conference on grey literature and repositories

proceedings 2015

CONFERENCE ON GREY LITERATURE AND REPOSITORIES

Proceedings

National Library of Technology, 2015

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

English conference website (**http://nrgl.techlib.cz/index.php/8th_year_of_the_conference**)

Czech conference website (**http://nusl.techlib.cz/index.php/8_rocnik_konference**)

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

Editors: Mgr. Hana Vyčítalová, Bc. Michaela Charvátová

ISSN: 2336-5021

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

Programme Committee:

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

Ing. Jozef Dzivák, Slovak Chemistry Library

Dr. Dominic Farace, GreyNet

Ing. Martin Lhoták, Academy of Sciences Library

Ing. Jan Mach, University of Economics, Prague

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

Dr. Dobrica Savić, Nuclear Information Section, IAEA

Organizing Committee

Bc. Michaela Charvátová, National Library of Technology

Mgr. Lenka Patoková, National Library of Technology

PhDr. Petra Pejšová, National Library of Technology

Mgr. Hana Vyčítalová, National Library of Technology

List of Reviewers:

Ing. Naděžda Andrejčíková, Ph.D., Cosmotron

Drs. Marnix van Berchum, DANS

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

Dr. Jan Dvořák, Charles University in Prague

Dr. Dominic Farace, Greynet

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

Mgr. Adéla Jarolímková, Ph.D., National Medical Library Prague

Prof. Keith Jeffery, Keith G Jeffery Consultants

JUDr. Pavel Koukal, Ph.D., Masaryk University

PhDr. Hana Landová, Ph.D., Czech University of Life Sciences Prague

Ing. Martin Lhoták, Academy of Sciences Library

Mgr. Lenka Němečková, Czech Technical University of Prague

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

PhDr. Petra Pejšová, National Library of Technology

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

PhDr. Radka Římanová, Charles University in Prague

Christiane Stock, The Institute for Scientific and Technical Information

Mgr. Václav Stupka, Masaryk University

Mgr. Daniela Tkačíková, VŠB-Technical University of Ostrava

Marcus Vaska, University of Calgary

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

Table of Contents

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

INIS: NUCLEAR GREY LITERATURE REPOSITORY

Dobrica Savić

d.savic@iaea.org

International Atomic Energy Agency (IAEA), United Nations

Abstract

As one of the world's largest collections of published information on the peaceful uses of nuclear science and technology, INIS represents an extraordinary example of world cooperation. Currently, as INIS members, 130 countries and 24 international organizations share and allow access to their valuable nuclear information resources, preserving them for future generations and offering a freely available nuclear knowledge repository. Since its creation in 1970, INIS has collected and provided access to more than 3.8 million bibliographic references to publications, documents, technical reports, non-copyrighted documentation, and other grey literature, as well as over a million full texts. Public interest throughout the years in accessing the INIS Collection has been remarkable. This paper deals with the challenges faced by INIS in its endeavour to increase the use, accessibility, usability and expandability of its on-line repository. It also describes document collection, the features and characteristics of implementing a new search engine, as well as the lessons learned.

Keywords

Nuclear Information, Document Repository, Grey Literature, INIS, IAEA

Introduction

The International Nuclear Information System (INIS) hosts one of the world's largest collections of published information on the peaceful uses of nuclear science and technology. It offers on-line access to a unique collection of 3.8 million bibliographic records and over a million full texts of non-conventional (grey) literature. This large repository collection suffered from most of the well-known shortcomings of the classic meta-data catalogue. Searching was complex and complicated, it required training in Boolean logic, full text searching was not an option, and response time was slow. An opportune moment to improve the system came with the retirement of the previous database software and the adoption of Google Search Appliance (GSA) as an organization-wide search engine standard. INIS was quick to realize the potential of using such a well-known application to replace its on-line search engine.

This paper deals with the challenges faced by INIS regarding its endeavour to increase the use, accessibility, usability and expandability of its on-line repository. It also describes document collection, the features and characteristics of implementing a new search engine, as well as the lessons learned. Although based on specific INIS practice and experience, this paper also offers guidelines on ways to improve classic collections comprised of millions of bibliographic and full text documents, while reaping the multiple benefits of a well-organized grey literature repository, such as increased use, accessibility, usability, expandability and improved user search and retrieval experiences.

International Atomic Energy Agency (IAEA)

The IAEA is regarded as the world's centre of cooperation in the field of safe, secure and peaceful uses of nuclear technologies. It was set up in 1957 as the world´s "Atoms for Peace"[1] organization within the United Nations system. As of September 2015, the IAEA has 165 Member States.

The IAEA Secretariat is headquartered at the Vienna International Centre in Vienna, Austria. It also operates liaison and regional offices in Geneva, Switzerland; New York, USA; Toronto, Canada; and Tokyo, Japan. The IAEA runs and supports research centres and scientific laboratories in Vienna and Seibersdorf, Austria; Monaco; and Trieste, Italy. The IAEA Secretariat is a team of 2300 multi-disciplinary professional and support staff from more than 100 countries.

The IAEA's mission is guided by the interests and needs of Member States, strategic plans and the vision embodied in the IAEA Statute. Three main pillars – or areas of work – underpin the IAEA's mission: Safety and Security; Science and Technology; and Safeguards and Verification.

The work of the IAEA is carried out through six departments[2]: Nuclear Energy, Nuclear Safety and Security, Nuclear Science and Applications, Safeguards, Technical Cooperation, and the

[1] **http://www.iaea.org/About/about-iaea.html**
[2] **http://www.iaea.org/Publications/Reports/Anrep2012/orgchart.pdf**

Department of Management. Although supporting the entire Agency, the Nuclear Information Section (NIS) is organizationally part of the Department of Nuclear Energy. The Department's main tasks are to foster the efficient and safe use of nuclear power by supporting interested Member States in improving the performance of nuclear power plants, the nuclear fuel cycle, and the management of nuclear wastes; catalysing innovation in nuclear power and fuel cycle technologies; development of indigenous capabilities for national energy planning; the deployment of new nuclear power plants; and the advancement of science and industry through improved operation of research reactors.

IAEA Nuclear Information Goals

One of the IAEA's goals is the collection, preservation and dissemination of nuclear information and knowledge, which in turn is the main responsibility of NIS.

NIS consists of the IAEA Library Unit, the INIS Unit and the Systems Development and Support Group. It fosters the exchange of scientific and technical information on the peaceful use of nuclear science and technology; increases awareness in Member States of the importance of maintaining efficient and effective systems for managing such information; provides information services and support to Member States and to the IAEA; and assists with capacity building and training.

INIS

INIS represents an extraordinary example of world cooperation, where 154 members allow access to their valuable nuclear information resources, including grey literature, in order to preserve world peace and further increase the use of nuclear energy for peaceful purposes. In addition to over a million full texts, more than 3.8 million bibliographic references to publications, documents, technical reports, non-copyrighted documentation, and other grey literature are made available. Overall, there are 800 GB of data in the INIS Collection.

Full text documents available from the INIS Collection almost entirely represent nuclear related non-conventional literature (NCL) or grey literature. This PDF collection contains some very important historic and technical documents collected by INIS during the last 45 years. Much of that documentation originated in paper form only, so digitization of that collection was a huge project which went back and converted millions of microfiche pages to electronic, fully searchable, files. Optical Character Recognition (OCR) was performed on all documents within the collection making it easy to index and search. Besides being a source of information when searching, the availability of full text gives INIS a special role in the area of nuclear information and documentation — being the main custodian of this world information heritage and preserving this codified, specialized, scientific and technical knowledge.

INIS Collection

On average, INIS adds around 120,000 bibliographic records and 13,000 full text PDF documents to its collection annually. The collection is freely accessible from the INIS Search

website[3]. Only a small portion of the full text documents is regarded as restricted and kept for internal use.

The INIS Collection covers around 50 well defined subject categories which are regularly maintained by INIS and provides the scope descriptions used by national and regional centres to categorize nuclear literature for INIS input. The INIS Joint Reference Series publications are also available on the INIS website[4].

The INIS Collection covers all aspects of the peaceful uses of nuclear science and technology such as nuclear reactors, reactor safety, nuclear fusion, applications of radiation and radioisotopes in medicine, agriculture, industry and pest control, as well as related fields of nuclear chemistry, nuclear physics and materials science. Special emphasis is placed on the environmental, economic and health effects of nuclear energy. Legal and social aspects associated with nuclear energy are also covered. Fig. 1 lists a complete set of INIS Subject Categories.


