Handles or DOIs - they could monitor social media sites for mentions of relevant URLs instead. Measuring the success of any online collections and physical collections, and collecting and interpreting these numbers in an assessment tool to gain new insights about collections and user attitudes are top priorities for any repository manager. Metrics are necessary for repository managers in assessing the services they provide to their university community. For a young repository, generating quick metrics is essential (Gibbons 2013).

Digital repositories collect many different types of statistics to add value to their collections of ETDs and increase their usefulness as rich scholarly resources for improving the quality of and disseminating research for undergraduate students and the whole research community. The integration of download metrics and social media tools might strengthen their already significant impact and make them a good choice for repository managers to better communicate with different stakeholders in the community, although both aggregators provide varying degrees of important qualitative data behind the numbers they report. For example, in addition to seeing that items in your repository have been mentioned 17 times on Wikipedia, you can also see exactly what has been written about them. PlumX, however, does report some metrics that do not have the underlying qualitative, auditable data available for review (Konkiel et al. 2015) due to the increasing importance of the social web in giving more insights to research and going beyond numbers to ways to assess and measure the impact of different types of IR content.

Methodology

Content analysis of the Algerian digital repositories has been adopted as an evaluation tool to assess the metrics used to evaluate electronic theses and dissertations with a sample size of twelve (12) repositories out of the total number of fifteen (15). Three were excluded due to technical problems. The approach sought to analyse the metrics used to increase the value of the electronic theses and dissertations through the observation method, which was used to observe each repository with its metrics, and to analyse and evaluate the different metrics used to increase the value of the ETDs in each of the repositories.

The following repositories were examined and evaluated according to the commonly used repository metrics mentioned below:

- Item downloads
- Number of items in the repository
- Item uploads
- Location of visitors
- Participating units
- Participating faculty

Note*: other metrics also have been analysed and mentioned besides those mentioned above:

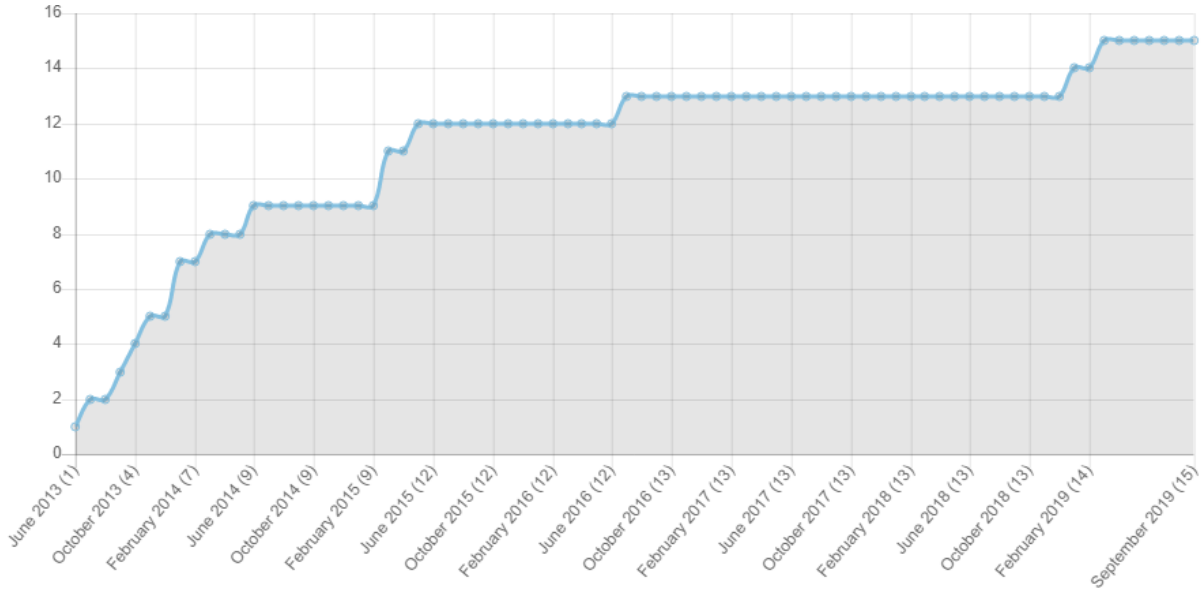

Figure 1: The growth in digital repositories in Algeria (OpenDOAR, 2019)[1].

The Development of the Algerian Digital Repositories

Digital repositories of the Algerian universities began to be established from 2013. Fifteen (15) repositories have been registered as of September 2019. The main purpose of those projects is to disseminate and facilitate access to the research outputs of the universities and research centres in the country. The different information resources being uploaded to the repositories are of huge importance to support the education in the universities and also the academic community. In order to make this work, achieving high visibility in the usage of the digital theses and dissertations is paramount. Moreover, ETDs have been extremely important in conducting research and, as rich information resources, ETDs tend to overcome many obstacles such as the growing size of paper dissertations and the related shortages of available storage capacity at traditional libraries, the difficulties in reporting and assessment the impact of those resources, and the need to provide access to the old theses and dissertations, which are a memory heritage for both the institutions and the users.

[1] OpenDOAR .**https://v2.sherpa.ac.uk/view/repository_by_country/dz.default.html** (accessed 13 September 2019).

Nowadays, theses and dissertations exist only electronically (as ETDs); it is often the IR that preserves and provides access to this valuable scholarly and institutional literature. In addition, the IR can provide support for a wide variety of "supplementary" files for ETDs, crucial information which, because of its format, cannot be included in the PDF submitted to ProQuest. These files often include datasets, software, and multimedia content, as well as research protocols in the case of technical reports and ETDs (Kennison et al. 2013).

ETDs are key factors in making the open access movement a reality. Most of the content in Algerian digital repositories is grey literature and ETDs represent 55% of all the content in Algerian digital repositories. Consequently, there is a definite need to establish metrics that measure their use. This has generated increased international interest from the academic community since 1987, when work was first begun to resolve the many obstacles, ranging from ease of access to the need for long-term preservation, while the reasons were mainly to resolve the many issues related to printed theses. For instance, Aquil et al. (2014) mentioned that the digital libraries of ETDs promise to be extremely useful to scholars, especially in developing countries. The greatest advantages of ETDs lie in avoiding duplication in research work, ensuring fast access to information, promoting resource sharing and providing a permanent solution to the problem of space.

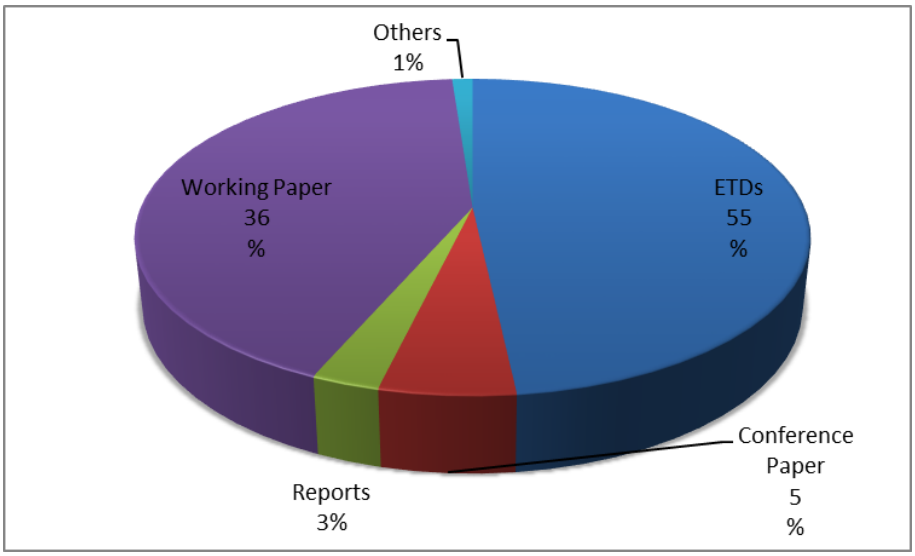

Figure 2: ETDs compared to the total content of Algerian digital repositories.[2]

Metrics to Enhance the Value of ETDs in Digital Repositories

The tracking of IR activity is of great importance in evaluating the sustainability of IRs, enhancing their ability to assess and predict the changing needs of users to then help repository managers evaluate their content and make decisions. Download statistics and other expressions of information use, like page visits, are regarded as additional metrics for information impact (Tsakonas and Paratheodorou 2009). As highlighted in Figure 2, 55% of the content in Algerian digital repositories is ETDs, and 50% of the repositories use metrics from third parties such as Google Analytics. As a new publishing system, repositories include many different tools to increase the value of the content. Once a faculty member has one article in a repository and begins to receive monthly statistics on downloads and use, they have greater motivation to ask libraries to include the rest of their materials (Giesecke 2011). Due

[2] Author is the creator of the picture.

to the importance of metrics in assessing the usage of online materials and the failure of traditional tools to resolve the multiplying issues and to keep abreast with new technologies in this arena, the advancement in technology which heralded ETDs brought leverage to efforts to address the problems of printed TDs.

Awareness of ETDs was increased with the introduction of the Database of African Theses and Dissertations (DATAD) by the Association of African Universities (AAU). Anunobi and Onyebinama (2011) posit that several projects are being developed in the European library community to set standards and develop tools for IR statistics reporting. These include PIRUS2, which has been funded as the IRUS-UK service (O'Brien et al. 2017). With the increasing importance of the social web, outreach has become more possible for electronic theses and dissertations with new metrics to spread ETDs in the social web environment. In this regard, Palmer (2013) argued that the benefits of introducing altmetrics into a repository include the possibility to deliver impact measures for publications that have not been published in scientific journals, such as posters, dissertations, datasets, and books. Moreover, altmetrics and PlumX can add more significant metrics for repositories, contribute towards disseminating EDTs, and measure impact and references on the social web. Konkiel et al. (2015) claimed that Altmetric collects altmetrics for any content in an institutional repository or digital special collection, including institutional repository items that have multiple versions stored elsewhere (i.e. the publisher's "version of record" of a journal article stored on the publisher's site, plus the author's preprint stored in the repository). The reports it generates include both metrics for the preprint in the repository as well as metrics related to the publisher's version on the journal's website. This might be a good option for IRs that do not issue permanent identifiers like Handles or DOIs that could monitor social media sites for mentions of relevant URLs instead.

Table 1: Metrics in the Algerian digital repositories[3]

Metrics in repositories	Percentage
Item downloads	50%
Location of visitors	58.33%
Number of items in repository	75%
Recent submissions	83.33%
Item uploads	25%
Total visits	58.33%
Total visits per month	58.33%
Top country views	58.33%
Participating units	75%
Participating faculty	66.66%
Scopus H-index for journals	10%
Impact factor	10%

As can be seen from Table 1, the recent submissions metric ranked first, followed by number of items in repository with 83.33% and 75% respectively. Then came participating faculty and participating units with 75% and 66%, which are considered in-house metrics. The metrics related to Google Analytics are represented through location of visitors, total visits, total visits per month and top country views, all with 58.33%, then item downloads with 50%. This can be counted at individual and community level as well. In addition, most repositories count the

[3] Author is the creator of the table.

number of ETDs in a specific community, and the ones included in each community. Regarding metrics related to the growth of the repository, represented through item uploads, 25% of the repositories provide item uploads in a specific period of time, which has an impact on determining the growth of the repositories. The Scopus H-index for journals and impact factor metrics have been found to be less commonly available in the Algerian repositories with 10% for each, although those metrics are specifically for journals rather than ETDs. According to (OpenDOAR, 2019), 73% of the repositories use DSpace software, which also offers statistics at the individual record and collection level, and just 7% use Eprints, whereas 20% use others, like WordPress and self-built CMS. Presumably, the use of non-standard software for repositories makes it difficult to set up reports and strong statistical services that fully meet the need to measure the usage of the repositories with statistical projects either with commercial services or academic ones and free from third parties such as Google Analytics services, which has been adopted by 46% of the repositories.

Discussion and Conclusion

The importance of repository metrics as value-added tools has been increasingly decisive in the online environment for determining and evaluating the success of repositories and content. Repository managers should be aware of the importance of metrics and their contribution in adding impact and value to their content. The results of the present study indicate that half the Algerian repositories have metrics related to usage statistics from item downloads, total monthly visits and top city views, counted at individual, collection or community levels, the same for location of visitors using metrics like item downloads, which has not been enabled by some repository managers, although these are actually integrated in their repositories, especially those using DSpace and EPrints.

Most other software packages are old versions, and this is a problem in view of the new trends in the complementary process of the repository, from collection to evaluation of the content, increasing the value of EDTs and broadening access to wider audiences worldwide, giving ETDs the maximum possible exposure. The need to establish a culture of metrics has been increasingly clear as the age of data has penetrated all fields. The lack of this culture prevents ETD and grey literature publications from being discovered and used. Online usage data has become vital as different metrics serve different purposes for different stakeholders in the digital repositories projects. Every stakeholder has its own significant metrics which demonstrate the value of its contribution to the repository in numbers. This is because metrics give vitality to the electronic theses and dissertations, and enable authors to see the impact of their theses. This consequently gives them motivation to place their work in repositories, and establish connections through the impact of their work. This process is beneficial for many parties in relation to IR projects.

Recommendations and Suggestions

The Algerian digital repositories lack a strong culture of metrics, since half the metrics relating to the measurement of use are less available compared to other metrics, at 50% and 58.33% respectively, while metrics of usage should be top priorities for repository managers. This might be a potential barrier to obtaining funding for enhancing the value of ETDs (the quantities of which are growing rapidly due to the mandatory depositing of digital copies in libraries) in particular, and grey literature in general.

Data standardization is becoming of significant importance for repositories when measuring the impact of research. Schufreider and Romaine (2008) demonstrated that most respondents felt the greatest challenges relating to usage statistics were the amount of time involved and a lack of consistency/absence of standards. Needham and Lambert (2019 a) emphasized that there is no single, perfect measure to assess value and impact, and institutions may use a range of metrics including citations, page views and altmetrics. However, download statistics are among several measures used to demonstrate value and are the focus of this article. Usage metrics are a key aspect in terms of understanding how publicly available research is being used. Tracking, monitoring and benchmarking the usage of scholarly resources supports an understanding of an institution's research. Over the past 15 years, the COUNTER standard has been integral to facilitating the recording and reporting of online usage statistics in a consistent, credible and comparable way, as current research is prompting an increasing interest in more granular metrics, including item-level and research data metrics. The COUNTER standard, now in its fifth iteration, has evolved over time in response to a changing environment and evolving requirements. COUNTER CoP release 5 (R5) standardizes usage metrics for e-resources, including journals, books, databases and platforms due to the increasing interest in more granular metrics, including the item level and research data metrics. (Needham and Lambert 2019b). Most importantly, the need for repositories to collaborate using standardised metrics today is the subject of great international interest, and many projects and initiatives like COUNTER have been released over the past decade, such as SUSH (Standardized Usage Harvesting Initiative) and the OpenAIRE Usage Statistics Service to meet the needs of use granularity and standardization usage metrics for e-resources.

Electronic theses and dissertations can add more value to the research and the institution. We have observed significant variations between repositories regarding the metrics they offer, generated using various platforms which in turn use different metrics to represent different meanings on impact, and the ability to use the Google Analytics service for this purpose without any cost. Based on the outcomes of the research, two suggestions have been made:

Use Metrics to Argue for Funding

Many repository networks like IRUS-UK (Institutional Repository Usage Statistics UK) have appeared to measure the use of online content. Funders, whether people or organisations, need to know that the money they invest has a positive impact on the project they have invested in and the outcomes the project intends to produce. A strong, targeted metrics infrastructure is needed to serve as a driver to increase the impact of the content, and to assess the sustainability of a project, its use, growth and success, and to measure the advantages for all the members of the academic community. Such services, like Plumx, need effort and financial support, which can only be received by advancing and spreading a real culture of metrics and establishing meaningful metrics to measure the impact of the content and to evaluate its use and reuse for members of the academic community members and others.

Altmetrics as a Value-Added Service

The further development of metrics technology in alignment with the social web is needed to expand access and referrals to ETDs in the social web; social media platforms are powerful altmetrics tools to ensure the outreach and measure the impact of ETDs and grey literature in general in a new context and in a much more complex environment. Altmetrics might be useful for repositories to measure the visibility of their content on social media, bookmarking platforms complementing download/citation metrics and ETDs, as these represent more than half the content that will be exposed to wider and global audiences.

References:

ANUNOBI, Chinwe V. and Colette O. ONYEBINAMA, 2011. ETD Initiatives at the Federal University of Technology, Owerri (FUTO): Successes, Challenges, Prospects. *Proceedings of the 14th International Symposium on Electronic Theses and Dissertations* [online], Cape Town, South Africa, 13-17 September. [Accessed 20 September]. Available from: **http://dl.cs.uct.ac.za/conferences/etd2011/papers/etd2011_anunobi.pdf**

AQUIL, Ahmed, Sulaiman ALREYAEE and Azizur RAHMAN, 2014. Theses and Dissertations in Institutional Repositories: an Asian Perspective. *New Library World* [online], **115(**9/10), 438 – 451. [Accessed 20 September]. Available from: **https://www.emerald.com/insight/content/doi/10.1108/NLW-04-2014-0035/full/html**

BOUDERBANE, A, T. BENKAID KESBA and N. GAMOUH, 2018. The University Teacher's Attitudes Towards Grey Literature: a Survey at the University of Constantine. *11th Conference on Grey Literature and Repositories: Proceedings* [online]. Prague: National Library of Technology, 2018 [Accessed 20 September]. ISSN 2336-5021. Available from: **https://nusl.techlib.cz/en/conference/conference-proceedings**

GIESECKE, J., 2011. Institutional Repositories: Keys to Success. *Journal of Library Administration* [online], **51**(5-6), 529-542. [Accessed 13 September 2019]. DOI: **10.1080/01930826.2011.589340**

GIBBONS, S., 2013. Benefits of an Institutional Repository. *Library Technology Reports* [online], **40**(4), 11–16.