S01 - Coal, lignite, and peat
S02 - Petroleum
S03 - Natural gas
S04 - Oil shales and tar sands
S07 - Isotopes and radiation sources
S08 - Hydrogen
S09 - Biomas fuels
S10 - Synthetic fuels
S11 - Nuclear fuel cycle and fuel materials
S12 - Management of radioactive wastes, and non-radioactive wastes
S13 - Hydro energy
S14 - Solar energy
S15 - Geothermal energy
S16 - Tidal and wave power
S17 - Wind energy
S20 - Fossil fueld power plants
S21 - Specific nuclear reactors and associated plants
S22 - General studies of nuclear reactors
S24 - Power transmission and distribution
S25 - Energy storage
S29 - Energy planning, policy and economy
S30 - Direct energy conversion
S32 - Energy conservation, consumption, and utilization
S33 - Advanced propulsion systems
S36 - Materials science
S37 - Inorganic, organic, physical and analytical chemistry
S38 - Radiation chemistry, radio chemistry and nuclear chemistry
S42 - Engineering
S43 - Particle accelerators
S46 - Instrumentation related to nuclear science and technolgoy
S47 - Other instrumentation
S54 - Environemntal sciences
S58 - Geosciences
S60 - Applied life sciences
S61 - Radiation protection and dosimetry
S62 - Radiology and nuclear medicine
S63 - Radiation, thermal, and other environmental pollutant effects on living organisms and biological materials
S70 - Plasma physics and fusion technology
S71 - Classical and quantum mechanics, general physics
S72 - Physics of elementary particles and fields
S73 - Nuclear physics and radiation physics
S74 - Atomic and molecular physics
S75 - Condensed matter physics, superconductivity and superfluidity
S77 - Nanoscience and nanotechnology
S79 - Astrophysics, cosmology and astronomy
S96 - Knowledge management and preservation
S97 - Mathematical methods and computing
S98 - Nuclear disarmament, safeguards and physical protection
S99 - General and miscellaneous

Figure 1 INIS Subject Categories

[3] **http://inis.iaea.org/search/**
[4] **http://nkp.iaea.org/INISSubjectCategories/**

A break-down of the amount of documents by major category is provided in FIG 2, while FIG. 3 gives a similar break-down according to various document types in the Collection.

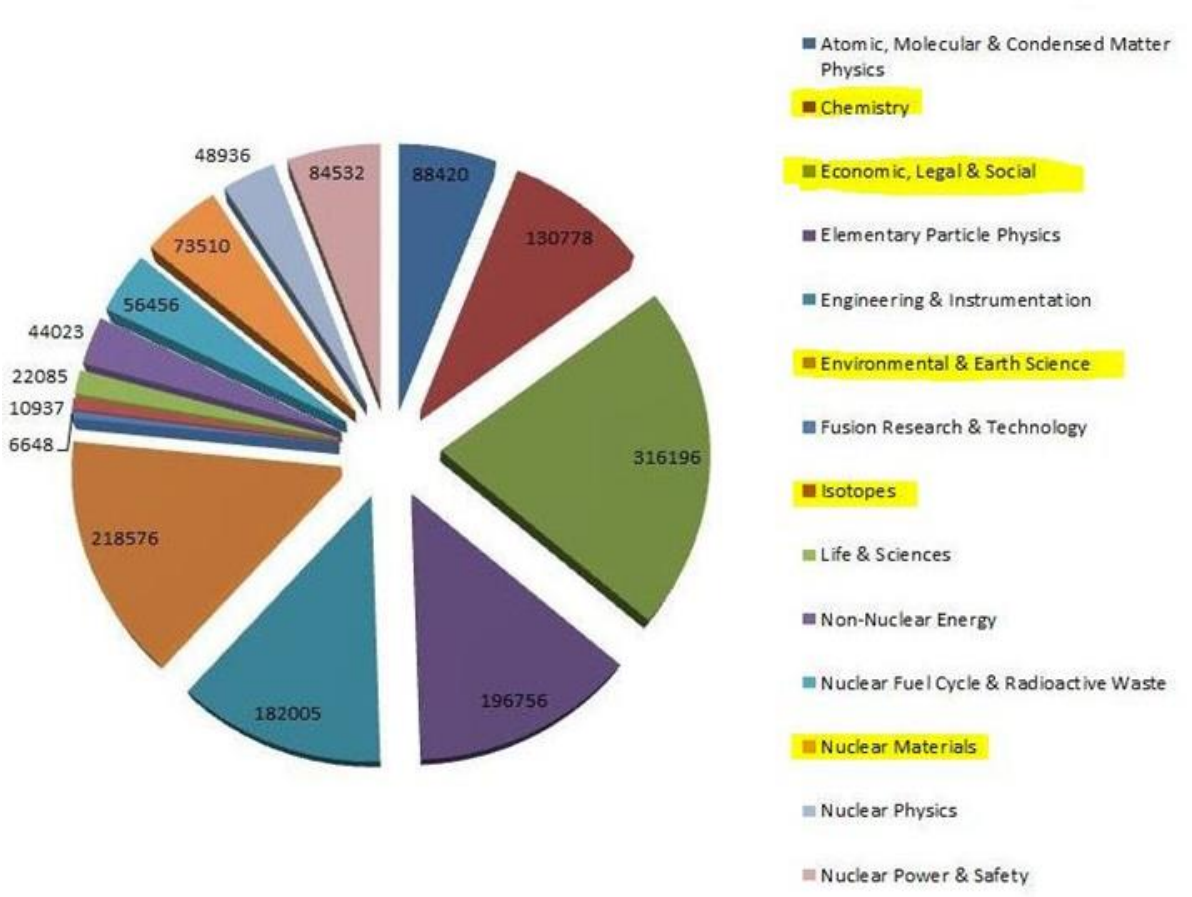

Figure 2 INIS Collection by Subject

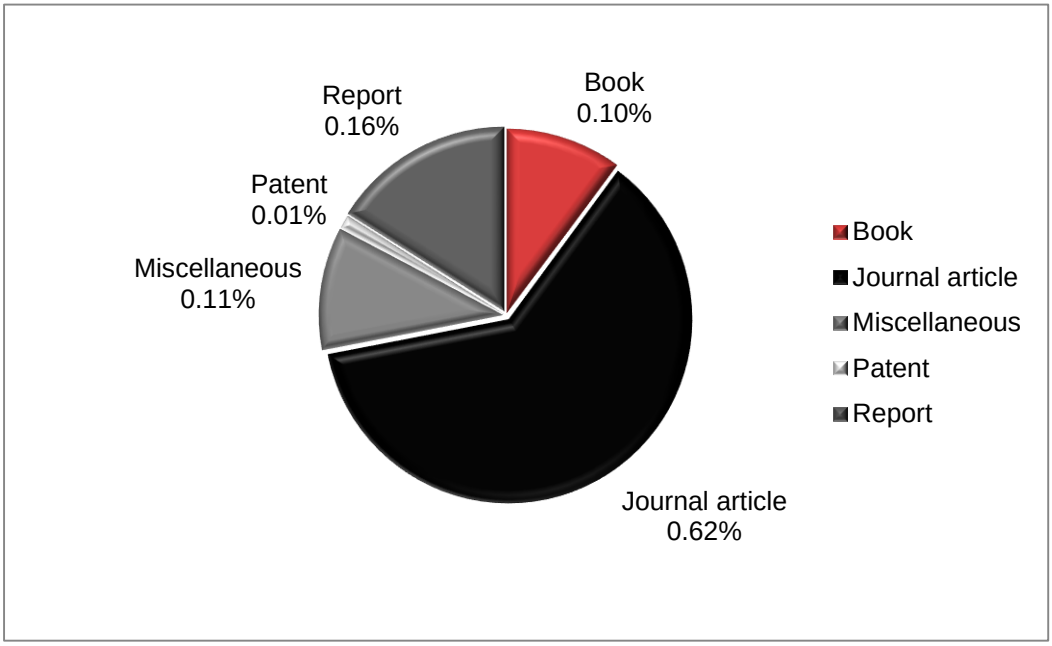

Figure 3 Bibliographic Records by Type (Status on 1 October 2015: 3,846,526 records)

INIS Search engine

Since its inception, the INIS Collection has operated in a controlled environment, requiring users to register through their national INIS centre, as well as the INIS Secretariat headquarters in Vienna, before being given access to the Collection. This changed in April 2009, when INIS became a free, open, and unrestricted information resource for internet users around the world. The opening of the Collection simplified access to reliable nuclear information on the peaceful uses of nuclear science and technology, including non-conventional literature, and made nuclear knowledge readily available worldwide for research, development and other uses. The opening of the Collection resulted in a significant increase of users.

The old search engine, a well-known BASIS Extended Relational Database Management System (ERDMS)[5], was in operation from almost the beginning of INIS until 2011. Eventually, BASIS merged with an RDBMS system called DM, and became known as BASISplus. Currently, BASISplus is owned and supported by OpenText.

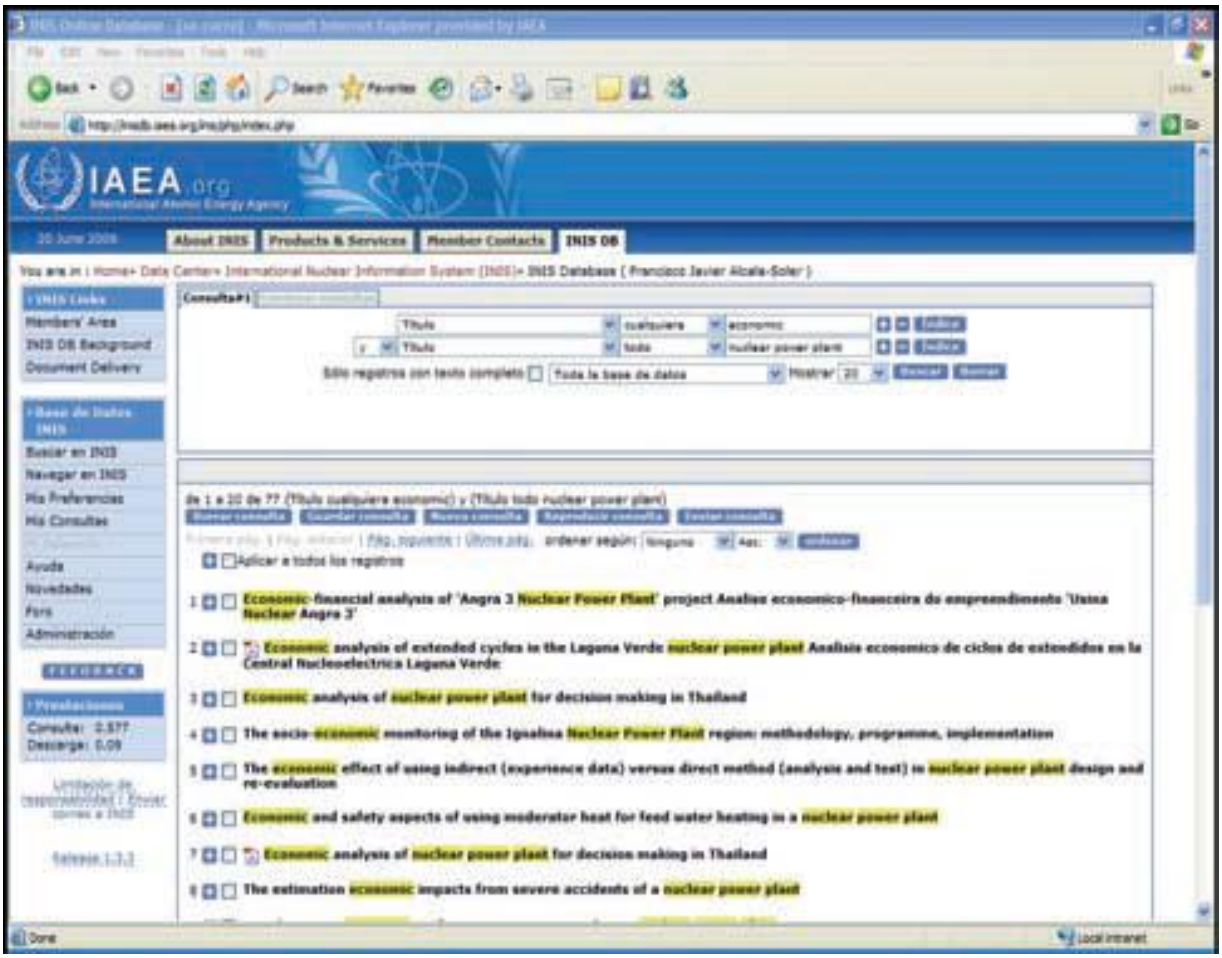

Figure 4 Old INIS BASIS Search Interface

[5] **http://en.wikipedia.org/wiki/Basis_database**

The problems INIS experienced with its version of BASISplus included slow response time, a cluttered interface made for librarians, only an advanced search option, no full text indexing, issues with support for a multilingual search interface, and lack of support for the thesaurus and different authorities The new search engine was installed in 2010, and became fully operation in April 2011. It was based on Google Search Appliance (GSA), which, at that time, became the IAEA-wide search engine standard. Although INIS reviewed a number of different search engine offerings, the decision was made to use GSA as a de facto on-line catalogue and to search the digital collection of INIS records.

This decision was based on the following characteristics offered by GSA: great speed and scalability; uncluttered and easy to use starting interface; the possibility to use advanced options and to broaden or tighten searches; powerful full text indexing; retrieval of more relevant results; and the existence of faceted/filtered search features. The images shown below are actual snapshots of two different versions of the INIS GSA-based standard (simple) search interfaces, both deployed on the INIS website.

Figure 5 ICS Standard Interfaces

The ICS Standard interface offers a Main Search Box for entering search terms; a Tool Bar linking to INIS Home, Thesaurus, Browse, and Advanced Search, and the Language Selector. Both ICS Standard and Advanced Search interfaces are available in eight languages (Arabic, Chinese, English, French, German, Japanese, Russian, and Spanish). RSS subscribe and email options are also available.

All words put into the query are used; searches are always case insensitive; punctuation is ignored, including @ # $ % ^ & * ( ) = + [ ] \ and other special characters; and common articles or determiners, i.e. stop words, such as 'the,' 'a,' and 'for,' are usually ignored. The inclusion of stop words for languages other than English is also possible.

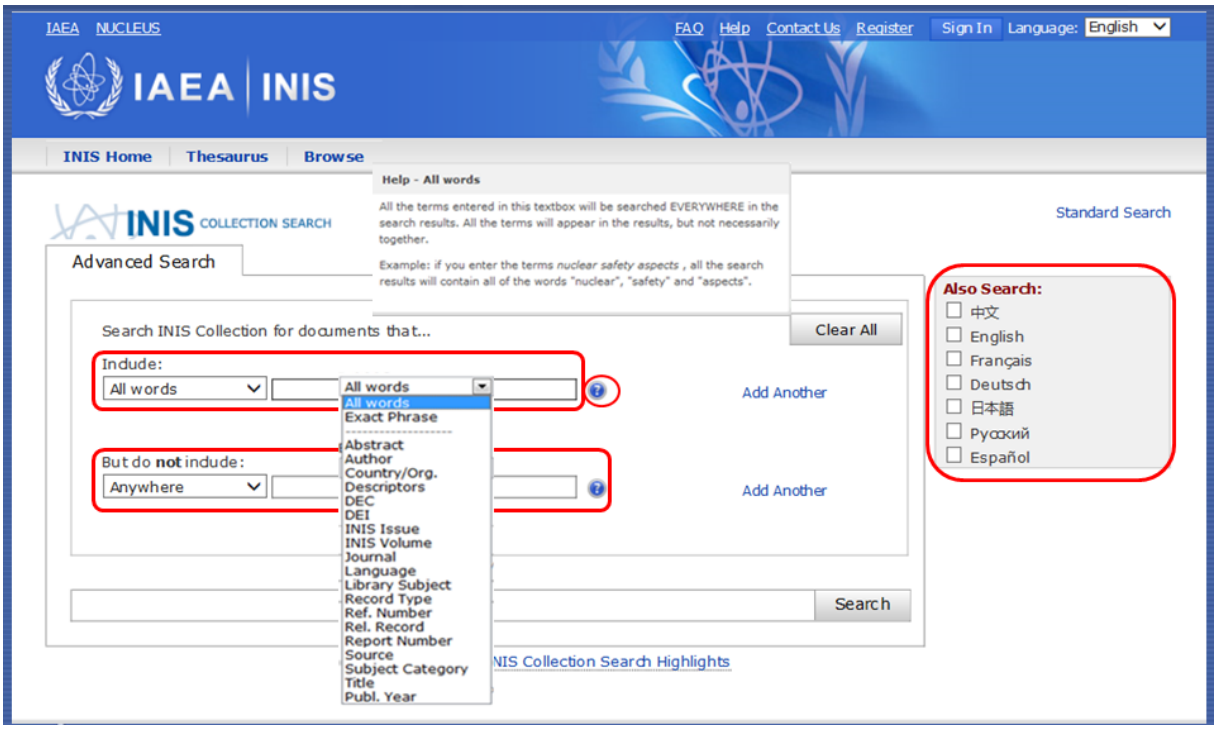

Figure 6 Features of the ICS Advanced Interface

The Standard Search interface provides excellent results for most search requirements. Our statistics show that the overwhelming majority of visitors (almost 99%) use the ICS Standard Search interface, or other channels, such as access through Google Scholar. Many other studies demonstrate that users are not inclined to become expert searchers (Novotni, 2004; Wallace, 1993; Valentine, 1993). However, if a more precise search is needed, a query can be constructed using the query builder form on the Advanced Search page. Query builder also generates a query syntax which appears in the Advanced Search query box. Alternatively, very experienced users can type the query directly into the Advanced Search query box. In this case, the query builder form becomes disabled.

The Advanced ICS Search (Fig. 6), offers the possibility to search all words or an exact phrase; to select (include or exclude) specific metadata fields; and to select the language of publications to be covered by the search. It also supports 'range queries' which enable the user to search for results where field values are between the lower and upper range specified by the query. Range queries are specified using the Range operator, written as two consecutive dots (e.g. year: 2007..2009). The dropdown menu of the Advanced Search offers the list of fields (metadata elements) in which one can search. This includes abstract, author, country/organization, descriptors, title, subject category, publication year, report number, etc.

Main Features of the INIS Collection Search (ICS)

The main features of the INIS Collection Search include:

Platform: Project, started in 2010, using Google Search Appliance© technology - renowned, simple, fast, flexible; Current version is 4.4 (May 2014); Virtual servers are within the IAEA, offering 24/7 availability.

Accessibility: Free and open web-based application; No installation required; Mobile Apps available for iPAD, iPHONE and Android; Widget prepared allowing searches from other websites, with or without filter.

Ease of use: Intuitive, self-explanatory initial screen; No previous knowledge of the INIS Collection is required; A single search box communicates confidence to users that the search tool can meet their information needs from a single point of entry.