HACHANI, S., 2017. Algeria's, Morocco's and Tunisia's Presence in the Directory of Open Access Repositories (DOAR) and the Registry of Open Access Repositories (ROAR): A Comparative Study of the Ratio of Open Access Material. *Open Information Science* [online], **1**(1), 56–70. [Accessed 13 September 2019]. ISSN 2451-1781, DOI: **https://doi.org/10.1515/opis-2017-0005**

KENNISON, R., S. L., SHREEVES and S. HARNAD, 2013. Point & Counterpoint - the Purpose of Institutional Repositories: Green OA or Beyond? *Journal of Librarianship and Scholarly Communication* [online], **1**(4), p.eP1105. [Accessed 13 September 2019]. DOI: **https://doi.org/10.7710/2162-3309.1105**

KONKIEL, S. and D. Scherer, 2013. New Opportunities for Repositories in the Age of Altmetrics. *Bulletin of the American Society for Information Science and Technology* [online], **39**(4), 22–26. [Accessed 13 September 2019]. DOI: **https://doi.org/10.1002/bult.2013.1720390408**

KONKIEL, S., M. DALMAU and D. SCHERER, 2015. *Altmetrics and Analytics for Digital Special Collections and Institutional Repositories* [online]. Figshare [Accessed 13 September 2019]. Available from: **http://dx.doi.org/10.6084/m9.figshare.1392140**

LAVOIE, B. et al., 2014. *The Evolving Scholarly Record* [online]. OCLC [Accessed 13 September 2019]. [Available from: **http://oclc.org/research/activities/scholarcomm.html**

NEEDHAM, P. and J. LAMBERT, 2019. Institutional Repositories and the Item and Research Data Metrics Landscape. *Insights* [online], **32**(1), 26. [Accessed 20 September] DOI: **http://doi.org/10.1629/uksg.478**

O'Brien, P. et al., 2017. RAMP – the Repository Analytics and Metrics Portal. *Library Hi Tech* [online], **35**(1), 144-158. [Accessed 13 September 2019]. Available from: **https://doi.org/10.1108/LHT-11-2016-012**2

OpenDOAR [online]. [Accessed 13 September 2019]. **https://v2.sherpa.ac.uk/view/repository_by_country/dz.default.html**

PALMER, L. A. (2013). Altmetrics and Institutional Repositories: A Health Sciences Library Experiment. University of Massachusetts Medical School. *Library Publications and Presentations* [online]. Paper 142. [Accessed 13 September 2019]. DOI: **https://doi.org/10.13028/qmbg-qc96**. Available from: **https://escholarship.umassmed.edu/lib_articles/142**

REHEMTULA, Salima, Maria Rosa DE LURDES, M., Paulo LEITÃO and Rosario Arquero AVILÉS. Altmetrics in Institutional Repositories: New Perspectives for Assessing Research Impact. *Libraries in the Digital Age (LIDA) Proceedings* [online], vol. 13. [Accessed 20 September]. Available from: **http://ozk.unizd.hr/proceedings/index.php/lida/article/view/141**

SCHUFREIDER, B. and S. ROMAINE, 2008. Making Sense of your Usage Statistics. *The Serials Librarian* [online], **54**(3-4), 223-227. [Accessed 20 September] Available from: **http://dx.doi.org/10.1080/03615260801974164**

TSAKONAS, G. and C. PAPATHEODOROU, 2009. *Evaluation of Digital Libraries - An Insight into Useful Applications and Methods*. Oxford: Chandos. p.182.

INCREASING THE VISIBILITY OF GREY LITERATURE IN ALGERIAN INSTITUTIONAL REPOSITORIES

Babori Ahcene

ahcene.babouri@univ-constantine2.dz

The Institute of Library Science and Documentation, Abdelhamid Mehri University Constantine 2, Algeria

Aknouche Nabil

nabil.aknouche@univ-constantine2.dz

The Institute of Library Science and Documentation, Abdelhamid Mehri University Constantine 2, Algeria

Abstract

The global trend of establishing digital repositories in institutions of higher education has accelerated as they are becoming important to research and scientific institutions for evaluating criteria. Institutional repositories can also help promote the practical impacts of research, which is not only beneficial to the community but can help demonstrate the importance of the work and enhance the researcher's reputation. As of August 2019, the Directory of Open Access Repositories holds 4 140 repositories, including 15 from Algeria, consisting theses, conference papers, books and journals. This paper aims to discover the benefits and impacts of institutional repositories and their role in supporting scientific research and enhancing the global visibility and impact of grey literature through answering the following questions: What are the best practices to improve the visibility of grey literature? How can we achieve high grey literature visibility in Algerian institutional repositories?

Keywords
Grey literature, institutional repositories, visibility, Algerian institutional repositories

---

Introduction
Grey literature is often the best source of up-to-date research on certain topics. While there are a number of sources where grey literature can be found, institutional repositories are often consulted first. For years, universities have been building their repositories to collect the grey literature published by scholars, and potentially increasing the public value, ranking, prestige, and visibility of the researchers. In addition, the availability of open source institutional repository systems has encouraged a proliferation of institutional repositories worldwide, particularly among academic and research institutions. Depositing grey literature in Open Access repositories will increase visibility and citations through the removal of barriers to knowledge sharing.

The development of the institutional repositories in Algeria is a result of the desire to communicate intellectual output, and increase visibility and impact, because the poor visibility of research findings coming out of institutional repositories is a major challenge for Algerian scholarship.

Objectives of the Research
In this paper we explore factors that enhance the visibility of grey literature in institutional repositories. As such, our study has sought to achieve the following objectives:

- Assess the status of grey literature in Algerian institutional repositories.
- Indicate best practices and means to increase grey literature visibility in institutional repositories.

The study addresses and responds to the following questions:

- Which are the best practices to improve the visibility of grey literature?
- How can we achieve a high grey literature visibility in Algerian institutional repositories?

Open Access Movement and Repositories
Many universities, government agencies, and other research funders are embracing the benefits of making works freely available and are adopting open access policies. These policies generally require works create in university faculties or developed under agency or foundation sponsorship to be made openly accessible (Rubow, Shen, & Schofield 2015). Open access repositories provide free research access for users and maximise the visibility and impact of the researches, while also focusing on “serving the interests of faculty researchers and teachers by collecting their intellectual outputs for long-term access, preservation, and management”. (Carr, White, Miles, & Mortimer 2008)

Visibility of Grey Literature:

Open access repositories play a variety of roles in the scholarly communication system, and these roles continue to expand and evolve. To date, their primary functions have been to provide visibility and open access to research outputs. Additional open access repositories significantly raise the visibility, use, and citation counts[1] of deposited materials. Numerous studies over the last 15 years have reported on the citation advantage of open access content in general (COAR 2015).

State of Algerian Repositories:

As of August 2019, OpenDOAR, a service that monitors repositories, listed 15 repositories in Algeria. These graphs (Figures 1 – 4) show the growth in numbers of repositories in Algeria, the types of software used, subject areas, and language content. Most Algerian institutional repositories collect grey literature, e-theses, conference proceedings, working papers, and reports. The vast majority of repositories in Algeria are institutionally hosted and managed by research institutions and universities, while 73% of these repositories used DSpace open source software, 7% used EPrints software, and 20% used CMS (Content Management System) software.

The subjects of repositories are ubiquitous, multidisciplinary, and widespread, including, social sciences, business, economics, science, and mathematics. In addition, French is the most common language used in Algerian institutional repositories, followed by English, then Arabic.

[1] Citation counts: referring to a higher citation count

OpenDoar Algeria Statistics[2]:

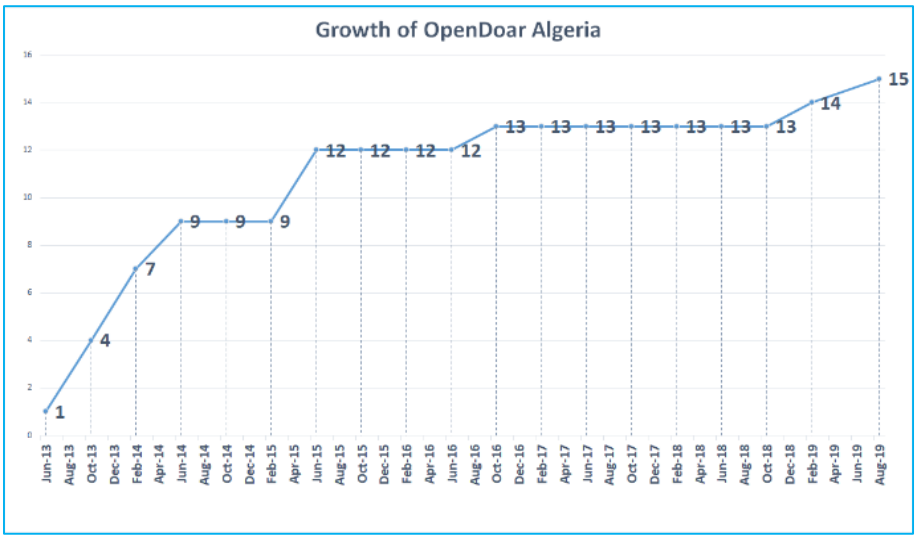

Figure 1: OpenDOAR Algeria Growth (data source: OpenDOAR)

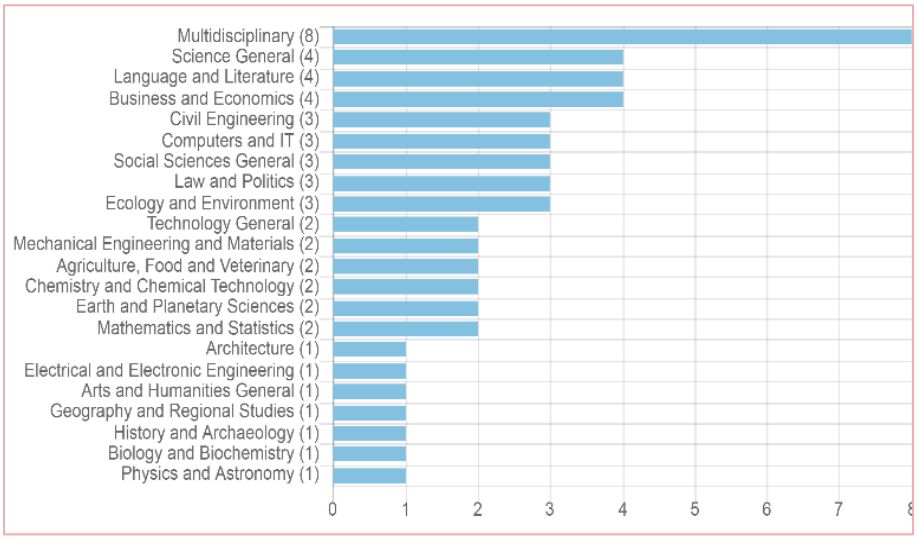

Figure 2: OpenDOAR Algeria Subjects (data source: OpenDOAR)

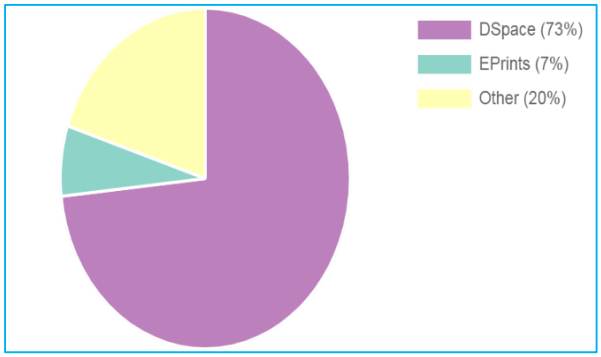

Figure 3: OpenDOAR Algeria Software (data source: OpenDOAR)

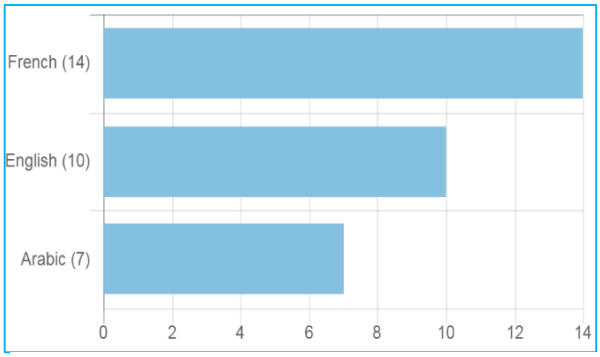

Figure 4: OpenDOAR Algeria Languages (data source: OpenDOAR)

[2] https://v2.sherpa.ac.uk/view/repository_by_country/dz.default.html

We could determine the exact date of creation of each of the 15 repositories, while the first repository in Algeria was launched in 2013. In comparison with arXiv, the first repository established in the world, and considering that the first repository launched in Africa was in 2005, Algerian institutions and universities have been extremely late to engage in the development and establishment of digital repositories for several reasons, including. "A l*ack of information on Open Access the concept is new and not popularized enough, thus implementation is not rapid; a lack of clear institutional and national policy on Open Access; the difficulty of securing long-term funding and getting commitments from more institutions to join the Open Access community*" (Hachani & Tennant 2017).

Grey Content in Algerian Institutional Repositories

According to OpenDOAR data, 55% of Algerian institutional repositories contain theses and dissertations, 36% working papers, 3% conference papers, 3% reports and 1% other resources (see fig. 5).

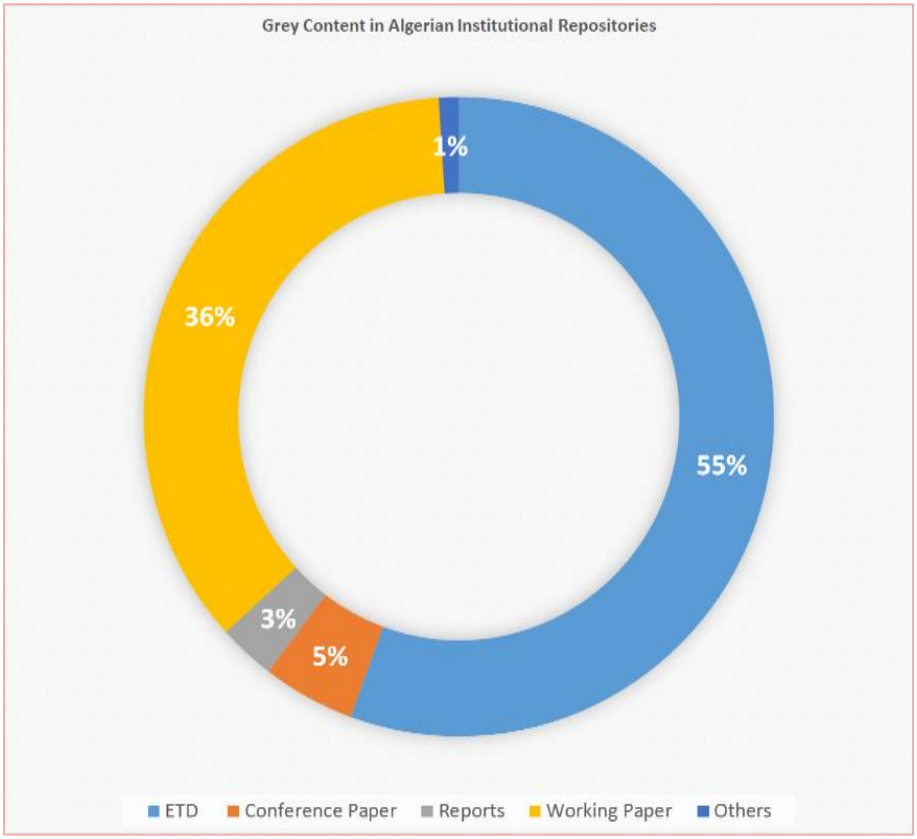

Figure 5: Grey Literature in Algerian Institutional Repositories

Current Best Practices and Ways to Increase Grey Literature Visibility in Institutional Repositories

International Registries

The best way to make a repository more visible and better known is to register it in international directories like OpenDOAR, OpenROAR, DSpace instance, OAIster, Repository Map, BASE

(Bielefed Academic Search Engine), DRIVER (Digital Repository Infrastructure Vision for European Research), SCIRUS and so on. The Figure 6 below shows the number of registered Algerian institutional repositories.

The Directory of Open Access Repositories - OpenDOAR (**https://v2.sherpa.ac.uk/opendoar/)** is an authoritative directory of academic open access repositories. Services offered by the directory include searches for repositories, searches of repository contents, lists of repositories, and repository statistics.

The aim of ROAR (**http://roar.eprints.org/**) is to promote the development of open access by providing timely information about the growth and status of repositories throughout the world. Open access to research maximizes research access and thereby also research impact, making research more productive and effective (Okpala 2013).

BASE (Bielefed Academic Search Engine, **https://www.base-search.net/**) is a registered **OAI service provider**. Database managers can **integrate the BASE index** into their local infrastructure (e.g. meta search engines, library catalogues). There are also several **tools and services** for users, and database and repository managers.

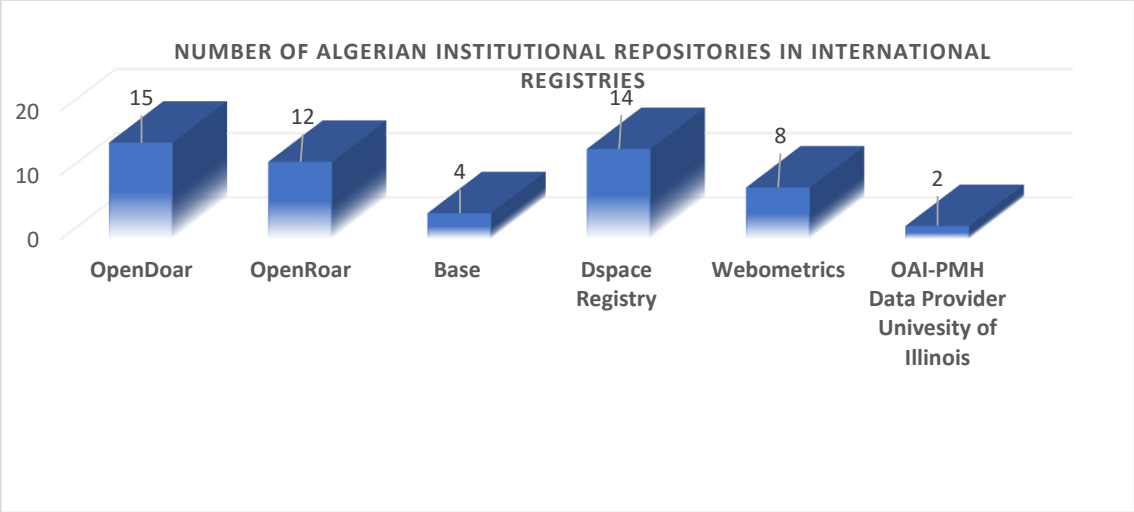

Figure 6: Number of Algerian Institutional Repositories in International Registries

Metadata standard and metadata quality

Metadata plays a key role in describing, accessing, and managing digital objects using different formats and media. The purpose of metadata is to offer the user multiple access points (e.g. author, title, and subject.). However, if the metadata are incorrect, the resources will not be adequately represented in institutional repositories and will remain invisible to users (Tmava & Alemneh 2012). Grey literature in institutional repositories requires specific metadata for identification and bibliographic description (Schöpfel, Prost, & Le Bescond 2011).

Interoperability

Interoperability is key to the success of your institutional repository, ensuring that content stays portable and compatible with on-campus systems, as well as complying with OAI (Open Archive Initiative) data harvesting, improving your SEO (Search Engine Optimization) capabilities. (Elsevier 2018)

Online Profile and Scholarly Identifier

Creating and maintaining online profiles will help to raise the impact of one's research outputs on the research community and the greater public. An online profile is an essential tool to disseminate one's research and publication output. Scholarly identifiers and online profiles such as ResearcherID and ORCID (Open Researcher and Contributor ID) provide a solution to the author ambiguity problem within the scholarly research community. (Ale Ebrahim 2017a). Creating a public Google Scholar profile is an easy way to increase a work's findability and also provides other benefits such as an author H-index, citation counts, and more.

ORCID (Open Researcher and Contributor ID) is a persistent digital identifier that distinguishes you from every other researcher. Having a unique identifier ensures that the bibliometric data about you and your body of work is accurate and correctly linked to your researcher profile. It also improves the visibility of the research. The academic social networking makes one’s work more widely discoverable and easily available. The two best known examples of academic social networking are ResearchGate and Academia.edu.

The ResearchGate and Academia academic social networking sites have become important components of the scholarly communication landscape.

Placing your publications and presentations on ResearchGate and Academia will make it easier for others to encounter your work, not only because they are available on a social network, but also because they improve the search engine optimization (SEO) Search Engine Optimization of your research. A recent study found that papers uploaded to Academia.edu receive a 73% boost in citations over 5 years (Ale Ebrahim 2017b).

Google Scholar Search Engine

Google Scholar (GS) has become the best free search engine available for institutional repository content (Arlitsch & S. O’Brien 2013). The Table 1 and Figure 7 below shows the GS index ratio for the repositories indexed by GS. We collected the data from the webometrics website, which shows how many records GS has been indexed in.