Advanced Search: Offers more precise searches and metadata, as well as Boolean searches; Query builder has a built-in syntax generator; Ability to search for all words and exact phrase; Users can include or exclude metadata, select language; Range queries (2007..2009) and dropdown menus are available.

Faceted search: Dynamic navigation through country, language, publication year, and INIS Volume.

Expandability: ICS can be expanded to new collections, databases, repositories, new formats and type of documents.

Multilingualism: ICS interface and INIS Thesaurus are available in 8 languages, as well as a cross-language search. Automatic Google translation into 8 languages is also present.

Authority files: Journal titles, CODEN, ISSN, Subject category, Descriptor, Country/Organization, Author, and Report number authorities are incorporated in the ICS. Report number authority files are not frequently encountered in other collections and data bases, but INIS maintains this unusual feature as a sign of its dedication to quality document management and in order to accommodate INIS users who know the report series.

Usability: Users can print or export results in different formats (PDF, HTML, Excel, XML), download citations in plain text, RIS format, RefWorks, EndNote, create RSS feeds, e-mail search results as a link; Stop words for languages other than English, and translation of bibliographic records into other languages using Google Translator.

User profiling: User registration offers cross-Agency Single Sign-On, interface language set up, number of displayed results per page, saving queries and search updates, as well as emailing query results; Workspace concept also offers to save documents and translate into other languages.

Help: FAQ (on INIS and on ICS), on-line help, pop-up hints (examples on how to build a query using metadata), and e-training course are available.

Conclusion

The decision to replace the previous, almost 40 year old, INIS classic bibliographic on-line database and search engine was neither easy nor cheap. Finding a best fit replacement was even more difficult. A number of available search engines were examined and evaluated by a team of various information and computer specialists. The decision was to go for a market leader, namely the Google Search Appliance (GSA). Although it might have seemed risky at the time, the years following its implementation have shown that it was the right decision. It has brought considerable benefits, primarily to INIS Collection end-users.

The overall goal was to increase use, accessibility, usability, and expandability of the INIS on-line repository with 3.8 million bibliographic records and over a million full text nuclear related documents. This was achieved in the following manner:

Increased use - By making the repository open and freely accessible to the public and by replacing a legacy database search with a top-of-the-line one. While working on this goal, it became obvious that provision of full text documents was the users' most sought-after feature, making the new interface particularly beneficial to nuclear grey literature.

Improved usability - Meeting user needs was a tricky requirement due to the users' diverse experiences and backgrounds. Some users desired advanced search capabilities, while the majority wanted a simple and straightforward search. Statistics have proven that users mostly use a simplified, basic search interface. Because only by a small number of users (less than a few percent of all users) use the advanced search, this feature was improved upon but made discrete.

Expandability (scalability) – The most efficient way of providing for expandability is by getting a top performance technical solution, while making sure that different record formats are part of the overall solution and that they can be used from the beginning or added at a later stage.

Increased accessibility - Locating and accessing a relevant repository requires preparation and substantial ground work. Incorporating the repository and the document collection with Google.com, Google Scholar and other search engines (e.g. worldwidescience.org), can dramatically increase both the number of searches and users looking for relevant information. Together with incorporating the repository and making it available through other search engines, promotion is also essential. Conferences, articles, presentations, and poster sessions, all provide a great channel to promote the repository to the relevant user groups.

The success of the upgraded INIS repository and user satisfaction were evident from the feedback received. The number of visits dramatically increased and most importantly, the number of full text document downloads went up. The number of INIS Collection users dramatically increased with the introduction of GSA, even more so by connecting INIS to WorldWideScience.org, and, finally, by making the INIS Collection available on Google.com and Google Scholar. The number of searches went up fivefold and the number of downloads more than tenfold within the first few months.

In conclusion, this example of the INIS grey literature repository clearly demonstrates the options and tools that are available to attract more users and successfully boost the use of available information resources.

References

NOVOTNI, Eric (2004). I Don't Think I Click: A Protocol Analysis Study of Use of a Library Online Catalog in the Internet Age. *College & Research Libraries* [online]. 2004, **65**(6), p. 525-537. Available from: **http://crl.acrl.org/content/65/6/525.full.pdf**.

VALENTINE, Barbara (1993). Undergraduate Research Behavior: Using Focus Groups to Generate Theory. *Journal of Academic Librarianship*. 1993, **19**(5), p. 300-304.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

REPOSITORY WORKFLOW FOR INTERLINKING DATA WITH GREY LITERATURE

Johanna Vompras, Jochen Schirrwagen

johanna.vompras@uni-bielelefeld.de, jochen.schirrwagen@uni-bielefeld.de

Bielefeld University Library, Universitätsstr. 25, Bielefeld, 33615, Germany

Abstract

Publishing data is more and more considered as part of the research process. While funder mandates and journal policies demand the disclosure of research data at the time of article publication there is still a lack of guidelines and workflows to reference data from grey literature. Based on multidisciplinary examples found in our repository 'PUB' we present a user friendly generalized framework for interlinking research data with grey literature. This way, we are not only increasing the number of 'grey' non-textual research outputs - including research data - but also foster awareness of its sharing and re-use in scientific communities.

Keywords

Research Data, Linking Data to Publications, RDM, Institutional Research Infrastructures

Introduction

Research data and other nontraditional output types are increasingly considered as valuable as research papers and other textual publications. Research data that can be persistently identified relate to increased visibility, citability and re-use [4]. They may allow for reproducibility and validation of research results. This imposes requirements on the research

data management, documentation using descriptive and disciplinary metadata standards, storage, interlinking with their publication, curation and optional anonymization.

Grey literature mostly describes how data was collected, generated, or processed. It is not only characterized by a large heterogeneity of publication types (dissertations, technical reports, data handbooks, working papers, methods reports, newsletters and bulletins etc.) but also by the question how research data is referenced therein. We analyzed grey literature in our repository PUB, in particular publications with supplementary material or linked to data sets. We found various types of supplementary material, among them research data, software and other kinds of research output that can be seen as discrete resources.

Aims

We propose certain measurements to improve publishing of research data and other research output linked to the underlying grey literature resulting in a better visibility and discovery of all research results. We see the following key requirements:

- **Citation:** authors should cite datasets in their underlying publications (enabling data location and validation by the reader) and vice versa (enabling the reader to understand context and methodology).
- **Awareness:** author guidelines for writing grey literature material should recommend the registration and deposit of research data if generated or used in the research process; design and implementation of corresponding publication workflows [2] need to be oriented towards intuitive and efficient user interfaces to minimize additional effort.
- **Reproducibility of Research:** publications should be complemented by any documents supporting interpretation and replication of the research data and helping to give insight into the resulting research findings.
- **Publication of Research Data:** research data (either created or re-used) that is needed to validate research results should be prepared for deposit and archiving.

State of the Art

Several funding bodies, as well as policy makers and research councils increasingly propagate that publicly funded research data should be openly available to the scientific community. The most German funders (like DFG[6]) and funders from the international landscape follow this guidance and encourage researchers to share and contextualize their data and research output. Through data disclosure policies formulated by journals (e.g. Nature, PLOS One) data, materials, codes, or scripts forming the basis for the respective publications have to be deposited within a repository at publication time.

In a meanwhile, policies imposed by funding agencies received wide recognition, but there is still a lack of institutional data policies. For example, guidelines on handling research data might be defined either at multiple administrative levels or they might not be anchored in the institutional policy at all. In addition, their liability might be misunderstood by researchers, or there might be a lack of technical and organizational support to fulfill them, e.g. by missing research data management services or library services supporting registration and

[6] DFG: German Research Foundation: **http://www.dfg.de/en/**

dissemination, and contextualization of research data. In Germany, the awareness of the professional handling on research data has already reached the universities. In May 2014, the General Meeting of the German Rectors' Conference (HRK) recommended the management of research data to be a strategic function of university management and calls the Universities upon to create the structural framework for efficient research data management for the whole institution.

Research Data as an integral part of grey literature

To illustrate the variety of publications with related or linked research results we picked some samples to showcase the current situation.

- *Enhanced Publication:* In 2011, a PhD student at TU Delft published his dissertation in the institutional repository and linked to the associated datasets published in the 3TU.Datacentrum[7].
- *Technical Report of the Collaborative Research Center 882 (SFB882)*[8]: The technical report itself contains parts which might be considered as research data (e.g. questionnaire, codebook in Appendix). The data is embedded within the document in a non-standardized way. Further metadata details on the registered data can be found as a DOI reference (here: http://dx.doi.org/10.4119/unibi/sfb882.2014.12), which links from the PDF to the data landing page.
- *Bielefeld Working Paper in "Economics and Management"*[9]: The landing page and metadata of the working paper relate to the source code and links to a stand-alone "research data set" http://doi.org/10.4119/unibi/2674041 which is in turn cited by other publications.
- *Software Publication:* Connecting version control systems, e.g. GitHub, with repositories, e.g. zenodo, allows for archiving of software snapshots or releases with persistent identifiers and thus makes code citable and put it into context with related publications and projects [3].