Table 1: Google Scholar Index Ratio of Algerian Institutional Repositories

Repositories	Index Ratio
Archives Numériques de l'Université Frères Mentouri Constantine	23.29%
Dépôt Institutionnel de l'Université Abou Bekr Belkaid Tlemcen	83.03%
Dépôt institutionnel de l'Université Mohamed Boudiaf de M'Sila	81%
Institutional Repository Université Hassiba Benbouali Chlef	0%
Production Scientifique de l'Université de Bouira	0%
Production Scientifique de l'Université M'hamed Bougara Boumerdès	0%
University of Biskra Repository	0%
University of Biskra Theses Repository	83.85

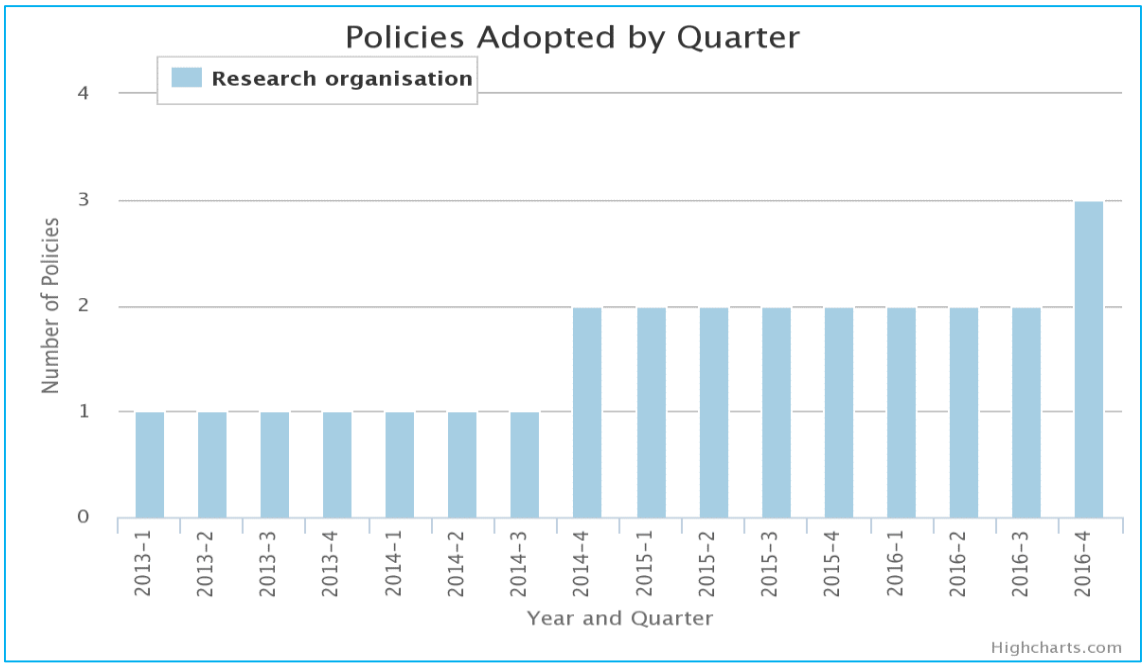

Figure 7: Number of Records vs. Records Indexed by Google Scholar

Mandatory Self-Archiving of Grey Literature:

The mandatory institutional or self-depositing of grey literature is promoted by Stevan Harnad: green road (self-deposit) to free online full-text access to peer-reviewed literature, through an explicit and institutional mandatory policy in order to obtain commitment by close to 100% of the authors (Schöpfel, Prost, & Le Bescond 2011). According to ROAR Map, only three universities in Algeria have adopted mandatory deposit policies (see Figure 8).

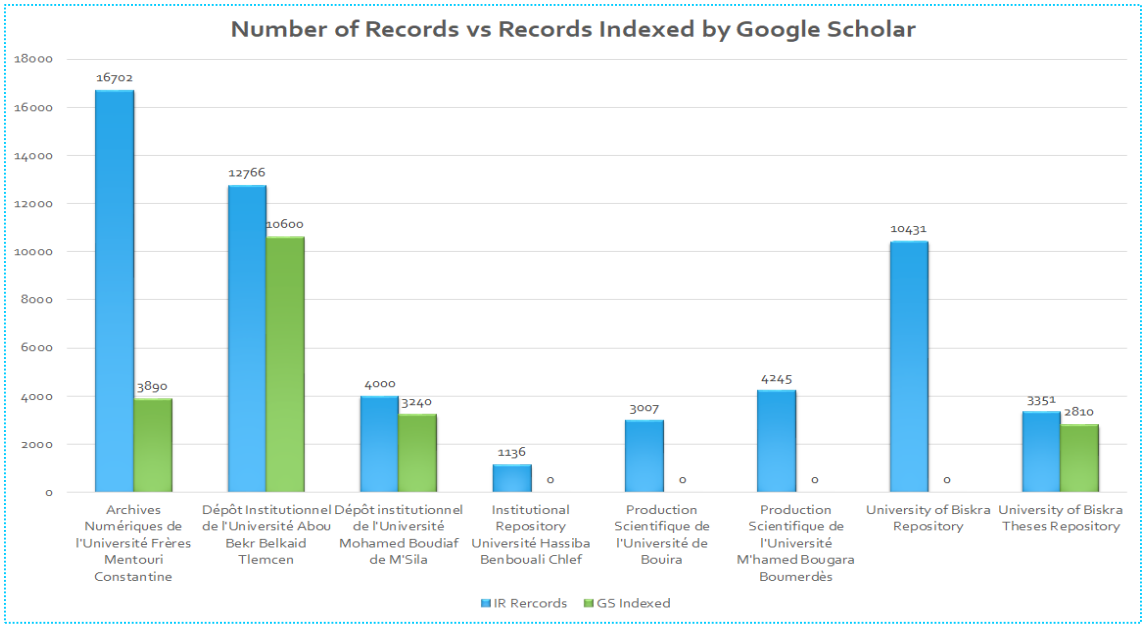

Figure 8: Number of ROAR Map Algeria policies

Conclusion and Recommendations

Many academic institutions in Algeria have great potential for the implementation of institutional repositories to preserve, increase visibility, and widen access to the research outputs of universities, including grey literature.

The use of institutional repositories and good institutional policies will enhance the availability and accessibility of grey literature as a key information resource that supports teaching research and learning at universities (Samzugi 2017).

Our study recommends the following:

- Upgrade the DSpace software version for institutions using an earlier version, because we have observed that Algerian institutional repositories are using DSpace version 1.7.
- Enable Google Scholar indexing features.
- Develop a national open access policy for institutional policies requiring mandatory deposit.
- Allow full-text downloading of the grey literature in the repositories.
- Register the repositories in international directories.
- Motivate faculty authors to self-archive.
- Raise awareness among scholars of the importance of improving their presence in academic social networks. (Ale Ebrahim 2017a)

References

ALE EBRAHIM, N., 2017a. *Create Online Researcher's Profile to Increase Visibility* [online]. [Accessed 10 September 2019] Available from: **https://doi.org/10.6084/m9.figshare.4959788.v1**

ALE EBRAHIM, N., 2017b. *ResearchGate and Academia: Networks for Researchers to Improve Research Impact* [online]. [Accessed 10 September 2019] Available from: **https://dx.doi.org/10.6084/m9.figshare.3464156.v1**

ARLITSCH, Kenning and Patrick S. O'BRIEN, 2013. *Improving the Visibility and Use of Digital Repositories Through SEO: A LITA Guide* [online]. Chicago: American Library Association, [Accessed 10 September 2019]. LITA Guide. ISBN 978-1-55570-906-8. Available from: **https://ebookcentral.proquest.com/lib/techlib-ebooks/detail.action?docID=1187219**

CARR, Less, Wendy WHITE, Susan MILES, and Bill MORTIMER, 2008. Institutional Repository Checklist for Serving Institutional Management. *Third International Conference on Open Repositories: 01 - 04 Apr 2008* [online]. Southampton: University of Southampton, [Accessed 10 September 2019]. Available from: **https://eprints.soton.ac.uk/23929/**

COAR, 2015. *Promoting Open Knowledge and Open Science: Report of the Current State of Repositories* [online]. Confederation of Open Access Repositories. [Accessed 10 September 2019]. Available from: **https://www.coar-repositories.org/files/COAR-State-of-Repositories-May-2015-final.pdf**

ELSEVIER, 2018. *Digital Commons™: 10 strategies to expand your institution's global research visibility with a next-generation IR* [online]. [Accessed 10 September 2019] Available from: **https://www.elsevier.com/__data/assets/pdf_file/0006/794022/PLS_DDD_EN_FS_WEB.pdf**

HACHANI, Samir and Jon TENNANT, 2017. The state of Open in Algeria: an in-depth view with Samir Hachani. *Open Science Interviews* [online]. [Accessed 10 September 2019] Available from: **https://blog.scienceopen.com/2016/10/the-state-of-open-in-algeria-an-in-depth-view-with-samir-hachani/**

OKPALA, Heler Nneka, 2013. *Access Tools and Services to Open Access: DOAR, ROAR, SHERPAROMEO, SPARC and DOAJ* [online preprint]. Lokoja: Librarians Registration Council of Nigeria (LRCN) Workshop, [Accessed 10 September 2019]. Available from: **https://www.academia.edu/25282158/ACCESS_TOOLS_AND_SERVICES_TO_OPEN_ACCESS_DOAR_ROAR_SHERPA-_RoMEO_SPARC_AND_DOAJ** Published version available from: **http://hdl.handle.net/10760/32498**

RUBOW, Lexi, Rachael SHEN, and Brianna SCHOFIELD, 2015. *Understanding Open Access: When, Why, & How to Make Your Work Openly Accessible* [online]. Berkeley, CA: Authors Alliance; Samuelson Law, Technology, and Public Policy Clinic [Accessed 2 September 2019]. ISBN 0-6925-8724-1. Available from: **https://authorsalliance.org/wp-content/uploads/Documents/Guides/Authors%20Alliance%20-%20Understanding%20Open%20Access.pdf** or **https://muse.jhu.edu/book/62064**

SAMZUGI, A., 2017. The Role of Institutional Repositories in Promoting Grey Literature in Academic Libraries in Tanzania. *University of Dar es Salaam Library Journal* [online], **12**(2), 3-21 [Accessed 15 August 2019]. ISSN 0856-1818. Available from: **https://www.ajol.info/index.php/udslj/article/view/184574**

SCHÖPFEL, Joachim, Hélène PROST and Isabelle LE BESCOND, 2011. Open Is Not Enough: Grey Literature in Institutional Repositories. *GL 13: Thirteenth International Conference on Grey Literature: The Grey Circuit from Social Networking to Wealth Creation. Washington, 5-6 December 2011* [online]. [Accessed 3 September 2019]. Available from: **https://archivesic.ccsd.cnrs.fr/sic_00908862**

TMAVA, Ahmet Meti and Daniel Gelaw ALEMNEH, 2012. *Enhancing Content Visibility in Institutional Repositories: Maintaining Metadata Consistency Across Digital Collections* [online poster]. University of North Texas Libraries, [Accessed 12 August 2019]. Texas Conference on Digital Libraries. Available from: **https://digital.library.unt.edu/ark:/67531/metadc86151/**

CLARIN-DSPACE REPOSITORY AT LINDAT/CLARIN

Pavel Straňák

stranak@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

Ondřej Košarko

kosarko@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

Jozef Mišutka

misutka@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

Abstract

CLARIN-DSpace is a fork of the well-known repository system DSpace, which is optimised for use as a data repository. It has been used by many centres in the CLARIN project, but its enhancements make it a good choice for other data repositories as well.

Keywords
Data repository, data citations, FAIR, community standards, language data, software tools, licensing, service integration, metadata exchange, open source

---

LINDAT/CLARIN and its Repository for Data and Software

LINDAT/CLARIN is a Czech project[1] which started as the national contribution to the European network CLARIN: "European Research Infrastructure for Language Resources and Technology." CLARIN, in the preparatory phase of the project, decided to inventorize language datasets and language processing tools in all member states, but quickly realised that something much more permanent than a list, in which records get outdated and links non-functional all the time, is needed. A living database of datasets and tools was needed, which would be always up to date, and data will be safely preserved. When CLARIN realised this, it was clear that repository systems are needed. Since CLARIN is an ERIC (European Research Infrastructure Consortium), infrastructure funded and organised by its member countries, these members decided to set up repositories in their national centres, and create a central discovery service: Virtual Language Observatory (VLO).[2] As an infrastructural project, CLARIN provides technical specifications and certification of compatible centres. There are several types of centres, but at the core there are Service and Data Providing Centres or so called CLARIN B-centres[3]; these run (among other services) repository systems in a way that should offer reliable long term data preservation, means to support direct data citation, and compatibility with CLARIN central discovery service VLO[4].

When the project started, there was no suitable repository system for hosting data and tools at any of the organisations that together form LINDAT/CLARIN (**https://lindat.cz**).

As the Czech CLARIN partner, LINDAT/CLARIN wanted to avoid building a system from scratch; instead, we looked for a repository system that was popular and robust, one we could believe will keep being updated and would allow us to modify it and share the modifications. The system would need to have a reasonable frontend that allows user submissions and offers standalone search functionality directly on the web, not relying solely on CLARIN's VLO. Ideally, it would be usable out of the box, while fulfilling CLARIN's requirements. Namely, support for persistent identifiers in the form of handles (this has recently changed and other PID[5] systems are allowed), support for CMDI[6] metadata harvested via OAI-PMH[7], support for federated authentication/authorization via SAML[8] protocol, and support for handling licenses of the data and tools submitted to the repository.

Around 2011, these requirements resulted in our choice of DSpace: the most popular repository system in the world that seemed easy to deploy and maintain and could do most of

[1] Project numbers: LM2015071 and CZ.02.1.01/0.0/0.0/16_013/0001781
[2] VLO available from: **https://www.clarin.eu/content/virtual-language-observatory-vlo**
[3] For more details about the types of centres see **https://www.clarin.eu/content/clarin-centres**
[4] The full list of requirements on CLARIN B-centre available from **http://hdl.handle.net/11372/DOC-78**
[5] Persistent Identifier
[6] Component MetaData Infrastructure, see **https://www.clarin.eu/content/component-metadata**
[7] The Open Archive Initiative Protocol for Metadata Harvesting
[8] Security Assertion Markup Language

what we needed out of the box. LINDAT team first modified DSpace to be compatible with the assignment of Handle PIDs via EPIC service[9], and added a simple CMDI metadata schema. When an option was added to harvest the metadata directly in the CMDI format, the repository was compatible with the CLARIN guidelines at the time.

The repository was further modified and upgraded in the following years, and it is run continuously at the LINDAT/CLARIN centre at Charles University (at **https://lindat.cz/repository/**). Its popularity is steadily growing, and it became a repository of choice for many international projects involving language datasets, like Universal Dependencies[10], or various NLP shared tasks (contests) like WMT of CoNLL. Currently, we preserve 356 datasets, 1.6TB in total[11]. There have been 310,000 downloads[12] since the start of 2019 (up to 27 September). At the moment, the repository has 724 user accounts, which are only used to either submit new datasets or sign licenses for restricted datasets.

In the following sections, we first describe further development of the CLARIN-DSpace software with motivation for the various modifications and extensions of DSpace we have implemented. Then we briefly mention CLARIN-DSpace install base and the move from the original LINDAT-DSpace as a project-internal[13] software development to CLARIN-DSpace, an open-source project involving several countries.

Evolving DSpace

The requirements for changes and improvements were coming from multiple directions. After the initial modification for using the EPIC handle system, we kept developing the system to best suit the needs of both users and administrators. Some changes were made to fulfil further CLARIN requirements for (what eventually became) B-Centres. Some were made to make administration more efficient, and yet another set of features was required by our users. Some were even our experiments because they seemed to offer interesting added value. In addition, we found and shared fixes for several bugs in the system, improved the user interface, enhanced the federated authentication system.

New Administrative Features

There are two features we have successfully merged into the main DSpace: our modified control panel (see Figure 1), and our health check system. The reason behind those improvements is that the system produces a lot of log messages that were not easy to manage; the whole repository infrastructure is not only the DSpace repository software, but also a database server, a web server, the single sign on federation service provider (Shibboleth service provider), and a handle server (standalone PID system). On the operating system level (or on the virtualization level) there are backups and periodical administrative tasks (cron). To get a good overview of the whole system setup, and to make this information readily available to repository administrators, we have substantially extended DSpace's control

[9] European Persistent Identifier Consortium, see **https://www.pidconsortium.eu/?page_id=112**

[10] Available from: **https://universaldependencies.org**

[11] We preserve also 710 metadata only records, mostly inherited from the early list of CLARIN resources called Linguistic Resources and Tools Inventory.

[12] These roughly match dataset downloads, but some datasets have multiple downloadable files.

[13] Even though, the internal project always was open source with a publicly available repository, wiki and issue tracker.

panel[14]. Originally, it showed just basic information like the uptime and some configuration; with our extensions, it also shows and searches log files, enables the admins to run some of the occasionally required reindex tasks, and allows us to inspect and edit metadata in bulk.

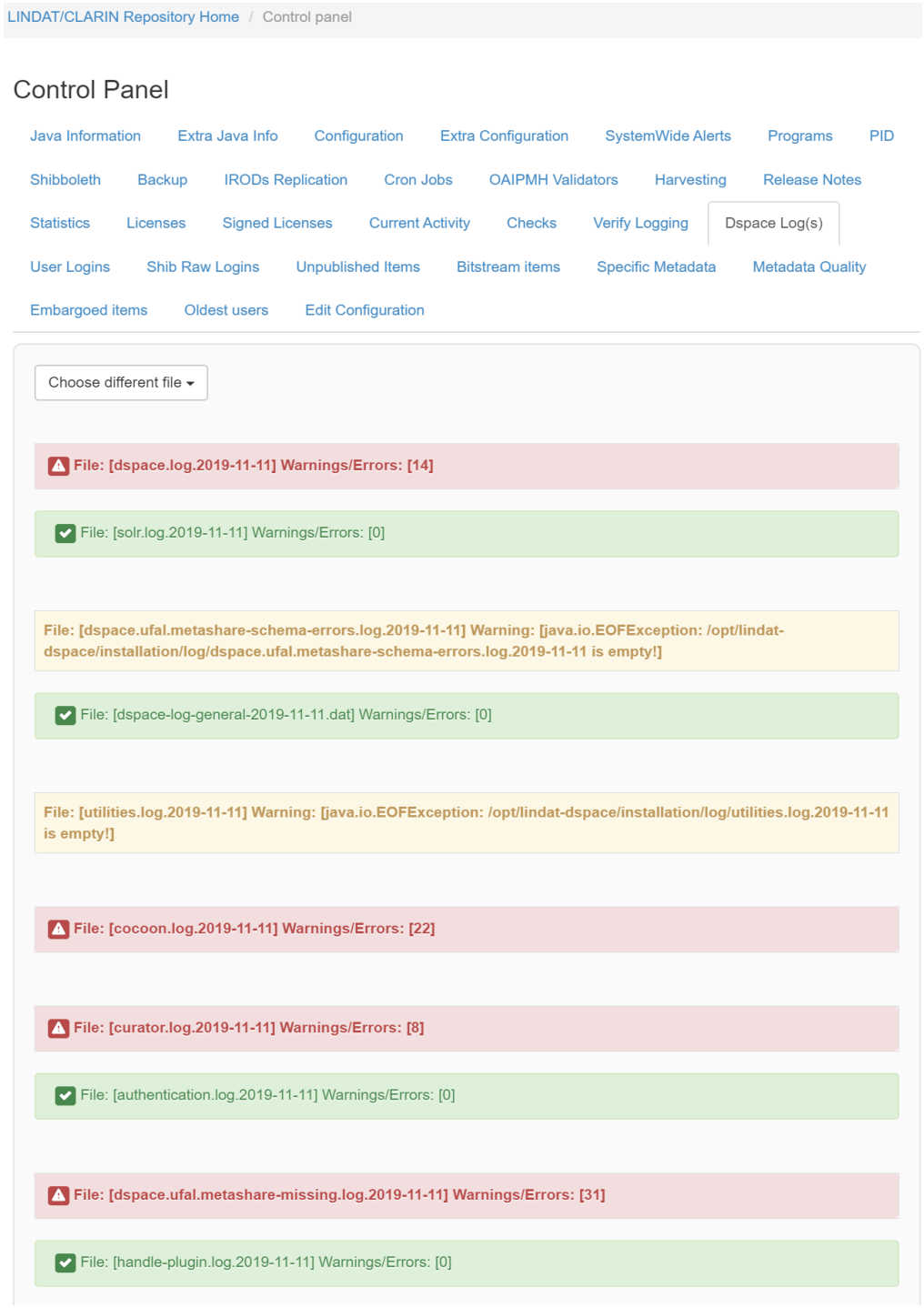

Figure 1: An illustration of control panel with logs tab selected. This provides a brief overview of various log files of the system and allows to inspect them without using the command line.