While publishing research data with the underlying publication by the authors themselves is a fundamental first step, services on top are needed which allow for aggregation, linking, knowledge extraction and discovery of those research artifacts across heterogeneous data sources. There are a number of initiatives that are striving to achieve such aims:

- *Data Literature Interlinking (DLI) service*[10]: The DLI service is a result of a collaboration between data centres, publishers, and research organizations that provides 'authoritative' links between datasets and underlying research literature.
- *InFOLiS I:* The aim of this DFG-funded project [8] was the development of techniques to discover links between publications and research data (in the Social Sciences) automatically and to retrospectively integrate them into the current retrieval systems.
- *OpenAIRE:* Aggregation and knowledge extraction from literature and data repositories are key components of the OpenAIRE scholarly communication infrastructure[11] that allows for contextualization of research results [9].

[7] **http://t1p.de/ep-dissertation-tudelft**
[8] **http://pub.uni-bielefeld.de/publication/2730392**
[9] **http://pub.uni-bielefeld.de/publication/2723277**
[10] **http://dliservice.research-infrastructures.eu/**
[11] **https://www.openaire.eu**

- *Research Data Switchboard:* By aggregation of links between datasets, publications and projects across different registries this service allows for discovery of datasets and related information[12].
- *DataCite Metadata Search*: This service[13] allows finding research data but also grey literature that has been minted with a DataCite DOI. In DataCite metadata relations to other resources can be stated, e.g. a dataset that relates with its underlying publication.
- *E-Science Funding Programme on the federal state level (Baden-Württemberg)*: The project bwDataDiss[14] develops an interdisciplinary infrastructure for describing, storing and linking research data with electronic thesis.

Standardized Relations in Metadata

The relation between the data and underlying publication needs to be manifested in metadata. One example is the DataCite metadata schema[15] that allows describing related resources of e.g. a research dataset using the element 'relatedIdentifier'.

```
<relatedIdentifier
  relatedIdentifierType="DOI"
  relationType="IsCitedBy">
  10.1234/fooBar
</relatedIdentifier>
```

In contrast, there is a variety of metadata standards to describe electronic theses and dissertations[16], making interoperability a challenging task. XMetaDissPlus is the metadata standard used in Germany to describe and transfer electronic dissertations [6]. Using Dublin Core elements the standard provides basic support for tagging relations to other resources that are part of a dissertation like research data.

```
<dc:relation xsi:type="dcterms:URI">http://dx.doi.org/10.1234/fooBar
</dc:relation>
<dcterms:requires xsi:type="dcterms:URI">http://dx.doi.org/10.1234/fooBar
</dcterms:requires>
```

Addressing Challenges

Grey literature often lacks standards for its creation, publication and distribution. Its target application and target audience may also vary, e.g. it may be shared only among a researcher group and never be published to the public. Grey literature is therefore often difficult to discover, access, and evaluate. Furthermore, it lacks of guidelines and good practices about research data linkage in grey literature. However, it offers opportunities for research libraries to contribute with services for research data management at institutional level [7]. In particular we encounter the following challenges:

- Data related to a publication (e.g. primary data, software products, or scripts used for data analysis) is mostly just mentioned in footnotes or bibliographies without a proper citation.

[12] **http://www.rd-switchboard.org/**
[13] **https://www.datacite.org/services/find-dataset.html**
[14] **http://www.alwr-bw.de/kooperationen/bwdatadiss (in German)**
[15] **http://schema.datacite.org/**
[16] **http://www.ndltd.org/standards/metadata**

Thus, this data is not searchable or accessible through public research data catalogues and academic search engines.

- Such discrete resources need to be registered and mutually linked with the grey literature document as shown in Figure 1.
- Authoring tools are needed that support guidelines for e.g. the production of scientific and technical reports [11] and support standards for citing data.
- Institutional policies – if they exist – often do not cover grey literature and research data as a regular component of the research output.
- Solutions for the Research Data Management (RDM) should be generic enough, that they can be applied to many disciplines.
- Libraries can take the role to promote awareness to researchers about a possible publication of research data and its sharing.

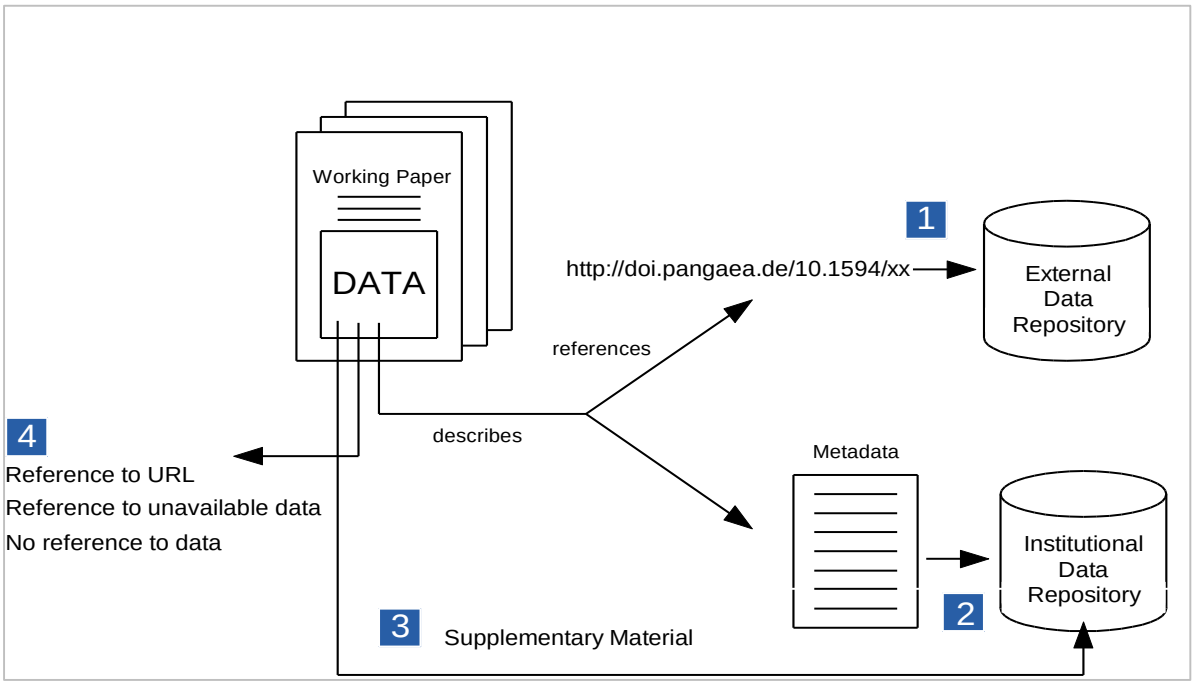

Figure 1 Use Cases for Linking Grey Literature with Data

Institutional Context with PUB

In 2009, the *Rektorat* of Bielefeld University launched the INFORMIUM[17] Initiative in order to identify the requirements for a university wide and generic research data infrastructure. As a result, Bielefeld University has passed guidelines and policies on research data management, being the first German university to form an institution-wide agreement upon standards for research data handling among all stakeholders. The university-wide policy calls on researchers

> (i) to take advantage of the university's advisory services for research data management and

[17] https://data.uni-bielefeld.de/en/informium

(ii) to publish research data through registered research data repositories[18]. To this end, Bielefeld University Library offers comprehensive advisory services on data management planning, data publication and preservation through its institutional repository PUB – Publications at Bielefeld University[19].

"PUB – Publications at Bielefeld University" is used to reflect the work of the university's researchers. Through easy-to-handle embedding possibilities, different views on the data (e.g publications of a single researcher, research group, department, or whole faculty) can be visualized as lists – optionally with faceted search – on the respective institution's websites. PUB is a hybrid institutional repository depositing and disseminating data and publications. The repository is compatible with OpenAIRE which supports and monitors the Open Access mandate in EU Horizon 2020.

Technically, PUB is based on the LibreCat[20] framework developed by the university libraries Gent, Lund and Bielefeld. Through its data processing routines for data oriented applications it facilitates the normalization of metadata and provides plugins for import and export. PUB is well integrated in the international scholarly communication infrastructure, e.g. by importing from large bibliographic databases and thematic repositories. It supports established machine interfaces (OAI-PMH, SRU, CQL) and metadata formats (Dublin Core, DataCite Metadata Kernel, MODS, XMetaDissPlus) to serve aggregative services like BASE[21], OpenAIRE, DataCite, EuropePMC[22] and DNB[23].

To ensure failure safety and reliability, PUB participates in the international distributed preservation repository network, SAFE Private LOCKSS Network[24], with the aim to preserve digital objects for future generations and to minimize the risk of data loss – caused by hardware breakdowns, obsolescence or natural disasters, or even human errors – over the long-term. The overall idea of SAFE-PLN is to make multiple copies (here: seven) as preservation strategy and to disseminate these copies throughout the world, in places considered to be safe. In the event of an unfortunate loss, data can be restored from one of the other preservation nodes, which all act in an autonomous and independent way at both financial and administrative level.