[14] An element of the web user interface, which is only visible to repository administrators, not regular users.

The health check exists for similar reasons: to generate periodical reports (we use a weekly schedule) describing the state of the system. Among other things, it shows the number of items, some distribution of items into collections based on type, it shows errors (if any) from the log files, and it also runs curation tasks. Curation tasks are usually submission level checks. One task checks that the links (URLs) in metadata work, and reports those that do not. Another check is a consistency check, which verifies the submitted data have not been modified. Some of the checks come with DSpace, some are our extensions. For example, we have a specific check for items that were funded by EU grants, to verify they contain a correct id and metadata for OpenAIRE[15] export.

One of the CLARIN requirements has always been for persistent identifiers to be handles[16]. DSpace comes with a handle server, so the only thing needed was to contact the Handle system administrators asking for your handle prefix, pay a small fee, and set up the handle server with the new prefix. However, our initial setup was using PID (handle) assignment from an external web service run at EPIC consortium[17], which required the first modification we made to DSpace. Our setup eventually became much more complex than that, however. Today CLARIN-DSpace has options to configure different handle prefixes for different communities[18], and we still provide a connector to the EPIC API. This means that some of the handles are hosted locally while others are minted by EPIC. We are using exactly this approach for a community called "LRT Inventory". It serves as a repository for countries, research groups or individuals who don't have their repositories, to be able to readily preserve and share language data. This community is connected to CLARIN ERIC, so we are using a handle prefix from EPIC owned by CLARIN ERIC. This gives CLARIN a fundamental level of control over the records. The other community in the repository serves for data and tools of the LINDAT/CLARIN project and has its prefix owned by the project itself. To be able to manage the handles efficiently, a new user interface was implemented into the CLARIN-DSpace.

Licensing

An item (a record) in the repository consists, in general, of 2 parts: data and metadata. For metadata our licensing policy is simple: we keep the perspective that metadata is not a "free creative work" within the scope of copyright, thus it doesn't require any license. In fact, it cannot even be licensed, it simply is in the public domain.

Data, however, is very often (and language data almost always) creative work that is within the scope of copyright law. This means that any handling of such data requires an explicit license[19]. Thus a repository system for language data must have a strong licensing support in two ways: the submitter must choose a license for end-users, which specifies how they can use the data, but they also must agree to a "deposition license" from the repository. This is an agreement, in which the submitter gives the repository right to distribute the data to end users and states

[15] OpenAIRE's mission is closely linked to the mission of the European Commission: to provide unlimited, barrier free, open access to research outputs financed by public funding in Europe. More details about OpenAIRE available from: **https://www.openaire.eu/**

[16] That includes DOI, because DOIs are also implemented using the Handle system.

[17] Today PID Consortium

[18] A community is a name of a top-level collection in DSpace.

[19] Unless there is an explicit exception in the copyright law for such use.

explicitly that he/she has checked the legal situation of the data and has the right to distribute the data under the chosen license and to pass this right to the repository.

For choosing and attaching a license to an item DSpace includes a small module that allows users to specify a Creative Commons (CC) license. This is nice, but by far not enough even if all the datasets could be licensed under some sort of public licenses. Thus CLARIN-DSpace implemented a completely new licensing framework, which allows the repository managers to "define" a license in the system and attach it to records. The license definition, in addition to the license text, has several other attributes. The key attributes specify, whether the license needs to be signed for each dataset it is attached to, or not. Public licenses – which allow redistribution – do not require signatures by their very nature, but many other licenses do. The licensing framework allows all kinds of licenses to be used, thus providing support for datasets that cannot be distributed under public licenses. For such restrictive licenses, the system blocks download attempts and redirects users to authentication. After they successfully login via their academic home institution (SAML2 system), the license for the particular dataset can be signed and the data downloaded. The licensing framework logs the information that this user signed this particular license for this particular dataset.

While the support for custom licenses and their signing is unique to CLARIN-DSpace, the emphasis is on Open Science. To support users in choosing an optimal license for their data or software,[20] the LINDAT/CLARIN project teamed with an expert lawyer and created a separate piece of software: the Public License Selector[21]. This small tool presents questions and explanations, and guided by the user's answers suggests the most open license for the given dataset (see Figure 2). The selector was integrated directly in the submission workflow of CLARIN-DSpace, so that users who want help with the choice of the license, can use it directly during data submission.

[20] In the choice of license data and software are not equal. Usually, licenses are suitable for data or software, but not for both.
[21] The software/the code available from: **https://github.com/ufal/public-license-selector**

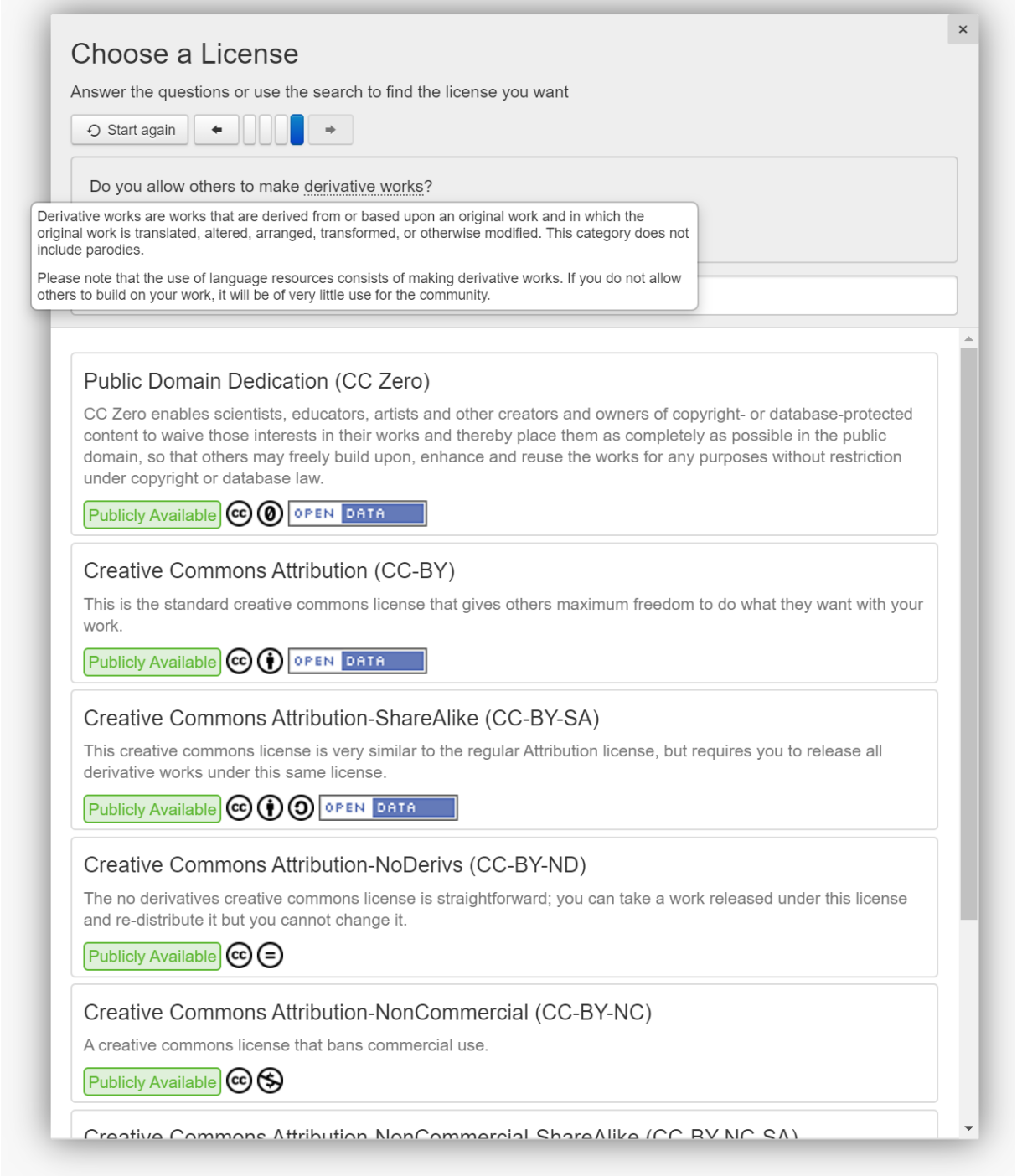

Figure 2: The public license selector asks a series of questions and based on the answers filters the suitable licenses. In this particular case we are at question number four "Do you allow others to make derivative works?" where the phrase "derivative works" is explained in detail as a mouse over hint.

Submission & Metadata

One of the reasons for choosing DSpace was its customizable submission workflow that allows us to easily define the metadata fields and choose, which of them are required, and which optional. Another aspect of metadata handling we could support with DSpace easily was the dissemination of the metadata in multiple formats and/or schemata. In the research domain of linguistics/language data, there are several schemata and frameworks related to metadata in use. There is the CMDI (required by CLARIN), which is not a schema, but rather a framework

that lets you create a schema suited for your particular items, and it also provides means of interoperability in this world of many schemata; there is the MetaShare project that prescribed a set of required minimal metadata; there is also OLAC,[22] and of course OpenAIRE for reporting all scientific results, including datasets. There is also the Clarivate Data Citation Index,[23] which CLARIN-DSpace fully supports, and as a result, it indexes all the data from LINDAT/CLARIN. We weren't required to support all of these, but we decided on the strategy of maximal dissemination as our distinguishing feature and a promise to the users: Your data will be visible. All of these services require their metadata schemata and often their metadata formats. However, implementing them was rather straightforward, because DSpace generates metadata for export (e.g. over OAI-PMH) by simple XSL transformations from the internal metadata. Thus adding one new format or simple schema for export was usually quite simple.

Some of the metadata formats, among other things, define a minimal set of required attributes. Our ability to disseminate into the multitude of formats also serves as a sort of verification that the schema we decided to implement (i.e. what we require at the submission time) is a good and sensible set. It fulfils the requirements of all the exports mentioned above.

A question of data citation and thus also export of item metadata in a bibliographic format can be also treated as a subset of the broader issue of metadata formats and dissemination. LINDAT/CLARIN decided to adopt the policy of direct data citations as it was pioneered by Force11[24] and implemented the “citation box” that is shown prominently on every item page. It contains a formatted text citation including the PID, conforming to the Force11 specification, and it also contains an option to export the citation in the BibTeX format. This BibTeX support was implemented via XSLT just like all the other metadata exports mentioned before. This means one can also get the BibTeX metadata over OAI/PMH from any CLARIN-DSpace repository.[25]

A positive side-effect of using DSpace is that it integrates well with Google Scholar.[26] LINDAT/DSpace made some significant changes and is optimised for datasets, not publications, but the development team made a conscious effort to keep this integration working. As a result, datasets of LINDAT/CLARIN are indexed by Google Scholar, just like any other scientific publications. When they are cited – which we promote as explained above – the authors of the data get the credit they deserve.

Versioning

One of our policies, coming from how we view persistent identifiers, is that a handle always resolves to one concrete item (its landing page), concrete dataset. When somebody cites data, we don't see a use case citing it in some abstract facility, not referring to a concrete dataset (i.e. its concrete version). Such vague use would break the principle of reproducibility in science.[27] We analysed how versioning was supported in various repository systems, including

[22] Open Language Archives Community, see **http://www.language-archives.org**

[23] Available from: **http://wokinfo.com/products_tools/multidisciplinary/dci/**

[24] For more details about Force 11 data citation principles see **https://www.force11.org/datacitationprinciples**. The policy of direct data citation is now also actively promoted by CLARIN.

[25] For example **http://lindat.mff.cuni.cz/repository/oai/request?verb=ListRecords&metadataPrefix=bibtex**

[26] Inclusion Guidelines for Webmasters available from: **https://scholar.google.com/intl/en/scholar/inclusion.html**

[27] Reproducibility represents the letter ‘R’ in the modern FAIR acronym (see the FAIR Data Principles **https://www.force11.org/group/fairgroup/fairprinciples**).

DSpace from its early attempts, and we decided to use a different approach. The implementation of versioning in CLARIN-DSpace is very simple. Each version of an item is a separate record and each has its handle. The only addition implemented is using the standard Dublin Core attributes ‘relation.replaces' and ‘relation.isreplacedby' to chain versions of the same item together. This information is visualised in the UI in two ways: a pop-up list of versions (see Figure 3) on each item that has the relations filled-in, and the fact that CLARIN-DSpace by default hides bitstreams of items that have a newer version, and instead shows an explanation that this dataset has newer versions (see Figure 4). Of course, the bitstreams can still be readily shown and downloaded, it is just a measure of pointing out to users who came to an older record, usually from a PID in a citation, that they can use the latest version if they want. The latest versions of items should also appear first in the search results. The submission process for new versions was also made very convenient by basically cloning the last version into the new one, and providing a user guide to do so.[28]

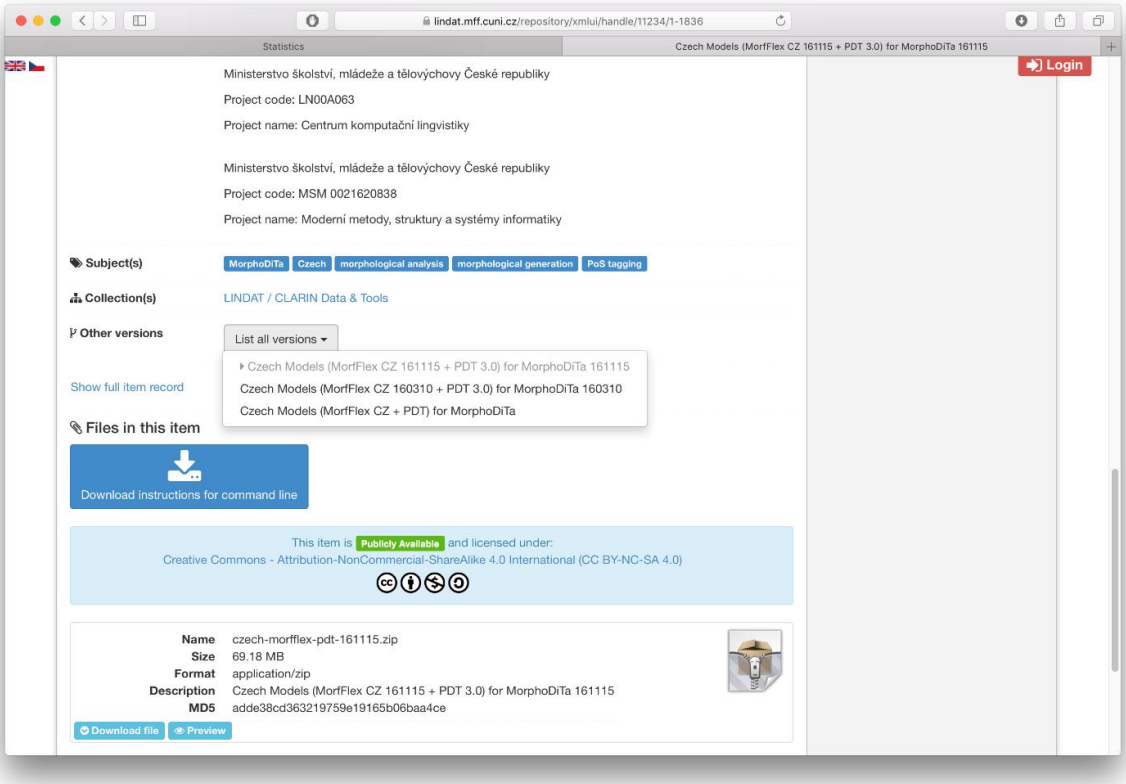

Figure 3: The latest version of a resource (if there are multiple versions) shows both the actual data files and links to all the previous versions.

[28] The user guide available from: https://github.com/ufal/clarin-dspace/wiki/New-Version-Guide

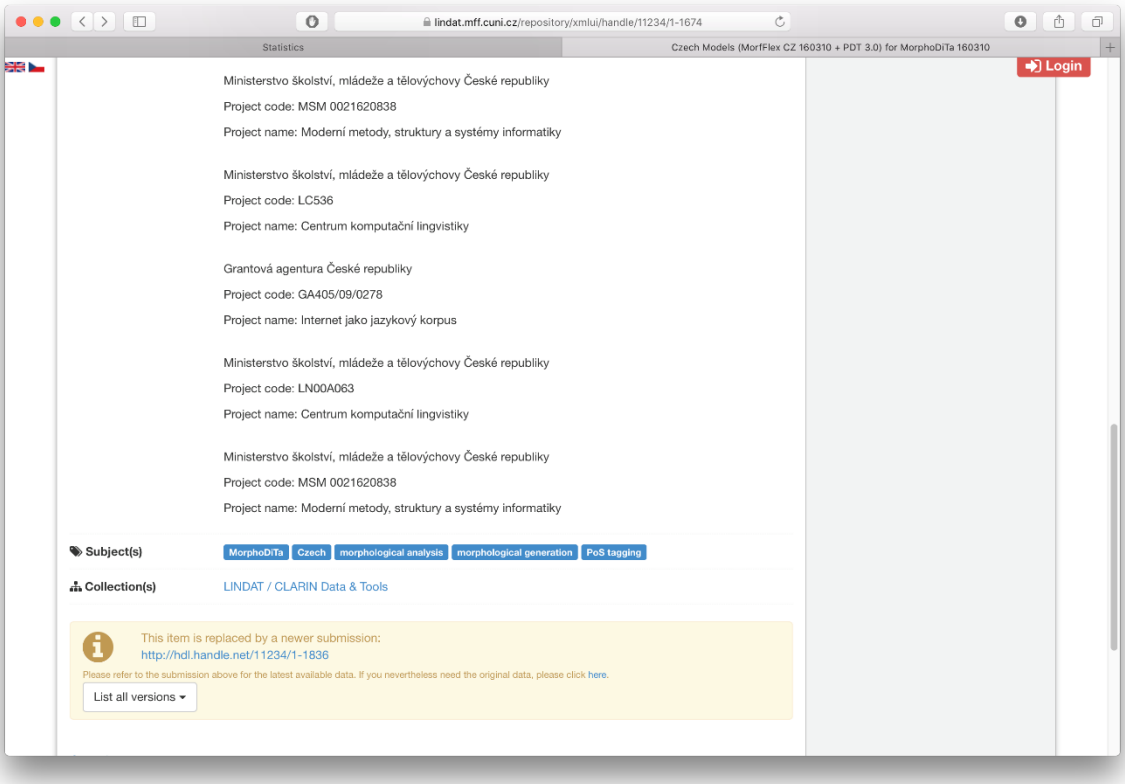

Figure 4: An illustration of what is shown to the users when they reach a resource that has a newer version in the system; ie. A link(s) to the different versions of the resource is shown instead of the files, but the original data can be downloaded when they follow the instructions.

Statistics

Any project running a repository has to prepare detailed reports to its stakeholders, including very detailed statistics of the actual usage of the repository. DSpace contains support for basic statistics but the support is not complex enough to be used as the basis for useful reports. Another option present in DSpace is to connect to Google Analytics (web analytics platform), but that has other implications, mainly sharing all the traffic data with Google. Eventually, the CLARIN-DSpace team chose to implement support of the Piwik (rebranded to Matomo) secure and open web analytics platform, which can be run in-house. At LINDAT/CLARIN we do just that. With this new feature, it is possible to provide meaningful and detailed statistics and do it without sharing information on visits of individual items with other parties. Submitters of data – or any other interested users – can also subscribe for monthly statistical reports of their items. These reports include the numbers of downloads and views, and graphs showing usage trends.

Working with Data

One crucial difference in how CLARIN-DSpace is used, at least the LINDAT/CLARIN instance, compared to many regular DSpace installations is the size of the files (bitstreams) being hosted. Our repository contains files with sizes in tens of GBs (at the time of writing, the largest single file is 70GB). Because a large portion of our users use fast academic or enterprise networks the file size itself is not viewed as a problem. What became a problem, however, is

the inefficient and naive implementation of the downloading process by DSpace stack[29]. It put a lot of stress on the CPU resources, and at the same time was not able to fully exploit the potential of very fast internet connection and storage. A workaround[30] was implemented that allows the webserver to handle the file downloads directly when the user is authorised by the repository systems (e.g., the requested item does not require any license signing). With this approach, CLARIN-DSpace added also a new feature – an essential one for a data repository – support for resuming of interrupted downloads.

On the other hand, we are taking a different approach when large files are being submitted to the repository. Uploads of less than 4GB are available directly through the submission workflow leveraging the http(s) protocol. Simple drag and drop of files onto the browser window. Larger files, however, need the cooperation of the repository staff. There are several reasons for that, one of them is that we want to check the users have thought about different ways of splitting the data and whether potential users are able to use big files efficiently. Another reason is to keep a level of control. In practice, this is not a problem, because language data are not commonly this large (when compressed), so in practice the load on repository administrators is minimal.

CLARIN-DSpace as an Open Source Project

DSpace is being developed mainly as a repository for publications, and the system is configured by default to serve this purpose. The requirements of a CLARIN data repository are different in various aspects. The core DSpace contains a number of features that a data repository might find useful, but they are disabled by default. The CLARIN-DSpace[31] project is not only a DSpace version with modified code, look, and some new features; it is also DSpace configured and optimised for data repositories. Furthermore, specifically for the CLARIN project, the CLARIN-DSpace meets the requirements of a CLARIN for B-Centres that run certified repositories for language resources.

What started at LINDAT/CLARIN as an attempt to fulfil CLARIN requirements with a simple and reliable solution, became in time appreciated by more CLARIN centres and several other institutions looking for data repositories with similar criteria. The consequence was that after several years there were about 10 installations. Consequently, the project was re-branded as CLARIN-DSpace and its install base is still steadily growing. The whole project has always been developed as an open source project under MIT license, and completely in the open: originally at the Redmine installed at the Institute of Formal and Applied Linguistics, and since 2015 at Github. The open approach to development and documentation seem to be key factors in the increasing adoption of the system.

CLARIN-DSpace system now meets all our requirements for a repository system for managing language data. While it can always be improved, all the critical parts are there for efficient use. Therefore, we focus on maintaining the repository and advocate for proper data preservation

[29] Where it doesn't leverage the features of a servlet container below. To copy the file to the Response, it loops through the InputStream with a blocking read() and a fixed buffer size.
[30] Essentially, the repository handles the authorisation and then tells the proxy what file to serve. The proxy uses more efficient means (system calls such as sendfile) to serve the content. More details available from: **https://github.com/ufal/clarin-dspace/wiki/Speeding-up-downloads**
[31] The code of Clarin-DSpace available from: **https://github.com/ufal/clarin-dspace**

and sharing, using the repository. We plan to keep the software stack in sync with DSpace development: as new versions of DSpace are developed and released, our team follows them and updates the CLARIN-DSpace to the new codebase.

THE SCOPE OF OPEN SCIENCE MONITORING AND GREY LITERATURE

Joachim Schöpfel

joachim.schopfel@univ-lille.fr

University of Lille, GERiiCO laboratory, France

Hélène Prost

helene.prost007@gmail.com

GERiiCO laboratory, CNRS, France

Abstract

One European research policy is to foster open science. The Open Science Monitor has been created as one particular source among many addressed before the European Commission makes proposals for policy in cooperation with Member States of the EU and stakeholders. The purpose of this paper is to assess the real and potential place of grey literature in the EC Open Science Monitor, in their data sources, in their methodology and indicators, in published surveys and case studies, etc. Additionally, as the creation of a monitoring system is among the objectives of the new French National Plan for Open Science, the paper provides comparative information about the French approach to open science monitoring.

Keywords

Open science, monitoring, evaluation, grey literature

Introduction

One European research policy is to foster open science. The Open Science Monitor[1] has been created as one particular source among many addressed before the European Commission makes proposals for policy in cooperation with Member States of the EU and stakeholders. The objectives are to provide qualitative and quantitative data and insights into understanding the development of open science in Europe, and to gather the most relevant and timely indicators on the development of open science in Europe and other global partner countries.

The Commission launched the Open Science Monitor in 2018. The first results have generally been acclaimed and widely shared on social media. However, at first sight, the underlying methodology of the Monitor focuses on journal publishing and repositories without data on grey literature such as conference papers, theses and dissertations, reports, working papers and so on. Are these simply out of the scope or beneath the radar of the European Open Science Policy? The only "boundary object" of the EC Monitor is the preprint, but only as preliminary versions of published journal articles.

The purpose of this paper is to assess the real and potential place of grey literature in the EC Open Science Monitor, in their data sources, in their methodology and indicators, in published surveys and case studies, etc. Additionally, as the creation of a monitoring system is among the objectives of the new French National Plan for Open Science, the paper provides comparative information about the French approach to open science monitoring. The paper does not assume that "more open science = better science" but assumes that open science monitoring will have a significant impact on the future development of the open science ecosystem as a main paradigm of scientific research in Europe.

Validity

Wiktionary defines "monitoring" as the "*carrying out of surveillance on, or continuous or regular observation of, an environment or people in order to detect signals, movements or changes of state or quality*"[2]. It means to check or to keep track of objects or people, usually for a specific purpose. One quality of any assessment is the reliability and consistency of the measurement. Yet the first requirement of quality is validity, i.e. a shared understanding or social agreement about what should be assessed.

[1] Open Science Monitor **https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor_en**
[2] Wiktionary **https://en.wiktionary.org/wiki/monitoring**

This is the first issue of open science monitoring, as "open science means different things to different people" (Daii et al. 2018). There are various approaches to defining open science, some broader and more inclusive, others more selective, and more or less useful (Bosman & Kramer 2017). Some definitions are presented as taxonomies of open practices or principles (e.g. open data, open access and open peer review), others are rather simple yet not very helpful because they raise new questions about concepts, meanings, limits etc. ("*right to use, reuse, modify, and redistribute scholarly knowledge*" and so on). In its broadest sense, "*open science (…) refers to efforts to make the scientific process more open and inclusive for all relevant actors, within and beyond the scientific community, as enabled by digitalisation*" (Daii et al. 2018).

The European Commission defined open science a couple of years ago as "*the transformation, opening up and democratisation of science, research and innovation, with the objective of making science more efficient, transparent and interdisciplinary, of changing the interaction between science and society, and of enabling broader societal impact and innovation*" (Ramjoué 2015). This EC approach is broad and inclusive rather than operational; today, significant emphasis is placed on research data, especially through the European Open Science Cloud (EOSC) and the preparation of the FP9 funding programme (Burgelman & Tsoukala 2018).

Probably the best explanation of the broadness of the EC approach is the understanding of open science as a policy process: "*Open science strengthens the link between science and society, for example increasing the transparency of evidence-based policymaking. It also enables society to handle the 'data deluge' more effectively as service providers may step in to curate and evaluate data for interested users*" (Ramjoué 2015). This "social agreement" aspect of social science implies the application of the usual policy instruments such as monitoring.

Policy Monitoring

Monitoring is an integral component of the policy process insofar as it "*describes the development and implementation of policies, identifies potential gaps in the process, outlines areas for improvement, and makes key implementing institutions accountable for their activities*" (Waterman & Brown 1993). Policy monitoring generally follows a two-step process:

- Identification of indicators measuring key activities related to the policy;
- Collection, analysis and dissemination of data on those indicators.

Policy monitoring can also include the identification of key operational policy barriers that should be addressed through policy or program reform. The purpose of policy monitoring is to allow "*policymakers and interested actors to systematically examine the process of creating a policy, implementing it, and evaluating its effects*" (Waterman & Brown 1993). Key indicators of policy impact are, for instance, service utilization, the adaptation of behaviour by the intended population, and even changes to policies. Typical examples of policy monitoring objects are international missions, health systems, and surveillance of climate change and biodiversity and all kinds of discrimination.

Regarding the European open science process, in 2016 the "Amsterdam Call for Action" recommended to "*implement monitoring and stocktaking at regular intervals about the progress made by all parties: the Commission, the Member States and stakeholders*". The idea was to agree on a "*100% target for 2020*", and for regular monitoring and evaluation based on standards, systems and services for monitoring and reporting to be established, and to "*regularly refine plans to achieve these targets based on information from monitorin*g", and to mainstream and further promote open science policies (Netherlands' EU Presidency 2016). Following the Amsterdam paper, national authorities and the European Commission are in charge of policy development and implementation and policy monitoring, while research funders and research-performing organisations should share expertise and provide data for the purpose of monitoring.

Of particular importance in the era of open science and big data, an OpenAIRE workshop in May 2019 investigated three main issues in research policy monitoring, i.e. the quality and validity of the assessment, the transparency of the process, and the collaboration among the different stakeholders[3].

The European "Open Science Monitor"

In line with the Amsterdam Action Plan, the Open Science Monitor was launched in 2018 as a pilot project initially developed by RAND Corporation[4] for the European Commission, and expanded and updated by a consortium led by the Lisbon Council think tank, with Elsevier as sole sub-contractor[5]. The tool is part of the EC website, managed by the EC Directorate-General for Communication. According to the information on the website, the Open Science Monitor aims to

- *"provide data and insight to understand the development of open science in Europe*
- *gather the most relevant and timely indicators on the development of open science in Europe and other global partner countries."*

However, the EC site stresses the pilot aspect of the project, which was developed to "*test the viability and value of assessing open science activity in Europe and beyond*" without being an assessment tool: "*The Commission may draw conclusions from the quantitative and qualitative trends in open science and its drivers to propose new policies for fostering open science. However, the Commission will not base its policies fully on it. The Open Science Monitor will be only one particular source among many (…)*". Criticisms were expressed in particular due to the use of proprietary data for the monitoring (i.e. Elsevier's Scopus).

Considering open science as an "*approach to research that is collaborative, transparent and accessible",* the consortium selected "*areas that have consistent and reliable data, specifically: open access, open research data, open scholarly communication and citizen science*" (Smith et al. 2017b). These three fields – open access, open research data and open collaboration – are covered by indicators, while other areas such as open educational resources, open peer review or open methodology are (at least for the moment) marginal or excluded. The consortium admits the limits and beta-character of the website: "*As long as we do not have*

[3] **https://www.openaire.eu/research-policy-monitoring-in-the-era-of-open-science-and-big-data-2**
[4] **https://www.rand.org/randeurope/research/projects/open-science-monitor.html**
[5] **https://www.scienceguide.nl/2018/07/elsevier-is-trying-to-co-opt-the-open-science-movement-and-we-shouldnt-let-them/**

yet an open data infrastructure(s) available, we are dependent on actors giving access to data sources, which are useful for the tracking and monitoring of open science practices. The current Monitor will therefore not be perfect, but the intention is to be as inclusive as possible in terms of drawing on data-sources and suggestions from experts".

"Open access" means open access to scientific publications, a concept meaning journal articles. In fact, the methodological notes and other preparatory documents make use of the terms 'article', 'paper' and 'publication' in a confusingly random way, as if they were all synonyms (Smith et al. 2017a; van Leeuwen et al. 2017). Obviously, the consortium does not consider any distinction between publication, paper and article as relevant for their purpose of monitoring open access in Europe. Also, the different sections of the Monitor make almost exclusive use of indicators of journal publishing:

- Trends in open access to publications: this section provides data and case studies covering access to scientific publications, with bibliometric data and data on the policies of journals and funders. It especially contains data on gold and green access to journal articles per country and field of science and technology, retrieved from Scopus and Unpaywall and double checked using different sources, such as the Web of Science (Osimo 2019).
- Funders' policies: the second section provides information about types of mandates established by research funders concerning open access publication and archiving, retrieved from the Sherpa Juliet database with a focus on journals.
- Research journal policies: the last section provides information about types of mandates established by research funders' journals concerning open access archiving policies, produced by the Sherpa Romeo database, also with a focus on journals.

Additional indicators are provided on the number of preprints, on articles published before peer review, and on surveys on the attitudes of researchers towards open access by 101 Innovations, Taylor and Francis, and Nature Publishing Group, along with 30 case studies on the drivers and barriers encountered regarding open science and the direct impact on three main areas, i.e. science, industry and society, including a comparison between the Web of Science and Scopus (Osimo 2019).

All these indicators and studies, except for the preprints, focus on journal publishing without any data on other peripheral or unconventional types of scientific documents such as theses and dissertations, reports, working papers or conference proceedings. Some of the Monitor's sources contain at least some of these documents but the Monitor does not make any distinction, as if they were all just the same kind of resource, with the same value. One probable reason for this is that the key indicators heavily rely on journal platforms, discovery tools and the DOI.

In a personal email to the first author, dated 24 October 2019, Thed van Leeuwen explains why they focus mostly on academic publishing in journals. "*The issue with (grey literature) is that it comes in a variety of appearances. We know of course internationally oriented grey literature, such as reports and policy brief from the EC. On the other hand, we have a wide variety of documents originating from various countries, in various formats and in various languages. Certainly, we see a parallel with academic publishing, as for example also in the social sciences and in particular the humanities, publishing occurs in local language journals, oriented towards a more local audience. However, the part of the academic publishing that is*

more similar, which means, internationally oriented and mostly written in English, has created more standardization. And although I clearly see the issues connected to standardization, in this case that allows for large scale cross-country analysis of academic publishing, in both closed and OA format. These studies are (…) based upon databases such as Scopus or Web of Science. And unfortunately, such systems do not exist in the realm of a number of other types of analyses we would like to conduct, for example around the study of societal relevance. This seriously limits our possibilities to study the wider circle of activities by academic communities."

While he can see the interest of grey literature for the study of societal relevance, "*as it would indicate interactions between academics and non-scholarly audiences, for example policy environments*", Thed van Leeuwen confirms that they "*do not include this type of outputs as we feel we cannot study them in a generic and consistent manner, but mostly due to the lack of sources that allow us to draw conclusions that do right to the type of outputs we study, preferably in an internationally comparative manner*."

During the preparation of the new version of the Open Science Monitor, the choice of the Scopus database was heavily criticized, especially because of the bias against arts, humanities and social sciences, non-English publications and other journal documents such as books, preprints, reports, conference papers and posters, etc.[6] As one commentary puts it, "*all these indicators should be named not 'publications' but 'journal articles' percentage*s".

The Open Science Monitor contains some other indicators relevant for publishing such as open collaboration (citizen science), open peer review, altmetrics and corrections or retractions. None of them appears to make use of other-than-journal data, and none of them even mentions any kind of grey literature.

In brief, the current version of the European Open Science Monitor does not assess the development and implementation of open science policies regarding grey literature, except for preprints which are directly related to journal publishing. Grey literature seems out of the scope of the European open science policy monitoring. It has become invisible.

The French Open Science Monitor

According to Burgelman & Tsoukala (2018), half the European Member States monitor the development and/or growth of open access, in particular with indicators on publications at national level. One example is the Dutch Open Science Monitor on the national open science platform[7], while other examples are the Danish Open Access Indicator produced by the Danish Agency for Science and Higher Education[8] and the German Open Access Monitor from the Forschungszentrum Jülich[9]. A fourth example is the French Open Science Monitor[10] developed by the French Ministry of Higher Education, Research and Innovation and presented recently at the ELPUP 2019 conference in Marseille (Jeangirard 2019).

[6] **https://makingspeechestalk.com/ch/Open_Science_Monitor/**
[7] **https://www.openaccess.nl/en/in-the-netherlands/monitor**
[8] **https://www.oaindikator.dk/en**
[9] **https://open-access-monitor.de**
[10] Baromètre français de la science ouverte **https://ministeresuprecherche.github.io/bso/**

While the Dutch and the German Monitors are limited to journal articles in gold and hybrid journals and open repositories, the Danish monitor includes one type of grey literature, published conference proceedings. Its scope is peer-reviewed, scientific publications - articles and conference contributions - registered in the research databases and institutional repositories of the participating institutions or in other recognized open repositories, and published in proceedings or journals with an ISSN. In the 2017 dataset of 20,645 items, conference contributions represent 5% of the total number of monitored publications, and they have been published via proceeding series from SPIE, IOP, Springer (Lecture Notes) etc., which are not really part of grey literature.

The French Open Science Monitor proceeds in a different way, one that is described as transparent, open and bottom-up, and based on the requirements of the French National Plan for Open Science launched in 2018 (Jeangirard & Weisenburger 2019). Instead of using affiliation data from the Web of Science or Scopus, the French Monitor applies a three-step method:

1) Identification of publications with a French author
    a) Constitution of a representative, exhaustive publication database;
    b) Identification of French researchers (authors);
    c) Identification and selection of publications (co-)authored by French researchers;
2) Enrichment of the selected publications' metadata (institutions, research domains);
3) Establishment of accessibility (open access).

Data from different sources undergo quality control and correction procedures. Most of the processing is automated, with manual checks if required. "The affiliations metadata are key for building an OA monitoring at a national level" (Jeangirard 2019). Based on a manual check of a random sample, the precision of the identification of French publications was estimated at 96%, i.e. 4% false positive errors (identified wrongly as French). The detection of accessibility with HAL and Unpaywall produces between 3% (for older publications) and 11% (for recent publications) false negative errors (identified wrongly as closed).

The current version of the French Monitor collects data from different, openly (publicly) available sources, in particular Unpaywall (>100m items with 24m OA items) and the French national HAL repository (>1.5m items). Other sources are used to identify French researchers (like the French SUDOC academic union catalogue with its IDRef author identifier, the French PhD portal thèses.fr, and the ORCID database) and to enrich the metadata (like the French PASCAL and FRANCIS databases and the RNSR national directory of scientific structures). The main challenges are the data volume and variety (DOI, affiliations…) and the dynamics of open accessibility, i.e. the evolution of the publications' status (from closed to open). The results will be made publicly available on the French open data platform (data.gouv.fr).