[18] **http://re3data.org**
[19] **http://www.pub.uni-bielefeld.de**
[20] **http://librecat.org**
[21] **http://base-search.net**
[22] **http://europepmc.org/LabsLink**
[23] **http://www.dnb.de/DE/Wir/Kooperation/dissonline/dissonline_node.html**
[24] **http://www.safepln.org**

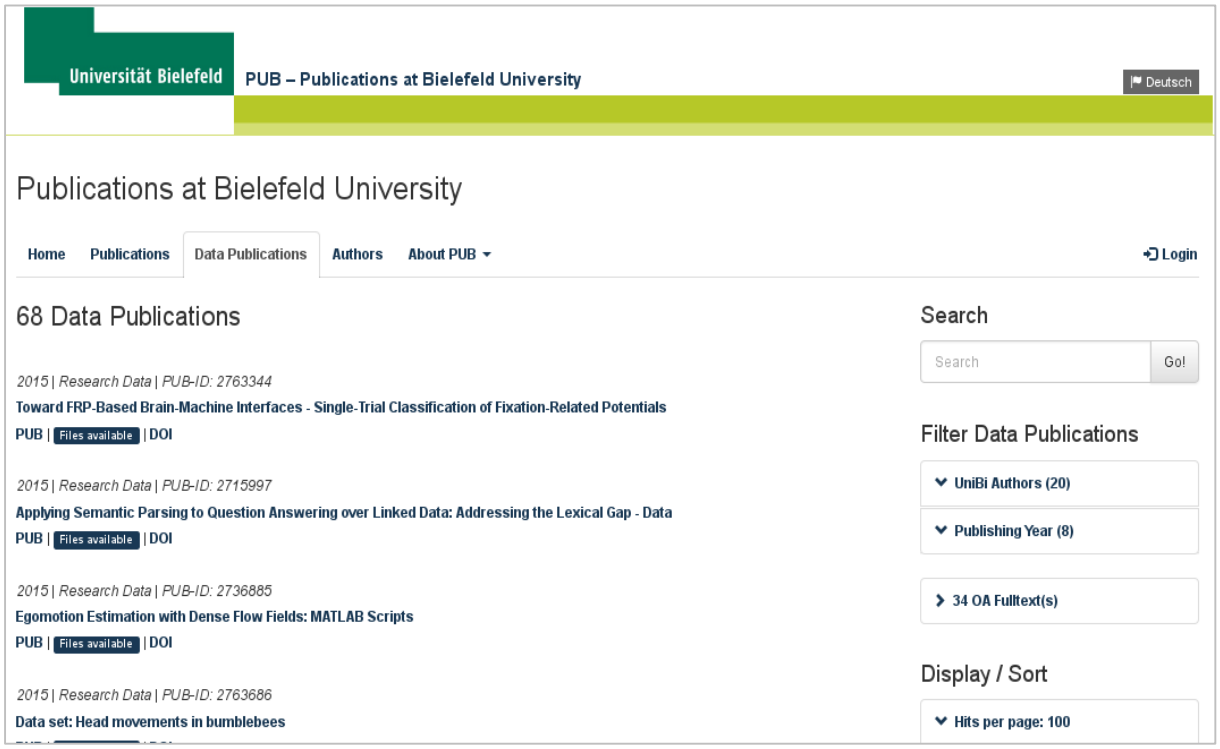

Figure 2 PUB Search Interface for Research Data

PUB has been extended by several aspects of data contextualization in the course of the introduction of the institutional research data policy[25] in 2013. As one measure Bielefeld University qualified as a publication agency for DataCite DOI. In PUB, the DOI registration of research data is part of the publication process. The persistent identification makes sure that the data stays available unchanged over time for later verification and re-use. Thus, the DOI can be used to cite the data in the manuscript. In general, the DOI resolves to a landing page in PUB, except for bilateral agreements made with research groups where the DOI may resolve to databases at their research institute. Research data in PUB have their own view as shown in Figure 2. They are part of more than 46,300 bibliographic references and publications in PUB. Thus, PUB provides increased access to a rich array of research output – among them grey literature and research data – as summarized in Table 1.

Quantity	Publication Type	With Suppl.Material
1670	thesis / dissertation	18
814	conference proceeding / paper	47
656	working paper	1
120	report	4
74	research data	
18	preprint	0

Table 1 Distribution of Open Access Grey Literature and Research Data in PUB

[25] https:ttps://data.uni-bielefeld.de/en/resolution

Some of these publications have supplementary material attached or data embedded in the publication, including research data (tabular data files, questionnaires and variables), software, algorithms, and multimedia files. This motivates for the design and implementation of a workflow for the discrete publication of literature, data and software.

Conceptual Design

In order to develop a common, discipline-independent model for interlinking grey literature with data, we have first analyzed concrete examples available in PUB to find out how research data is put into context. We found discipline-specific aspects based on differences in the particular research process and diversity in interpretation of the data life cycle. In addition, the maturity and the stages of how data citation principles are applied among disciplines are deeply divided. For example, in the Social Sciences, it is quite common to cite external data (e.g. census data), which have been used to respond the own research question. Conceptually, we consider the following cases in PUB:

A Classical supplementary material attached to the main manuscript: Any material (tables, figures, descriptions) that provides additional information and is published together with the main manuscript so that it depends on it and is not a discrete publication by itself.

B Stand-alone "research data": A dataset, software etc. that was generated, implemented, analyzed or otherwise used in the context of a research question and therefore constitutes a discrete resource. It benefits from referencing to a publication and can be cited in other works and gives credit to its authors or contributors.

C Links to external data: References to data either used to derive findings for the own research (e.g. any third party data) or data generated during the research (e.g. raw data, processed data, results, etc.). The data has been published or archived externally in a discipline specific repository.

In case researchers consider publishing their “own” data, the Library Services might support them in the decision making process. For example, questions about any restrictions placed on sharing the data (e.g. ethical, commercial, protection of personal data, intellectual property) are discussed.

Implementation

Advisory Support for Publication

Another aspect of supporting researchers in publishing their data within grey literature (e.g. PhD thesis) is to incorporate data publication workflows to the thesis submission process. After a successful thesis defense, researchers have to accomplish a publication of their thesis – as part of the graduation requirements – either in printed form (e.g. book) or online through the University’s Library publication services (pdf-file). Since the latter is the most frequently used choice, we are building on this point to sensitize the PhD candidates to publish their data attached to the printed thesis (e.g. on storage media) or enquire about the existence of data worth being published or reused within a broader community.

Thesis Submission Process

The PUB system supports a general user interface to define relations between resources of different type. A dissertation can be associated with both an already existing and registered research data in PUB and any kind of external resources, like deposited in a disciplinary repository. It is also possible to make plain references to software sources in version control systems like GitHub[26].

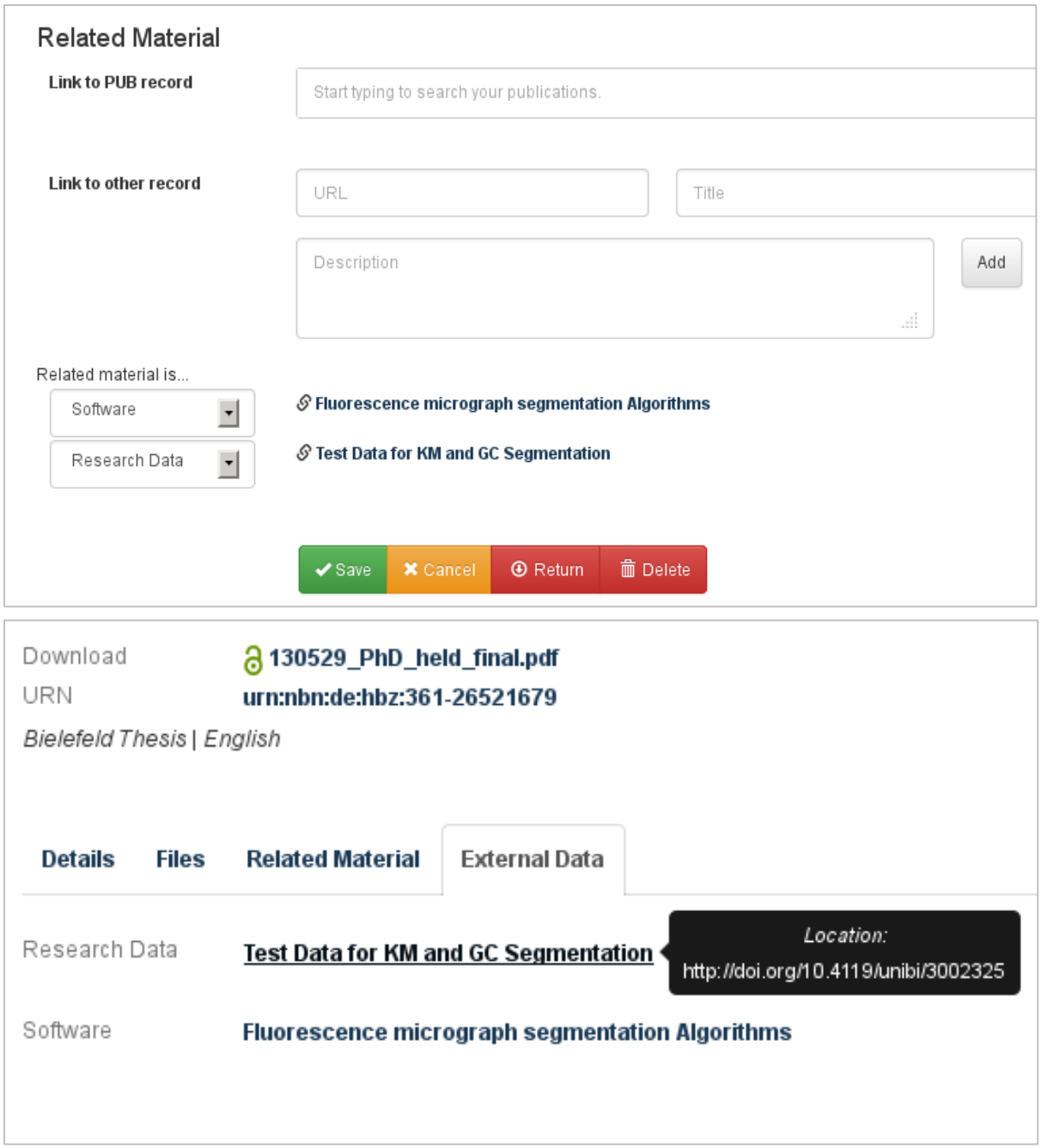

Figure 3 Submitting a PhD-Thesis with References to Research Data (a) and the corresponding view in the Landing Page (b)