The inclusion of grey literature is conditioned by the DOI. Insofar as the assessment of accessibility via Unpaywall requires a DOI, all documents with DOI are considered and nothing else. The attribution of DOIs is not limited to commercial journal publishing; the metadata of the Crossref database show evidence of conference papers, reports and other "posted content". Yet the Crossref 2017/2018 annual report[11] shows that this part is rather small; out of the more than 101m Crossref records, 0.08% are preprints, 0.2% are dissertations, 0.6%

[11] **https://www.crossref.org/annual-report/**

are reports and 5.5% are conference papers, most of the latter being published in commercial conference proceedings series and not as grey literature (see above).

Therefore, "*as a consequence of our first choice to reduce the perimeter to publications with a DOI (…) the majority of publications that we analyze are journal-articles (86.7%)".* The French dataset contains 7,004 "*proceedings-articles*" (sic), representing 5.3% of the total number of 132,970 identified publications by French (co-)authors in 2017 – "*probably an underestimation of the reality*" (Jeangirard 2019). The other categories are either marginal ("others") or not grey (books and chapters).

The French Monitor does not currently collect (meta)data from publications from other sources, but this remains an option. According to the project team, the future version may include French dissertations from 1990 on (source: theses.fr portal), books (source: SUDOC) and perhaps some HAL collections, such as the LARA collection with more than 30,000 scientific and technical reports[12]. Including the theses.fr and LARA data would potentially add about 13,000 PhD dissertations and 1,500 reports per year to the Monitor data, increasing the grey part of open science monitoring from its current 5% to 15%.

Discussion and Conclusion

Open science monitoring is part of policy monitoring, and seeks to describe the development and implementation of open access and open science. The analysis of the European Open Science Monitor and similar tools in the Netherlands, in Denmark and especially in France shows that they all produce indicators on open access to scientific publications mainly or exclusively in the field of journal publishing, neglecting and marginalizing other types of scientific publishing, grey literature above all. The only grey resources considered so far are preprints and conference papers but both are, as shown above, directly related to journal publishing.

Yet as the French initiative seems to confirm, there is real potential for including grey literature, especially for theses and dissertations and scientific and technical reports, but probably also for other categories such as conference papers, posters and presentations, and working papers. How can we explain the current situation?

Common Issues of Grey Literature

The first reason for the lack of grey literature in the tools of open science monitoring is probably that grey literature is… grey, with a large diversity of formats and languages, poor standardization and recording, and few generally accepted and harmonized identifiers. In particular, many grey resources still lack a DOI, which is a barrier not only for the application of altmetrics (Schöpfel & Prost 2017) but also today for monitoring based on data resources like Crossref and Unpaywall.

[12] **https://hal-lara.archives-ouvertes.fr/**

Also, particularly in institutional and other open repositories, grey resources are not always easy to identify because of lacking, misleading or ambiguous metadata describing document types. Thed van Leeuwen points out that large international reservoirs of grey literature are missing.

A third "common issue" of grey literature is the supposed lack of quality control. All monitoring tools lay emphasis on "certified content", i.e. on peer-reviewed journal articles, which is another, well-known and often addressed handicap of grey literature in this environment.

Tools in Transition

Political pressure and the speed of change may be another reason. In fact, technical and organisational feasibility is another quality criterion of measurement – the best indicator is without interest or value if it cannot be achieved with reasonable resources and delays. Prioritizing the most important and the (relatively) easy-to-produce indicators is a realistic approach when a project team or consortium has to meet tight deadlines.

The methodology of the Open Science Monitor is not definitive, and will be updated on a regular basis in the course of the project until the end of 2019. The EC has announced that new indicators and data will be uploaded over the next few months. They also described the Monitor as a "collaborative effort" and invited the community to contribute.

The national Monitors have adopted similar strategies, describing the current version as experimental, a test, a draft or transitional tool that will continue to develop. Thus the feedback loop of policy monitoring impacts and shapes not only policy development and implementation, but also the monitoring devices themselves, depending on policy changes, outcome evaluation of the existing tools, new technologies, new data sources and new requirements from the community.

The French Monitor clearly expects to exploit more data sources, including grey literature reservoirs, increase the variety and diversity of data on open access and the representativity and exhaustivity of the publication database. Following the comments from the European consortium, a similar agenda for the European Open Science Monitor is unlikely, for the reasons mentioned above. As for the community criticism and recommendations during the initial project phase, the updated methodological note published 4 April 2019 provides some answers to these comments but limits further exploitation of open access data to the Web of Science and Unpaywall, arguing *inter alia* that only few received proposals were "*immediately actionable*", and that "*most proposals need additional effort, and that some are not deemed relevant*"[13]. OpenAIRE[14] could be such a new data resource, as it contains more than 7.5m grey items, preprints, conference objects, reports, theses and dissertations out of over 30m total publications. For the moment, however, the Open Science Monitor methodological note does not mention this option.

Due to the diversity of languages and data reservoirs (databases, catalogues, repositories…), it may be easier to include grey literature in national than European monitoring.

[13] **https://ec.europa.eu/info/sites/info/files/research_and_innovation/open_science_monitor_methodological_note_april_2019.pdf**

[14] **https://www.openaire.eu/**

The "Seepage" of Grey Literature

As mentioned above, the quality of monitoring depends on the reliability (consistency) and the validity of the assessment. Considering open access, the validity criterion requires a kind of shared understanding – a social agreement of what accessibility of scientific publications means. Such a shared understanding on open access does not exist, either on gold or green roads, on business models (APCs or platinum), or on licensing (libre or gratis?) or reuse conditions… While many publishers, funding bodies, research-performing organizations and authorities focus on journal publishing as the mainstream dissemination of research results, other initiatives and communities argue for a larger variety of knowledge production and business models ("bibliodiversity")[15].

However, in the public debate, as in scientific literature about academic publishing, the impression prevails that non-conventional publications (= not published as journal articles) do not exist or at best are not relevant for evaluation and monitoring. For instance, a recent monograph on scholarly communication published by one of the most important academic publishers simply omits speaking of other kinds of scientific literature except journals (De Silva & Vance 2017). Does grey literature become invisible or, to use a geological term, are we witnessing a kind of "seepage" of grey literature in the mainstream of academic publishing? What is invisible often does not exist, at least in political strategies. Grey items are still somewhere outside but who really cares?

There is a famous quote, often (mis)attributed to Albert Einstein: "Not everything that counts can be counted, and not everything that can be counted counts". Nevertheless, without any reliable data monitoring of grey literature, how will the European Commission (and the French Government) conduct an inclusive and comprehensive open science strategy and foster the production, discovery and curation of the grey part of scientific production in the new research ecosystem?

Acknowledgments

We would like to express our gratitude to Eric Jeangirard and Emmanuel Weisenburger from the French Ministry of Higher Education, Research and Innovation for their helpful advice, and to Thed Van Leeuwen from the Centre for Science & Technology Studies, Leiden University, for his reply concerning the European Open Science Monitor.

References

BOSMAN, Jeroen and Bianca KRAMER, 2017. *Defining Open Science Definitions* [online]. I&M / I&O 2.0. [Accessed 15 September 2019]. Available from: **https://im2punt0.wordpress.com/2017/03/27/defining-open-science-definitions/**

[15] See the French *Jussieu Call for Open Science and Bibliodiversity* **https://jussieucall.org/jussieu-call/**

BURGELMAN, Jean-Claude and Victoria TSOUKALA, 2018. Open Science in Europe: The Perspective from the European Commission. *OPERAS Conference: Open Scholarly Communication in Europe. Addressing the Coordination Challenge* [online], Athens, 31 May 2018. [Accessed 15 September 2019]. Available from: **http://helios-eie.ekt.gr/EIE/bitstream/10442/15720/2/BURGELMAN.pdf**

DAII, Qian, Eunjung SHIN, and Carthage SMITH, 2018. Open and Inclusive Collaboration in Science: A Framework. *OECD Science, Technology and Industry Working Papers* [online], No. 2018/07. [Accessed 15 September 2019]. ISSN 1815-1965. Available from: **https://doi.org/10.1787/2dbff737-en**

DE SILVA, Pali U.K. and Candace K. VANCE, 2017. *Scholarly Communication: The Changing Landscape*. Springer: Berlin. ISBN 978-3-319-50626-5.

JEANGIRARD, Eric, 2019. Monitoring Open Access at a National Level: French Case Study. *ELPUB 2019 International Conference on Electronic Publishing* [online], June 2019, Marseille, France. [Accessed 15 September 2019]. Available from: **https://doi.org/10.4000/proceedings.elpub.2019.20**

JEANGIRARD, Eric and Emmanuel WEISENBURGER, 2019. Utilisations de bases et référentiels ouverts pour aider au pilotage de politiques publiques: Exemples de ScanR et du Baromètre de la Science Ouverte. *Journées de l'ABES* [online], Montpellier, 29 May 2019 [Accessed 15 September 2019]. Available from: **https://www.slideshare.net/abesweb/jabes-2019-session-plnire-baromtre-de-la-science-ouverte-et-scanr-moteur-de-la-recherche-et-de-linnovation-deux-outils-au-service-de-lactivit-scientifique**

NETHERLANDS' EU PRESIDENCY, 2016. Amsterdam Call for Action on Open Science. *Open Science, from Vision to Action* [online report]. Amsterdam, 4-5 April 2016 [Accessed 15 September 2019]. Available from: **https://www.government.nl/topics/science/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science**

OSIMO, David, 2019. Building Together an Open Science Monitor: Tracking Trends for Open Access, Collaborative and Transparent Research Across Countries and Disciplines. *Open Science Conference* [online], Berlin, 19-20 March 2019. [Accessed 15 September 2019]. Available from: **https://www.open-science-conference.eu/wp-content/uploads/2019/03/01_osc2019_presentation.pdf**

RAMJOUE, Celina, 2015. Towards Open Science: The Vision of the European Commission. *Information Services & Use* [online], **35**(3), 167–170. [Accessed 15 September 2019]. **https://doi.org/doi:10.3233/isu-150777**

SCHÖPFEL, Joachim and Hélène PROST, 2017. Altmetrics and Grey Literature: Perspectives and Challenges. *The Grey Journal* [online], **13**(1), 5-22. [Accessed 15 September 2019]. Available from: **https://hal.archives-ouvertes.fr/GERIICO/hal-01405443**

SMITH, Elta, et al., 2017a. *Monitoring Open Science Trends in Europe* [online]. RAND Corporation: Santa Monica, CA. [Accessed 15 September 2019]. Available from: **https://doi.org/doi:10.7249/TL252**

SMITH, Elta, et al., 2017b. *Open Science Monitoring: Methodological Note* [online]. RAND Corporation: Santa Monica, CA. [Accessed 15 September 2019]. Available from: **https://www.rand.org/pubs/external_publications/EP67274.html**

VAN LEEUWEN, Thed, Ingeborg MEIJER, Alfredo YEGROS-YEGROS, and Rodrigo COSTAS, 2017. Developing Indicators on Open Access by Combining Evidence from Diverse Data Sources. *Proceedings of the 2017 STI Conference* [online], 6-8 September, Paris, France. [Accessed 15 September 2019]. Available from: **https://arxiv.org/ftp/arxiv/papers/1802/1802.02827.pdf**

WATERMAN, Richard W. and B. Dan WOOD, 1993. Policy Monitoring and Policy Analysis. *Journal of Policy Analysis and Management* [online], **12**(4), 685–699. [Accessed 15 September 2019]. Available from: **https://www.jstor.org/stable/3325346**

EXCEPTION FOR TEXT AND DATA MINING FOR THE PURPOSES OF SCIENTIFIC RESEARCH IN THE CONTEXT OF LIBRARIES AND REPOSITORIES

Jakub Míšek

Jakub.Misek@law.muni.cz

Institute of Law and Technology, Masaryk University, Czech Republic

Abstract

The Copyright Directive in the Digital Single Market aims to update copyright legislation to meet the needs of the information society. This paper will deal with its Articles 3 and 4, which introduce a mandatory exception to copyright and the sui generis database right, allowing text and data mining. Article 3 specifically serves the purposes of scientific research and Article 4 covers a general, albeit limited, exception. The first part of the paper will introduce this new institute and the context of its adoption. The second part will critically evaluate it and analyse its strengths and weaknesses. The third part will then cover the activities of libraries and repositories, including the context of public sector information publication.

Keywords
Text and data mining, copyright, database protection, public sector information, grey literature

Introduction[1]

Text and data mining (hereinafter referred to as "TDM") collectively identifies techniques by which a large amount of information[2] can be gathered from a large number of documents by revealing the links between various sources.[3] In the context of grey literature,[4] i.e. documents from various originators available online and easily processed by automated means, this is an extremely useful technological procedure that can be widely applied.[5] As Truyens and Van Eeck point out, typical TDM activities include sorting text and data into one or more categories, grouping text and data together, extracting a common concept (for example, finding what the documents say), analysing document sentiment and modelling relationships between the entities contained in the text and data.[6] Compared to standard data analysis, TDM practices differ in that data are not used to test prepared models or hypotheses but rather to discover new, currently hidden, contexts, patterns and models.[7] Traditionally, the TDM process is carried out in three steps: accessing the analysed content, mining or copying the analysed content, and finally analysing the content itself.[8]

Apart from the fact that processed documents may be protected by special information rights such as protection of personal data, confidential information or trade secrets, intellectual property law constitutes a fundamental obstacle to TDM. Even if the data analysed or acquired during analysis are not protected in any way,[9] TDM may infringe the author's rights in the case of copyrighted works that are part of the analysed database or may infringe the database

[1] The author gratefully acknowledges the aid from the Masaryk University under grant No. MUNI/A/1006/2018.

[2] Information in this context means information in a semantic sense, that is data put into a particular context within which they create knowledge (cf. e.g. ADRIAANS, Pieter. Information. In: ZALTA, Edward N. (ed.). *The Stanford Encyclopaedia of Philosophy* [online]. Stanford: Metaphysics Research Lab, Stanford University, 2013 [Accessed 30 June 2019]. Available from: **https://plato.stanford.edu/archives/fall2013/entries/information/**; FLORIDI, Luciano. *Information: a very short introduction*. Oxford; New York: Oxford University Press, 2010, 130 p. Very short introductions, 225. ISBN 978-0-19-955137-8.

[3] See ROSATI, Eleonora. Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity. *Asia Pacific Law Review* [online]. Rochester, NY: Social Science Research Network (SSRN), 2019, p. 2 [Accessed 8 October 2019]. Available from: **https://papers.ssrn.com/abstract=3452376**.

[4] The paper works with the concept of grey literature in the sense of the so-called Prague Definition, chosen because of its timeliness and comprehensive nature. See SCHÖPFEL, Joachim. Towards a Prague Definition of Grey Literature. *Twelfth International Conference on Grey Literature: Transparency in Grey Literature: Grey Tech Approaches to High Tech Issues*. 2010, pp. 11-26.

[5] Cf. MYŠKA, Matěj. Text and Data Mining of Grey Literature for the Purpose of Scientific Research. *The Grey Journal*. 2017, Vol. 13, p. 32.

[6] TRUYENS, Maarten; VAN EECKE, Patrick. Legal aspects of text mining. *Computer Law & Security Review* [online]. 2014, Vol. 30, No. 2, p. 153 [Accessed 8 October 2019]. ISSN 0267-3649. Available from: **https://doi.org/10.1016/j.clsr.2014.01.009**

[7] See LUPAŞCU, Monica. Text and Data Mining Exception - Technology into Our Lives. International Conference: CKS - Challenges of the Knowledge Society. 2019, p. 906. ISSN 2068-7796.

[8] See ROSATI, Eleonora. Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity. *Asia Pacific Law Review* [online]. Rochester, NY: Social Science Research Network (SSRN), 2019, p. 8-15 [Accessed 8 October 2019]. Available from: **https://papers.ssrn.com/abstract=3452376**

[9] Cf. e.g. KOŠČÍK, M. Duševní vlastnictví k výzkumným datům. In: KOŠČÍK, Michal et al. Výzkumná data a výzkumné databáze. *Právní rámec zpracování a sdílení vědeckých poznatků*. Prague: Wolters Kluwer ČR, 2018, p. 34. ISBN 978-80-7552-952-7.

copyright.[10] It may also affect sui generis database right within the meaning of Article 7 et seq. of Directive 96/9/EC.[11] The main problem for TDM thus lies in the wide scope in which are interpreted the concepts of reproduction (in the context of copyrighted works),[12] and mining[13] (in the context of the sui generis database right of the database maker).[14]

In order for a modern society to fully benefit from the application of TDM techniques, it is essential to ensure the comprehensibility and clarity of the relevant legislation. It must enable this analytical procedure to be carried out without the risk of infringement of intellectual property rights and the consequent legal sanctions. Senseney and Koehl, for example, argue that in the context of the use of TDM for scientific research, researchers experience a 'chilling effect'[15] when exposed to legal uncertainty about the consequences of their actions.[16] An effective system of exceptions and limitations to copyright and the sui generis database right is thus essential for the effective application of TDM techniques.

This paper introduces the applicable exceptions and limitations to copyright and the sui generis database right of the database maker that may apply to TDM, in particular in the context of the new EU Directive 2019/790 on copyright and related rights in the Digital Single Market (hereinafter referred to as the "DSM Directive"), which newly introduced exceptions for the implementation of TDM. It also critically assesses the practical applicability of this exception and, in its third part, places the new exception in the context of the activities of libraries and repositories, in conjunction with the legislation on the re-use of public sector information.[17]

[10] A database is protected by copyright in accordance with Article 3 of Directive 96/9/EC if it constitutes the author's own intellectual creation by way of selection or arrangement of the content.

[11] See TRUYENS, Maarten; VAN EECKE, Patrick. Legal aspects of text mining. *Computer Law & Security Review* [online]. 2014, Vol. 30, No. 2, p. 163-164 [Accessed 8 October 2019]. ISSN 0267-3649. Available from: **https://doi.org/10.1016/j.clsr.2014.01.009**

[12] Within the meaning of Article 2 of Directive 2001/29/EC.

[13] Within the meaning of Article 7(1) of Directive 96/9/EC.

[14] Cf. DUCATO, Rossana; STROWEL, Alain. Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility". *IIC - International Review of Intellectual Property & Competition Law*. 2019, Vol. 50, No. 6, p. 658. ISSN 0018-9855.*;* MYŠKA, Matěj; HARAŠTA, Jakub. Omezení autorského práva a zvláštních práv pořizovatele databáze v případě datové analýzy. *Časopis pro právní vědu a praxi.* 2016, Vol. 23, No. 4, pp. 378. ISSN 1805-2789.

[15] This is an undesirable form of self-censorship, where one would rather not take certain steps or activities to make sure that any negative effect, such as a legal sanction, will be avoided. For more information on the chilling effect, see e.g. PENNEY, J.W. Chilling Effects: Online Surveillance and Wikipedia Use. *Berkeley Technology Law Journal* [online]. 2016, Vol. 31, No. 1, pp. 117–182 [Accessed 8 October 2019]. ISSN 1086-3818. Available from: **https://doi.org/10.15779/Z38SS13**

[16] See SENSENEY, Megan; DICKSON KOEHL, Eleanor. Text data mining beyond the open data paradigm: Perspectives at the intersection of intellectual property and ethics. *Proceedings of the Association for Information Science & Technology* [online]. 2018, Vol. 55, No. 1, p. 891 [Accessed 30 June 2019]. ISSN 2373-9231. Available from: **https://doi.org/10.1002/pra2.2018.14505501162**

[17] In particular Directive 2003/98/EC, as amended by Directive 2013/37/EC, which will soon be replaced by EU Directive 2019/1024.