Figure 3 (b) shows the respective landing page of the publication with the visualized interlinked data which is public to all users. Here, all of the interlinked resources are represented

[26] https://github.com

by hyperlinks where the target depends on the relation and resource type (see column "example" in Table 2).

	Relations	**Identifier**	**Resource type**	**Target example**
Related material	suppl. Materials	PUB-ID, FILE-ID	any	any uploaded file
Publications in PUB	is-part-of, earlier-version, cites, is-cited-by ...	PUB-ID	research data (as stand-alone publication)	PUB landing page
External Data	is-part-of, earlier-version, cites, is-cited-by, uses ...	DOI, URL	software, data sets	link to GitHub, Dryad, biological databases

Table 2 Possibilities of relating data to a PhD thesis in PUB

6. Conclusion and Future Work

In this paper we discussed challenges of linking research data with grey literature and presented an organizational and technical workflow that enables its publication, inter-linking and preservation through our institutional repository. By implementing and promoting institutional research data policy, technical infrastructure and organizational support to our researchers we increase their awareness of possibilities to publish and share data. In the context of the just started DFG funded CONQUAIRE project [5] interdisciplinary research teams, the CITEC Semantic Computing Group[27] and the University Library collaborate towards an infrastructure supporting analytical reproducibility of scientific data tightly integrated in the research process.

References

[1] BRINEY, Kristin, Abigail GOBEN, and Lisa ZILINSKI. Do You Have an Institutional Data Policy? A Review of the Current Landscape of Library Data Services and Institutional Data Policies. *Journal of Librarianship and Scholarly Communication* [online]. Pacific University Libraries, 2015, **3**(2). E-ISSN: 2162-3309 Available from: **http://jlsc-pub.org/articles/abstract/10.7710/2162-3309.1232/**.

[2] BLOOM, Theodora et al. *Workflows for Research Data Publishing: Models and Key Components (Submitted Version)*. 2015. DOI: **10.5281/zenodo.20308**.

[3] POTTER, M. and T. SMITH. Making code citable with Zenodo and GitHub. In: *Software Sustainibility Institute* [online]. 2015. Available from: **http://www.software.ac.uk/node/1720**.

[4] PIWOWAR H. A and T.J. VISION. Data reuse and the open data citation advantage. *PeerJ* [online]. PeerJ, 2013. ISSN: 2167-8359. Available from: **https://dx.doi.org/10.7717/peerj.175**.

[27] **http://www.sc.cit-ec.uni-bielefeld.de/**

[5] CIMIANO, P. et al. *CONQUAIRE: Continuous quality control for research data to ensure reproducibility: an institutional approach*. 2015. DOI: 10.5281/zenodo.31298. Project Proposal to Deutsche Forschungsgemeinschaft in the Programme.

[6] KOORDINIERUNGSSTELLE DISSONLINE (Ed.). *Referenzbeschreibung XMetaDissPlus v2.2*. Leipzig: Deutsche Nationalbibliothek, 2012. Available from: **http://nbn-resolving.de/urn:nbn:de:101-2012022107**.

[7] AYRIS, P., P. ACHARD and S. FDIDA, et al. *LERU Roadmap for Research Data.* LERU Advice Paper. Leuven: LERU, 2013, vol 14. LERU. Available from: **http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf**.

[8] BOLAND, K., D. RITZE, K. ECKERT and B. MATHIAK. Identifying references to datasets in publications. In: *Theory and Practice of Digital Libraries*. Berlin: Springer, 2012, p. 150-161. ISBN: 978-3-642-33289-0. ISBN 978-3-642-33290-6. DOI: 10.1007/978-3-642-33290-6.

[9] KOBOS, M., Ł. BOLIKOWSKI, M. HORST, P. MANGHI, N. MANOLA, & J. SCHWIRRWATEN. Information Inference in Scholarly Communication Infrastructures: The OpenAIREplus Project Experience. In: *Procedia Computer Science*. Elsevier, 2014, vol. 38, p. 92-99. DOI: 10.1016/j.procs.2014.10.016. ISSN: 1877-0509.

[10] HRK: GERMAN RECTORS' CONFERENCE. *Recommendation of the 16th General Meeting of the HRK, 13 May 2014: Management of research data – a key strategic challenge for university management*. Bonn: HRK, 2014. Available from: **http://www.hrk.de/uploads/tx_szconvention/HRK_Empfehlung_Forschungsdaten_13052014_EN.pdf**.

[11] DE CASTRO, P., & S. SALINETTI. Grey literature: challenges and responsibilities for authors and editors. In: *Science Editors' Handbook*. European Association of Science Editors, 2013. ISBN 978-0-905988-11-5. Available from: **http://www.ease.org.uk/sites/default/files/6-4.pdf**.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

OA TO SCIENTIFIC PUBLICATIONS AND RESEARCH DATA IN HORIZON 2020

Jana Kratěnová

kratenova@tc.cz

Technology Centre ASCR, Czech Republic

Abstract

Open Access to scientific information, which means to scientific publications and research data, is related to broader concept of Open Science. The aim of Open Science is to allow anybody, which means university researchers, business sector as well as society, to access some types of publicly funded R&I project results free of charge and on-line. EU framework programme for supporting research, development and innovation titled Horizon 2020 has taken a lead on this issue at the European level. Horizon 2020 grant beneficiaries are obliged to allow open access to their peer-review scientific articles. This obligation is in some case extended to underlying data as well as other non-published data.

Keywords

Open Access, Scientific Publications, Research Data

Introduction

Open access (”OA”) to scientific information through non-binding documents issued by the European Commission (“EC”), defined as the provision of free on-line access to scientific information to end users, is closely related to the obligation of the beneficiaries of Horizon 2020

grants ("H2020")[28] to disseminate the results of projects using suitable means as soon as it is possible, if this is not at odds with their legitimate interests. As ensues from the definition above, however, OA only concerns a specific category of project results, meaning scientific information, specifically scientific publications and research data, as explained below. The obligation to provide OA access to scientific publications only arises in the situation that the H2020 grant beneficiary chooses publication as the means of disseminating the project results, which is a question of setting plans in a successful project proposal. This changes nothing of the fact that any H2020 grant beneficiary that decides to disseminate project results in the form of scientific publications must do so within the regime of OA without the opportunity to extricate itself from this obligation.

The above therefore differs considerably from the FP7, in which OA to scientific publications only concerned certain beneficiaries of FP7 grants and was moreover regulated as a "best effort" obligation. Consequently, there was considerable room here for not publishing within the regime of OA without actually being in breach of the grant agreement.[29] The obligation to provide OA to research data that H2020 introduces for certain areas did not exist at all in the FP7. The subchapters below therefore look in detail at the rules of providing OA to scientific publications and to research data.

Open Access to scientific publication

As stated above, all H2020 grant beneficiaries must ensure OA to peer-reviewed scientific publications relating to their project results. Specifically-speaking, grant beneficiaries must ensure that the relevant publication can at least be read on-line, downloaded and printed out. As far as the term reviewed scientific publication itself is concerned, this assumes open access to reviewed scientific articles. Grant beneficiaries are nonetheless called upon to provide open access to other types of publication which relate to the project results, whether already having been reviewed or not; for example, books, reports and papers from conferences.

Open access to scientific publications is regulated by Article 29.2 of the model grant agreement, based on which the H2020 project is undertaken and from which it can be inferred how to achieve OA to scientific publications.

The first step in discharging this obligation is to deposit the reviewed scientific article in a repository in the form of a machine-readable, published version of the article or a version accepted for publication after review procedure ("post-print"). The publication in question must be deposite d not later than at the time of publication. The grant beneficiary is not limited in terms of the repository in which it may deposit the relevant publication, meaning that it might decide to use its own institutional repository. If the institution does not have its own repository, the scientific article may be deposited in a thematic or central repository. A list of different repositories can be found at, for example, **http://roar.eprints.org** (Registry of Open Access Repositories) or **http://www.opendoar.org/** (Directory of Open Access Repositories). The storage of reviewed scientific articles in a repository is tied to the obligation to make efforts

[28] Horizon 2020 is the largest and most significant funding programme in European-level science, research and innovation between 2014-2020; Horizon 2020 was preceded by the 7th Framework Program ("FP7"), which ran from 2007 to 2013.
[29] The grant agreement lays down the rights and obligations of the grant provider (the EC or its executive agencies) and the grant beneficiary, inter alia, the OA obligation, as described in detail in this paper.

to deposit the research data required to verify the results presented in the relevant publication ("underlying data"), ideally in a repository intended for research data.