Exceptions for TDM

The possible application of the exceptions and limitations of copyright resulting from Directive 2001/29/EC and of the exceptions and limitations of sui generis database right resulting from Directive 96/9/EC has already been addressed by a number of authors in the context of TDM implementation.[18] In the context of copyright, consideration has been given in particular to the limitation of copyright for temporary copies (based on Article 5(1)) and to the limitation of the right of reproduction for scientific purposes.[19] In the context of the sui generis database right, it was then possible to consider the rights of authorized users who could mine non-essential parts of the database,[20] and finally an exception allowing the mining of a publicly accessible database for scientific research purposes.[21] However, as the texts cited in the introduction to this section have already extensively demonstrated, neither of these exceptions and limitations is broad and flexible enough to act as a suitable tool for implementing TDM.

In the context of grey literature and Czech law, we can also rely on an exception for an official work (at least in part of the examined documents), both in the case of copyright protection[22] and, appropriately, in the context of the protection of the sui generis database right.[23] However, even these exceptions are not broad enough to handle a large amount of analysed data efficiently, as there are also many copyrighted works in the context of grey literature.

The factual usefulness of the implementation of TDM and the shortcomings mentioned[24] have inspired the European legislator to introduce, into the DSM Directive, exceptions to copyright and database rights to allow TDM. The TDM exceptions appear in the new directive as two variants, enshrined in Articles 3 and 4. Both exceptions are mandatory for Member States and must therefore be transposed into national law. Article 3 establishes the exception of text mining for scientific research. Article 4 then provides for a general exception allowing TDM, but also contains a huge number of limitations that make it difficult to apply.

[18] See e.g. BROOK, Michelle; MURRAY-RUST, Peter; OPPENHEIM, Charles. The social, political and legal aspects of text and data mining (TDM). *D-Lib Magazine* [online]. 2014, Vol. 20, No. 11–12 [Accessed 30 June 2019]. ISSN 1082-9873. Available from: **https://doi.org/10.1045/november14-brook**; CASPERS, Marco; GUIBAULT, Lucie. A right to 'read' for machines: Assessing a black-box analysis exception for data mining. *Proceedings of the Association for Information Science & Technology* [online]. 2016, Vol. 53, No. 1, pp. 1-5 [Accessed 30 June 2019]. ISSN 2373-9231. Available from: **https://doi.org/10.1002/pra2.2016.14505301017**; GEIGER, Christophe; FROSIO, Giancarlo; BULAYENKO, Oleksandr. Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?: Legal Analysis and Policy Recommendations. *IIC International Review of Intellectual Property and Competition Law* [online]. 2018, Vol. 49, No. 7, pp. 814-844 [Accessed 8 October 2019]. ISSN 2195-0237. Available from: **https://doi.org/10.1007/s40319-018-0722-2**; HANNAY, William M. Legally Speaking - Of Mindfields and Minefields: Legal Issues in Text and Data Mining. *Against the Grain* [online]. 2014, Vol. 26, No. 1, pp. 52-55 [Accessed 30 June 2019]. ISSN 10432094. Available from: **https://doi.org/10.7771/2380-176X.6663**; MYŠKA, Matěj. Text and Data Mining of Grey Literature for the Purpose of Scientific Research. *The Grey Journal*. 2017, Vol. 13, pp. 32-37. ISSN 1574-1796.; MYŠKA, Matěj; HARAŠTA, Jakub. Omezení autorského práva a zvláštních práv pořizovatele databáze v případě datové analýzy. *Časopis pro právní vědu a praxi*. 2016, Vol. 23, No. 4, pp. 375-384. ISSN 1805-2789.; TRIAILLE, Jean-Paul et al. *Study on the legal framework of text and data mining (TDM)*. Luxembourg: Publications Office, 2014, p. 41. ISBN 978-92-79-31976-1., et seq.

[19] See Article 5(3)a) of Directive 2001/29/EC.

[20] See Article 8 of Directive 96/9/EC.

[21] Ibid, Article 9b).

[22] See Section 3a) of Act No 121/2000

[23] Ibid, Section 94.

[24] These issues are analyzed in detail by e.g., TRIAILLE, Jean-Paul et al. *Study on the legal framework of text and data mining (TDM)*. Luxembourg: Publications Office, 2014. ISBN 978-92-79-31976-1.

Critical Evaluation of the New TDM Exception

Ducato and Strowel argue that the creation of a new comprehensive TDM exception was not appropriate and, in their view, the European legislature should have set out to formulate a new right to "electronic reading of documents" to cover TDM with sufficient flexibility.[25] However, if we continue to move within the current intellectual property rights paradigm, which includes a broad interpretation of the concepts of reproduction and mining, and accept that there is no right to automated text reading, the existence of an exception to copyright and the sui generis database right is necessary for the effective application of TDM procedures. Unfortunately, the inclusion of this exception in the DSM Directive was not without some critical moments.

The fundamental problem of Article 3 of the DSM Directive is the very narrow scope of its application resulting from the necessary simultaneous fulfilment of conditions relating to the entity carrying out the TDM (research organizations[26] and cultural heritage institutions), conditions relating to the purpose of the TDM (only scientific research purposes) and the necessity of legal access to the protected content.[27] Another problem lies in the third paragraph of Article 3, which states that rightsholders "shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted", while these measures "shall not go beyond what is necessary to achieve that objective". This is a very unclear provision whose interpretation is likely to be difficult. In practice, for example, a provider of information (datasets, copyrighted works) may limit the number of accesses to its database, justifying this by the necessity to ensure technical integrity. In fact, however, this will make it technically impossible to implement TDM and thus apply this exception. A third and fundamental problem is that the exception laid down in Article 3 of the DSM Directive concerns only the right to reproduce works or mine databases but not any other rights of the author or database maker.

In assessing the TDM exception enshrined in Article 4, we must conclude that despite the benefits resulting from the exception being ultimately mandatory for Member States,[28] and that the exception applies to any TDM implementation regardless of purpose or executor (assuming it has legitimate access to the protected rights), the applicability of this exception will be very problematic for two reasons. Firstly, the DSM Directive considerably limits the possible period of reproduction and extraction retention to the time required directly for the purpose of performing TDM. This greatly limits the possibility of future analyses or, for example, checks on the repeatability of the activity. This may be particularly problematic in the case of purely private research which does not benefit from the exception laid down in Article 3 of the DSM Directive. The second major problem then lies in Article 4(3), which allows rightsholders to exclude the application of this exception.[29]

[25] Ibid.

[26] Recital 11 of the DSM Directive states that companies in the private sector may also be exempted if they do research in cooperation with the public sector.

[27] See also DUCATO, Rossana; STROWEL, Alain. Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility". *IIC - International Review of Intellectual Property & Competition Law*. 2019, Vol. 50, No. 6, p. 665. ISSN 0018-9855.

[28] The Council and Parliament proposals included a version which made the exception optional. Cf. ibid.

[29] See also ROSATI, Eleonora. Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity. *Asia Pacific Law Review* [online]. Rochester, NY: Social Science Research Network (SSRN), 2019, p. 20-21 [Accessed 8 October 2019]. Available from: **https://papers.ssrn.com/abstract=3452376**

These shortcomings show that although the new TDM exception has the potential to help in certain narrow areas, it still suffers from a number of problems. In legal terms, these make the implementation of TDM only a little more certain compared to the previous legislation. This view is further supported by the fact that when performing TDM, interference with the original work is minimal. There is no further publication of the work or any parts of it. What is extracted from the work for further use is new information that would otherwise remain hidden and that is not copyrighted.[30] In view of this, I am in favour of the viewpoint of Ducato and Strowel that, instead of this new exception, the European legislator should have set out to formulate a new "electronic document reading" right that would cover TDM adequately and at the same time correspond more to the technical and legal reality.[31]

TDM, Libraries, Repositories and PSI

In the context of the activities of libraries and repositories, two options need to be considered in terms of how the new TDM exception can affect them. The first option is that the library or repository, as an institution, carries out its own TDM analysis of the documents it manages. Although many of these documents will be covered by the official work exception, repositories and libraries still function essentially as hosting ISPs for a number of documents that are copyrighted and for which the library does not need to have a direct licence to use the work further.[32] However, if the TDM analysis of managed documents is carried out for scientific research purposes, libraries and repositories, as cultural heritage institutions,[33] can rely on the TDM exception enshrined in Article 3 of the DSM Directive. Moreover, in this case there would be no problems related to the sui generis database right or copyright protection of the database, since in the vast majority of cases these rights are held by libraries and repositories.[34]

The second, more interesting, option is where an external entity wants to process resources that the library or repository manages using TDM techniques. In this case, as libraries and repositories are often set up by the public sector, we are dealing with the field of legislation concerning the provision and re-use of public sector information (PSI). This is expressed at European level by Directive 2003/98/EC as amended by Directive 2013/37/EC (hereinafter referred to as the "PSI Directive"), which will be replaced in 2021 by the recently adopted EU Directive 2019/1024 on open data and the re-use of public sector information (hereinafter referred to as the "Open Data Directive").[35] In general, the obligations under the PSI Directive

[30] Copyright protects only the specific expression of the work, not the ideas behind it.

[31] See DUCATO, Rossana; STROWEL, Alain. Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility". *IIC - International Review of Intellectual Property & Competition Law*. 2019, Vol. 50, No. 6, pp. 649-684. ISSN 0018-9855.

[32] For more information on a grey literature repository like an ISP, see MYŠKA, Matěj; ŠAVELKA, Jaromír. A Model Framework for publishing Grey Literature in Open Access. *Journal of Intellectual Property, Information Technology and E-Commerce Law*. 2013, Vol. 4, p. 109. ISSN 2190-3387.

[33] Cf. recital 13 of the DSM Directive.

[34] A similar case would be where a library or repository wishes to use the resources of other libraries (or, in general, resources held by third parties) to carry out TDM for scientific purposes. Again, given the nature of libraries as cultural heritage institutions, it would generally be possible to carry out a TDM analysis provided that the conditions set out in Article 3 of the DSM Directive are met.

[35] The Open Data Directive builds on current legislation in basic procedures and principles, and only introduces a few new obligations that *de facto* only reinforce them.

can be summarized in such a way that if any information (in terms of documents)[36] is provided, it must be done in such a way as to enable and facilitate the subsequent re-use of that information to the greatest extent possible. Of course, TDM is also an example of re-use.

From the point of view of libraries and repositories, the crucial question is whether the library can publish its databases and effectively enable TDM to be performed. Again, the main issue will be the copyrighted works that are part of the collection because the sui generis database right and database copyright can be licensed by the library or repository to allow TDM without the need for an exception. In Article 1(6), the Open Data Directive even forbids the regulated entities to exercise the sui generis database rights if this prevents the re-use of documents. The PSI Directive does not affect documents protected by copyright (and related rights) of third parties.[37] However, this conclusion must be interpreted as meaning only cases where the provision of the works in question would infringe the copyrights of third parties.[38] If the library or repository is licensed to distribute the work and communicate it to the public, it may make the works available online. In that case, the obligations arising from the PSI or the Open Data Directive will be fully applied.

However, in addition to publication, it is necessary to consider the possibility of the obligation to provide information upon request under the Act on Free Access to Information[39] for libraries and repositories that are also regulated subjects under that Act. The second possibility for disseminating copyrighted works and other resources for TDM is therefore to comply with a request for that information. This can not only provide a copy of the database with copyrighted works but also provide direct access to the database of works through the API.[40] In the context of Czech law, in certain cases it is possible to request information pursuant to Act No 106/1999 Sb. and provide third-party copyrighted documents even without a license from the author or rights executor. This is due to the statutory license to use the work for official use enshrined in Section 34 of Act No 121/2000 Sb., the Copyright Act.[41] However, a necessary requirement for using the statutory license is to pass a three-step test.[42] The third step of the three-step test is that the use of the work must not unduly affect the legitimate interests of the author. In the context of providing a copyrighted database of works for the purpose of TDM, the legal situation of the information applicant can be taken into account when evaluating the third step of the three-step test. We will therefore examine whether it qualifies for the possible application of the TDM exception and, if so, it can be argued that the provision of the works in question is not contrary to copyright protection because the rights of the author

[36] The *PSI* legislation works with the concept of information in the sense of a "document that records content", which corresponds to Buckland's concept of information. See BUCKLAND, Michael Keeble. Information as a Thing. *Journal of the American Society for Information Science and Technology*. 1991, Vol. 42, No. 5, pp. 351-360. ISSN 2330-1643.

[37] See Article 1(2)b) of the PSI Directive.

[38] See also Czech legislation, namely Section 11(2)c) of Act No 106/1999, on Free Access to Information, as amended.

[39] Typically, in the context of Czech law, these will be 'public institutions', i.e. institutions established by the state or self-governing units or fully owned by them. For more information, see Ruling of the Constitutional Court of 24 January 2007, file No I. ÚS 260/06, No N 10/44 SbNU 129 [available from **http://nalus.usoud.cz**, accessed 10 November 2019].

[40] For more information on access to and re-use of public sector information, in particular copyrighted works, see e.g. EECHOUD, Mireille Van; JANSSEN, Katleen. Rights of Access to Public Sector Information. *Masaryk University Journal of Law and Technology*. 2013, Vol. 6, No. 3, pp. 471-499. ISSN 1802-5951.; GILCHRIST, John. Accessing and Reusing Copyright Government Records. *Queensland University of Technology Law & Justice Journal*. 2010, Vol. 10, No. 2, pp. 213-232. ISSN 1445-6230.

[41] Cf. TELEC, Ivo; TŮMA, Pavel. *Copyright Act - Commentary*. 2nd edition. Prague: C. H. Beck, 2019, p. 414. ISBN 978-80-7400-748-4.

[42] See Article 5(5) of Directive 2001/29/EC. Cf. also e.g. GEIGER, Christophe; GERVAIS, Daniel; SENFTLEBEN, Martin. The Three-Step Test Revisited: How to Use the Test's Flexibility in National Copyright Law. *American University International Law Review*. 2013, Vol. 29, No. 3, pp. 581-626. ISSN 1520-460X.

are not unduly affected by this process. In conclusion, however, it must be pointed out that the TDM exception itself is not sufficient as legal title for the publication of copyrighted works.

In the context of the evaluation of the possibility of subsequent use of works through TDM, this will undoubtedly depend on the entity in question, the purpose of the TDM and, where relevant, whether copyright executors have restricted the possibility of applying the TDM exception within the meaning of Article 4(3) of the DSM Directive. If such a decision has been made, it is essential that this information is sufficiently accessible to other potential users, for example through a metadata record attached to the work in question. An important conclusion, however, is that access to works deposited in the repository, whether the works have been published in accordance with a license or under a statutory official license pursuant to Section 34 of the Copyright Act, will constitute legitimate access to the works, which is a prerequisite for the application of TDM exceptions under the DSM Directive.

Conclusion

With the DSM Directive and the exception allowing the implementation of TDM, the European legislator has embarked on a journey, the success and feasibility of which will only be decided over time. As the preliminary analyses show, the two formulated exceptions for TDM are applicable only to a very narrow number of cases. This is either because of the narrow focus, as in Article 3, or because of the factual option of the rightsholder to exclude the application of that exception, as in Article 4. The exception will have no far-reaching impact for libraries and repositories. It can be expected that this exception will facilitate the work, legal status and certainty of libraries wishing to conduct TDM analysis of the documents they manage. For external entities interested in carrying out TDM with respect to grey literature documents in depositories and libraries, there is an interesting synergy with the PSI legislation, as the TDM exception establishes a specific way to legally handle copyrighted documents and also slightly facilitates access to such documents.

References

ADRIAANS, Pieter. Information. In: ZALTA, Edward N. (ed.). *The Stanford Encyclopaedia of Philosophy* [online]. Stanford: Metaphysics Research Lab, Stanford University, 2013 [Accessed 30 June 2019]. Available from: **https://plato.stanford.edu/archives/fall2013/entries/information/**

BROOK, Michelle; MURRAY-RUST, Peter; OPPENHEIM, Charles. The social, political and legal aspects of text and data mining (TDM). *D-Lib Magazine* [online]. 2014, Vol. 20, No. 11–12 [Accessed 30 June 2019]. ISSN 1082-9873. Available from: **https://doi.org/10.1045/november14-brook**

BUCKLAND, Michael Keeble. Information as a Thing. *Journal of the American Society for Information Science and Technology.* 1991, Vol. 42, No. 5, pp. 351-360. ISSN 2330-1643.

CASPERS, Marco; GUIBAULT, Lucie. A right to 'read' for machines: Assessing a black-box analysis exception for data mining. *Proceedings of the Association for Information Science & Technology* [online]. 2016, Vol. 53, No. 1, pp. 1-5 [Accessed 30 June 2019]. ISSN 2373-9231. Available from: **https://doi.org/10.1002/pra2.2016.14505301017**

DUCATO, Rossana; STROWEL, Alain. Limitations to Text and Data Mining and Consumer Empowerment: Making the Case for a Right to "Machine Legibility". *IIC - International Review of Intellectual Property & Competition Law*. 2019, Vol. 50, No. 6, pp. 649-684. ISSN 0018-9855.

EECHOUD, Mireille Van; JANSSEN, Katleen. Rights of Access to Public Sector Information. *Masaryk University Journal of Law and Technology*. 2013, Vol. 6, No. 3, pp. 471-499. ISSN 1802-5951.

FLORIDI, Luciano. *Information: a very short introduction*. Oxford; New York: Oxford University Press, 2010, 130 p. Very short introductions, 225. ISBN 978-0-19-955137-8.

GEIGER, Christophe; FROSIO, Giancarlo; BULAYENKO, Oleksandr. Text and Data Mining in the Proposed Copyright Reform: Making the EU Ready for an Age of Big Data?: Legal Analysis and Policy Recommendations. *IIC International Review of Intellectual Property and Competition Law* [online]. 2018, Vol. 49, No. 7, pp. 814-844 [Accessed 8 October 2019]. ISSN 2195-0237. Available from: **https://doi.org/10.1007/s40319-018-0722-2**

GEIGER, Christophe; GERVAIS, Daniel; SENFTLEBEN, Martin. The Three-Step Test Revisited: How to Use the Test's Flexibility in National Copyright Law. *American University International Law Review*. 2013, Vol. 29, No. 3, pp. 581-626. ISSN 1520-460X.

GILCHRIST, John. Accessing and Reusing Copyright Government Records. *Queensland University of Technology Law & Justice Journal.* 2010, Vol. 10, No. 2, pp. 213-232. ISSN 1445-6230.

HANNAY, William M. Legally Speaking - Of Mindfields and Minefields: Legal Issues in Text and Data Mining. *Against the Grain* [online]. 2014, Vol. 26, No. 1, pp. 52-55 [Accessed 30 June 2019]. ISSN 10432094. Available from: **https://doi.org/10.7771/2380-176X.6663**

KOŠČÍK, Michal et al. Výzkumná data a výzkumné databáze. *Právní rámec zpracování a sdílení vědeckých poznatků*. Prague: Wolters Kluwer ČR, 2018, 180 p. ISBN 978-80-7552-952-7.

LUPAŞCU, Monica. Text and Data Mining Exception - Technology into Our Lives. *International Conference: CKS - Challenges of the Knowledge Society*. 2019, pp. 905–914. ISSN 2068-7796.

MYŠKA, Matěj. Text and Data Mining of Grey Literature for the Purpose of Scientific Research. *The Grey Journal*. 2017, Vol. 13, pp. 32-37. ISSN 1574-1796.

MYŠKA, Matěj; HARAŠTA, Jakub. Omezení autorského práva a zvláštních práv pořizovatele databáze v případě datové analýzy. *Časopis pro právní vědu a praxi*. 2016, Vol. 23, No. 4, pp. 375-384. ISSN 1805-2789.

MYŠKA, Matěj; ŠAVELKA, Jaromír. A Model Framework for publishing Grey Literature in Open Access. *Journal of Intellectual Property, Information Technology and E-Commerce Law*. 2013, Vol. 4, pp. 104-115. ISSN 2190-3387.

PENNEY, J.W. Chilling Effects: Online Surveillance and Wikipedia Use. *Berkeley Technology Law Journal* [online]. 2016, Vol. 31, No. 1, pp. 117–182 [Accessed 8 October 2019]. ISSN 1086-3818. Available from: **https://doi.org/10.15779/Z38SS13**

ROSATI, Eleonora. Copyright as an Obstacle or an Enabler? A European Perspective on Text and Data Mining and Its Role in the Development of AI Creativity. *Asia Pacific Law Review* [online]. Rochester, NY: Social Science Research Network (SSRN), 2019 [Accessed 8 October 2019]. Available from: **https://papers.ssrn.com/abstract=3452376**

SENSENEY, Megan; DICKSON KOEHL, Eleanor. Text data mining beyond the open data paradigm: Perspectives at the intersection of intellectual property and ethics. *Proceedings of the Association for Information Science & Technology* [online]. 2018, Vol. 55, No. 1, pp. 890-891 [Accessed 30 June 2019]. ISSN 2373-9231. Available from: **https://doi.org/10.1002/pra2.2018.14505501162**

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

TELEC, Ivo; TŮMA, Pavel. *Copyright Act - Commentary*. 2nd edition. Prague: C. H. Beck, 2019, 1295 p. ISBN 978-80-7400-748-4.

TRIAILLE, Jean-Paul et al. *Study on the legal framework of text and data mining (TDM)*. Luxembourg: Publications Office, 2014. ISBN 978-92-79-31976-1.

TRUYENS, Maarten; VAN EECKE, Patrick. Legal aspects of text mining. *Computer Law & Security Review* [online]. 2014, Vol. 30, No. 2, pp. 153-170 [Accessed 8 October 2019]. ISSN 0267-3649. Available from: **https://doi.org/10.1016/j.clsr.2014.01.009**

EXCEPTIONS FOR CULTURAL HERITAGE INSTITUTIONS UNDER THE COPYRIGHT DIRECTIVE IN THE DIGITAL SINGLE MARKET

Michal Koščík

koscik@med.muni.cz

Masaryk University, Czech Republic

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

Abstract

The paper introduces the Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market and analyses its impact on repositories of grey literature. The focus is on the provisions of Article 2, Article 6 and Article 8, which are the most relevant parts from the perspective of cultural heritage institutions. The paper concludes that repositories of grey literature and cultural heritage institutions that store and share pieces of grey literature online will benefit from the new legislation, which will bring more legal certainty in dealing with digital cultural heritage. The new rules will also improve legal certainty in cross border cooperation.

Keywords

Cultural heritage, Digital Single Market, EU law, grey literature

Introduction

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC (hereinafter the "DSM Directive"), which entered into force on 6 June 2019, aims to modernise the EU copyright framework and adapt it to the ever-evolving technologies. The DSM Directive is not a comprehensive European copyright codex for the twenty-first century. Rather, it is just another piece of a heterogeneous mosaic of approximately a dozen directives that relate to the exercise of copyright rights and access to information and cultural heritage. The DSM Directive complements the existing framework[1] and updates the framework only in specific narrow fields. It complements and updates the system of copyright exceptions granted under the InfoSoc Directive by transforming certain optional copyright exceptions to mandatory exceptions and also by making them more specific with regards to the use of works by cultural heritage institutions. The InfoSoc Directive remains in effect and serves as the ultimate limit of how far the copyright exceptions granted by national laws can reach.[2]

The preservation of European cultural heritage is firmly within the sights of the DSM Directive. The directive extends and updates the existing framework of copyright exceptions applicable to cultural heritage institutions. The copyright exceptions of the DSM Directive are designed to unburden cultural heritage institutions that collect digital works, digitise their collections and make these collections accessible online. The DSM Directive also amends the wording of the InfoSoc Directive so that Member States may introduce copyright exceptions for temporary reproductions made by cultural heritage institutions in their daily activities. However, this exception is not mandatory.

The paper aims to describe the most relevant parts of the newly adopted directive from the perspective of cultural heritage institutions. Below, the paper provides an explanation of Article 2, defining cultural heritage as a beneficiary of the exception, Article 6, setting forth rules for copyright exceptions related to cultural heritage, and Article 8, dealing with so-called out-of-commerce works.

The Cultural Heritage Institution as the Beneficiary of Copyright Exceptions

A cultural heritage institution (hereinafter a "CHI") is defined by Article 2(3) of the DSM Directive as "*a publicly accessible library or museum, an archive or a film or audio heritage institution*". The acquis does not contain any specific definition of what a library, museum or archive actually is. This definition of a CHI partially overlaps with the definition of a research organisation [see Article 2(1) of the DSM Directive] but this does not cause any practical problems because there are plenty of institutions that serve both purposes. The acquis also does not anticipate any specific legal form, funding source or organisational structure of a CHI. The status of a CHI, such as archive or library, is therefore assigned by national law. For example, a central national library can simultaneously serve as a cultural heritage institution, research organisation and even teaching organisation, meaning it could potentially rely on the

[1] The most notable directives which are partially amended or affected in their practical application are Directive 96/9/EC on the legal protection of databases, Directive 2000/31/EC on electronic commerce, Directive 2001/29/EC - InfoSoc, Directive 2012/28/EU on certain permitted uses of orphan works, and Directive 2014/26/EU on collective management of copyright and related rights.

[2] See Article 25 of the DSM Directive.

exception for text and data mining for scientific research[3], the exception to use the protected subject matter in digital and cross-border teaching activities[4], and the exception for preservation of cultural heritage[5]. Whether a respective national library in a respective Member State would actually be eligible to rely on the abovementioned exceptions would depend on its role envisaged by the laws of the respective Member State.

Mandatory Exceptions for Making Copies of Protected Content in Permanent Collections

The role of a CHI is to preserve cultural heritage for future generations. The concept of cultural heritage is gradually evolving from tangible objects, such as archaeological heritage, to broader concepts, such as folklore, intangible heritage and digital heritage[6]. Physical objects deteriorate over time, and digital storage is often a suitable place to store a backup copy and preserve the work or a document it contains[7]. Such a backup copy might be analogue or digital. The preservation of an object in a collection might require reproduction and consequently authorisation from the relevant rightsholder[8]. The copyright framework prior to the DSM Directive enabled the creation of backup digital copies[9], but the use of such backup copies was limited and basically restricted to being displayed on onsite terminals[10]. Directive 2001/29/EC on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc Directive) regulated copyright exceptions for CHIs and generally gave Member States rights to create statutory exceptions. However, the InfoSoc Directive neither obliged Member States to introduce such exceptions nor harmonised the extent of such exceptions.

The DSM Directive aims to take another step towards the harmonisation of national laws.

One objective of the DSM Directive is to unify the rules applying to digital preservation in order to promote the establishment of cross-border preservation networks in the internal market[11]. These cross-border networks should enable the pooling of resources and expertise across Europe. Items of cultural heritage of a Member State that lacks expert knowledge in a certain area of digitization could be digitised by a specialised institution from another Member State without worrying about the compatibility of copyright exceptions in the two countries. The pooling of the resources does not have to be based only on expertise. The cross-border preservation of certain forms of cultural heritage can also make sense for economic reasons.

[3] See Article 3 of the DSM Directive - Article 3(1) Member States shall provide for an exception to the rights provided for in Article 5(a) and Article 7(1) of Directive 96/9/EC, Article 2 of Directive 2001/29/EC, and Article 15(1) of this Directive for reproductions and extractions made by research organisations and cultural heritage institutions in order to carry out, for the purposes of scientific research, text and data mining of works or other subject matter to which they have lawful access.
[4] See Article 5 of the DSM Directive.
[5] See Article 6 of the DSM Directive.
[6] See KOŠČÍK, Michal and Matěj MYŠKA. Copyright Law Challenges to the Preservation of "Born-Digital" Digital Content as Cultural Heritage. *European Journal of Law and Technology* [online]. 2019, **10**(1) [Accessed 8 September 2019]. ISSN 2042-115X. Available from: **http://ejlt.org/article/view/664**
[7] See Koščík.
[8] See recital 25 of the DSM Directive.
[9] See Article 5(2)(c) of the InfoSoc Directive.
[10] See Article 5(2)(n) of the InfoSoc Directive.
[11] See recital 26 of the DSM Directive.

In accordance with Article 6 of the DSM Directive, all Member States have to introduce statutory exceptions that would enable CHIs to make copies of any works or other subject matter "*that are permanently in their collections, in any format or medium, for purposes of preservation of such works or other subject matter and to the extent necessary for such preservation"*[12]*.* Article 6 of the DSM Directive provides an exhaustive list of the rights affected by this mandatory statutory exception. As a result, CHIs can make temporary or permanent copies of any original work protected by copyright[13] or neighbouring right. The protected objects of intellectual property that will be covered under such exceptions are:

- Temporary or permanent reproductions of databases, regardless of whether these databases are protected by copyright[14] or sui generis database rights;[15]
- Any performance of a performing artist, or its fixation;[16]
- Copies of phonograms and films;[17]
- Reproductions of any audio and audio-visual broadcasts, whether those broadcasts are transmitted by wire or over the air, cable or satellite;[18]
- Copies of computer programs, and their translations and adaptations;[19]
- Press publications and online news.[20]

CHIs will be able to make backup copies of any item in their collections and freely convert the formats of such backup copies. Works on canvas can be digitized, and digital works can be printed. However, the exceptions enacted under the DSM Directive do not authorise CHIs to create backup copies of items that are not part of a permanent collection. For example, it is not possible to cover reproductions of items that are temporarily borrowed from other CHIs.

The distinction between a permanent and temporary collection can be interpreted in accordance with recital 29 of the DSM Directive. According to recital 29, works are permanently in the collection of a CHI *"when copies of such works or other subject matter are owned or permanently held"* by the CHI*, "for example as a result of a transfer of ownership or a licence agreement, legal deposit obligations or permanent custody arrangements"*. The definition of a "permanent collection" is another step towards legal certainty, as the use of this term created interpretational problems in the past[21]. The existence of a license or license agreement can be sufficient for an CHI to consider an item part of its permanent collection. If a provision of the license agreement, or any other type of contract, forbids the CHI from taking advantage of the statutory exceptions under Article 6, it should be considered unenforceable[22].

[12] See below for the definition of "other subject matter".
[13] Article 2 of Directive 2001/29/EC.
[14] Article 5(a) of Directive 96/9/EC.
[15] Article 7(1) of Directive 96/9/EC.
[16] Article 2 of the InfoSoc Directive.
[17] Article 2 of the InfoSoc Directive.
[18] Article 2 of the InfoSoc Directive.
[19] Article 4(1) of Directive 2009/24/EC.
[20] Article 15 of the DSM Directive.
[21] This problem was discussed in the previous works of the author, see KOŠČÍK, Michal and Matěj MYŠKA. Copyright Law Challenges to the Preservation of "Born-Digital" Digital Content as Cultural Heritage. *European Journal of Law and Technology* [online]. 2019, **10**(1) [Accessed 8 September 2019]. ISSN 2042-115X. Available from: **http://ejlt.org/article/view/664** and KOŠČÍK, Michal. Legal Framework for the Digitisation and Storage of Digital Works by Public Archives. *The Grey Journal*. Amsterdam: GreyNet, 2019, **15**(Special Winter Issue), 52-57. ISSN 1574-1796.
[22] Article 7 of the DSM Directive.

The wording of the DSM Directive is ambiguous regarding whether a CHI can rely on a third party (subcontractor or other CHI) to digitize or even archive a digital copy of the object of cultural heritage. Recital 28 of the DSM Directive explicitly states that *"the cultural heritage institutions should be allowed to rely on third parties acting on their behalf and under their responsibility, including those that are based in the other Member States, for the making of copies"*, but this objective is not directly reflected in the normative part of the directive. It is however possible to use, by analogy, the interpretation in the InfoSoc Directive, which has a similarly worded provision and has already been favourably interpreted by the Court of justice of the European union (CJEU). Article 5(2)(d) of the InfoSoc Directive conferred copyright exception to temporary copies made by broadcasting institutions. The CJEU ruled that this article had to be interpreted *as "meaning that a broadcasting organisation's own facilities include the facilities of any third party acting on behalf of or under the responsibility of that organisation"*[23]. It can be concluded that the exception under Article 6 of the DSM Directive does not prohibit CHIs from relying on contractors (third parties) which would act on behalf of a CHI and under the responsibility of that CHI in the process of cultural heritage digitisation. These parties could also rely on the new exception under the newly formulated Article 5(2)(d) of the InfoSoc Directive.[24]

Use of Out-of-Commerce Works by CHIs

The exceptions granted by the InfoSoc Directive and also Article 6 of the newly introduced DSM Directive are suitable for making backup copies of items of a permanent collection, however they are not really useful for making the content available online. Article 8 of the DSM Directive partly addresses this gap in the system of copyright exceptions. It obliges all Member States to create a mechanism under which a CHI is allowed to reproduce, distribute, communicate and make available online all the out-of-commerce works in its collection, providing that they conclude a contract with the representative collective management organisation.[25]

Out-of-commerce works should not be confused with "orphan works" or "works in the public domain". Orphan works are works for which the author cannot be defined or found[26], whereas out-of-commerce works are works no longer offered to consumers via the media. Works in the public domain are works no longer under copyright protection and which can be digitized, distributed and shared without any restrictions.

Certain types of protected content such as databases, computer programs, records of TV broadcasts or news articles often do not have a representative collective management organisation[27]. If such organisation does not exist in a respective Member State, the CHI will

[23] European Court of Justice, Case C-510/10, DR, TV2 Danmark A/S v NCB - Nordisk Copyright Bureau.

[24] See above.

[25] See par. 1 of Article 8.

[26] See Article 2 of the Orphan Works Directive: "A work or a phonogram shall be considered an orphan work if none of the rightsholders in that work or phonogram is identified or, even if one or more of them is identified, none is located despite a diligent search".

[27] On questions of legitimacy and representativeness of CHIs, see: GUIBAULT, Lucie, and Simone SCHROFF. Extended Collective Licensing for the Use of Out-of-Commerce Works in Europe: A Matter of Legitimacy Vis-a-Vis Rights Holders. *IIC-International Review of Intellectual Property and Competition Law* [online]. 2018, **49**(8), 916-939 [Accessed 8 September 2019]. ISSN 2195-0237. Available from: **https://doi.org/10.1007/s40319-018-0748-5**, and STRAKOVÁ, Lucie. Changes in the Area of Extended Collective Management in Relation to Memory and Educational Institutions in the Light of the Czech Amended Copyright Act. *The Grey Journal*. Amsterdam: GreyNet, 2018, **14**(Special Winter Issue), 61-65. ISSN 1574-1796.

be able to rely on exceptions based on the wording of the second paragraph of Article 8 of the DSM Directive. The CHIs will be allowed to make works available to the public even without the explicit consent of the (non-existent) collective management organisation. The works have to be made available on a non-commercial basis and identify the name of the author or rightsholder to such content.

Relevance of the DSM Directive for Grey literature

Repositories of grey literature can rely on copyright exceptions under Article 6 of the DSM Directive and also on the legal framework of collective rights management under Article 8. The DSM Directive brings good news for repositories of grey literature since it directly addresses works previously not associated with cultural heritage institutions such as databases, software and records.

Repositories of grey literature contain works not created with the intention to be made available to the general public or distributed by the media. The fact that the DSM Directive also covers "never-in-commerce works" improves the legal certainty for the storage of grey literature and making it available online. The recitals of the DSM Directive directly address works never intended for commercial use such as posters, leaflets, Trench Journals, amateur audio-visual works, unpublished works or other subject matter[28]. Therefore, grey literature repositories can, in principle, rely on the same legal framework as other CHIs.

CHIs, however, need to be careful when dealing with works never intended for commerce but designed as preparatory versions of works published subsequently. A different manifestation of a work commercially available (such as digital and printed formats of the same work)[29] cannot be qualified as an independent out-of-commerce work. The work may be considered to be out of commerce even if its adaptations, such as translations, are commercially available[30].

Conclusion

This paper has explained the concepts behind the newly introduced Directive on Copyright and Related Rights in the Digital Single Market relevant to repositories of grey literature and other cultural heritage institutions that archive and make out-of-commerce works available. The concepts of copyright exceptions and collective rights management mechanisms are not completely new and are based on already existing frameworks. The DSM Directive, however, turns certain copyright exceptions from optional to mandatory. Until today, the Member States could introduce exceptions for cultural heritage institutions. After 7 June 2021, all Member States will have to introduce them. This will most likely have a positive impact on cross-border cooperation in the digitisation and digital distribution of digitised copies between Member States.

[28] See recital 37 of the DSM Directive.
[29] See recital 37 of the DSM Directive.
[30] See recital 37 of the DSM Directive.

In general, the new directive is favourable towards cultural heritage institutions and especially repositories of grey literature, as it explicitly includes "other subject matter" such as databases, computer programs and records, as well as works never intended for commercial use.

The directive is an act of secondary legislation and will not have a direct effect on relationships between cultural heritage institutions and rightsholders. Repositories will, therefore, have to wait until national implementations take effect to be able to take advantage of the new rules.

References

MYŠKA, Matěj, 2018. Orphan and Out-Of-Commerce Works after the Amendment of the Czech Copyright Act. *The Grey Journal*. Amsterdam: GreyNet, 2018, **14**(Special Winter Issue), 55-60. ISSN 1574-1796.

GUIBAULT, Lucie, and Simone SCHROFF. Extended Collective Licensing for the Use of Out-of-Commerce Works in Europe: A Matter of Legitimacy Vis-a-Vis Rights Holders. *IIC-International Review of Intellectual Property and Competition Law* [online]. 2018, **49**(8), 916-939 [Accessed 8 September 2019]. ISSN 2195-0237. Available from: **https://doi.org/10.1007/s40319-018-0748-5**

KOŠČÍK, Michal and Matěj MYŠKA. Copyright Law Challenges to the Preservation of "Born-Digital" Digital Content as Cultural Heritage. *European Journal of Law and Technology* [online]. 2019, **10**(1) [Accessed 8 September 2019]. ISSN 2042-115X. Available from: **http://ejlt.org/article/view/664**

KOŠČÍK, Michal. Legal Framework for the Digitisation and Storage of Digital Works by Public Archives. *The Grey Journal*. Amsterdam: GreyNet, 2019, **15**(Special Winter Issue), 52-57. ISSN 1574-1796.

STRAKOVÁ, Lucie. Changes in the Area of Extended Collective Management in Relation to Memory and Educational Institutions in the Light of the Czech Amended Copyright Act. *The Grey Journal*. Amsterdam: GreyNet, 2018, **14**(Special Winter Issue), 61-65. ISSN 1574-1796.

Case law:

European Court of Justice. Case C-510/10, Judgment of the Court (Third Chamber), 26 April 2012, DR and TV2 Danmark A/S v NCB – Nordisk Copyright Bureau, ECLI:EU:C:2012:244.