The next step after storage is to make the relevant publication accessible, to "open" it. Access is provided either by auto-archiving, meaning "green OA", or by way of publication in OA journals, meaning "gold OA". In the second case, the article is made accessible at the moment of publication in such openly-accessible journals since the costs relating to publication in an open journal are covered by the author (or employer or grant provider) and the reader consequently has immediate, free access to the relevant article without any payment being required of him.

Eligible costs in H2020 project and gold OA publishing. The costs of publication of one article through gold OA might reach EUR 1,500 and are eligible costs in H2020 projects, if incurred within the period of duration of the grant. Should there be a publication issued after the H2020 project is over, the publication author cannot rely on grant budget to cover APC (Article Processing Charges – costs related to publishing in gold OA, see e.g. above mentioned 1,500 EUR). This situation is at least partly solved for FP7 projects by launching *FP7 post-grant Open Access publishing funds pilot (***https://www.openaire.eu/postgrantoapilot***).*

If a 7th FP grant beneficiary wants to publish project results after the end of the project, it can obtain the resources required for publication within the regime of gold OA from the above-mentioned fund. The costs of gold OA can be paid from the fund for a maximum of three scientific publications from one FP7 project on the condition that publication comes not later than within two years of the end of the project. It can be expected that similar instrument may be launched for H2020 grants provided the said pilot will be successful.

Embargo period in case of green OA including types of openly accessible publications.In the case of green OA, the relevant article can be made accessible not later than within a time limit of six months, or twelve months in the case of social sciences and humanities. This time limit (*embargo period*), i.e. the delay between storing an article in a repository and this article becoming accessible, might prove to be a complication for the grant beneficiary. The grant beneficiary is consequently bound on the one hand by the terms and conditions of the grant agreement, which lays down an *embargo period* of 6/12 months, and is frequently bound on the other by the terms and conditions of a licence contract signed with the publisher of a "paid/commercial/not OA" journal (a journal that is traditionally based on obtaining money from payments by readers for access to the articles published there and consequently, in contrast to open/gold OA journals, does not offer readers free access to the article), which might allow it free access to an article published in a paid journal after the passing of another embargo period, frequently longer than 6/12 months. Another possible problem might lie in which version of the relevant article the publisher allows the grant beneficiary to make accessible in the repository – as specified above, the model grant agreement requires the storage and accessibility of the post-print, the final, published scientific article. In the contract, the publisher might only allow the grant beneficiary to deposit the pre-print version of the article in the repository and make it accessible, meaning the version before review procedure. This, however, does not comply with the conditions laid down in the grant agreement and would represent failure to discharge the obligation to provide OA to reviewed scientific articles. It is therefore up to the grant beneficiary to monitor accord between the licencing conditions of the publisher and the terms and conditions of its grant agreement. The SherpaRomeo (**http://www.sherpa.ac.uk/romeo/**) website can be used to this end; i.e.

to find a clearly-arranged description of the licensing policies of different publishers and compare them with the terms and conditions of the model grant agreement.

The final step in complying with the OA obligation is storage and ensuring open access to bibliographic metadata, the aim of which is to allow and simplify searching for and identifying an article concerning the results of an H2020 grant and to monitor, generate statistics and evaluate the impact of H2020. Article 29.2.c) of the model grant agreement describes in detail the sort of metadata that is to be deposited and made accessible. What is more, it must be remembered that the obligation to deposit such articles in a repository with the aim of their long-term storage remains for both green as well as gold OA.

The technical level of OA to scientific publications described above is accompanied by the legal level, which consists of a suitably chosen licensing tool for making scientific publications accessible so that the obligation to provide OA is discharged; i.e. the opportunity to at least read a publication free online, download it and print it out. To this end the EC makes reference to the existence of Creative Commons (**http://creativecommons.org/licenses/**) licences and mainly suggests considering the use of CC:BY or CC:0 licences.

Open access to research data

Open access to research data is new in comparison with the FP7. This is also one of the reasons why the obligation to provide open access to research data emanating from an H2020 grant does not apply to all H2020 grant beneficiaries and currently takes the form of “pilot projects”. Of course, the rules of Horizon 2020 allow those H2020 grant beneficiaries that do not have the obligation laid down in the work programme to provide open access to the research data created to voluntarily take part in pilot projects regarding OA to research data. The aim of these pilots is to allow everyone free access to digital research data according to the terms and conditions laid down in the model grant agreement, to subsequently exploit, mine, reproduce it and disseminate it. The guide to providing OA to scientific publications and research data in Horizon 2020 defines research data as “information, specifically facts or numbers, collected with the aim of examination. For example statistics, the results of measurements, observation or experiments within the context of research. Research data should be accessible in digital format”.

Providing OA to research data concerns two types of research data – the data required to verify the results presented in a scientific publication (Article 29.3.a.i. of the model grant agreement) and other data (Article 29.3.a.ii. of the model grant agreement). Article 29.3 of the model grant agreement sets out detailed procedure for discharging the obligation to provide OA to research data.

Steps to fulfil OA obligation. The first step is to deposit research data in a repository designated for research data. As in the case of providing OA to scientific publications, it stands here that it is up to the H2020 grant beneficiary which specific repository it chooses. It might therefore decide in favour of its own institutional repository or a thematic or central repository. A useful list of usable repositories can be found at, for example, **www.re3data.org**. The second step is for the H2020 grant beneficiary to take measures to ensure that third parties have the possibility to access the deposited research data without charge and mine, exploit, reproduce it and disseminate it as soon as this is possible. This is the legal level of the provision

of OA to research data. Even though the EC again suggests considering the use of a Creative Commons licence to ensure OA to research data, specifically the use of a CC:BY or CC:0 licence, and the fourth version of Creative Commons now allows databases to be licensed, it is important to point to the fact that CC:0, for example, need not be without its problems within the context of national legislation and that these issues should not be underestimated. Moreover, the grant beneficiary has the obligation to provide information via the repository about the tools and instruments which it has at its disposal and which are required to verify results. However, the grant beneficiary is not obliged to provide these tools and instruments to the end user. The model grant agreement merely states that it should provide them, if this is possible.

Exceptions to OA obligation. The obligation to provide OA to research data is unlimited and the model grant agreement and guide to providing OA to scientific publications and research data specify cases in which one can be released from this obligation. If the obligation to provide OA to research data is in conflict with the obligation to ensure protection for project results which it is supposed can be used for commercial purposes or is in conflict with the obligation to maintain confidentiality of information or is in conflict with the rules to concern personal data protection, the H2020 grant beneficiary need not provide OA to the relevant research data. The grant beneficiary may also be released from this obligation in the situation in which the provision of OA to research data could jeopardise achievement of planned project objectives, if no research data is generated or collected in the project or if another legitimate reason exists. Nonetheless, grant beneficiaries should count on the fact that the use of this "opt-out" option must be preceded by justification and reference to the actual status of project preparation/implementation. There consequently exists the possibility of being released from this obligation, if the above conditions are met, at the stage of the preparation of a project proposal or at any time during its implementation.

Data Management Plan. Grant beneficiaries will also come across a new document termed the Data Management Plan ("DMP") in connection with providing OA to research data. A DMP will become compulsory output in H2020 projects to which OA to research data relates, if there has been no release from this obligation, and for projects voluntarily participating in the OA to research data pilot. A DMP can be compiled voluntarily in other projects in which data is generated/collected. It should describe how the generated/collected research data will be handled during and after the H2020 project, which methodology and standards will be used with the aim of disseminating and providing access to this data and how its long-term storage will be ensured. Details regarding the DMP are regulated in the guide to data management in Horizon 2020. A DMP is not a static document; on the contrary, it is expected that it will be updated regularly in connection with the data generated/collected during project implementation. If the grant beneficiary decides not to take part in the OA to research data pilot during the project, it should justify its decision in the DMP. In contrast to the DMP, research data management is a process which concerns projects in which data is generated/collected without such projects voluntarily or mandatorily participating in the OA to research data pilot. If relevant to the project, a description of research data management should be part of the project proposal in the "*Impact*" section and is evaluated together with the subsequent description of the planned impact of the project.

It stands here too that the costs invested in connection with providing access to research data throughout the duration of the project are eligible costs. Moreover, the EC plans to provide technical and methodical support services in this regard.

Conclusion

As is clear from a detailed explanation and description of the rules of OA in H2020, the European Commission has chosen the policy of OA as one of the key areas of the H2020 programme, with other activities at an EU level adapting themselves to this – for example, establishing national reference contacts in all EU countries with the aim of supporting and coordinating the creation of OA policies; the initiatives of the European Research Area Committee (ERAC) relating to monitoring and implementing OA at national levels; financing the series of projects in the FP7 and H2020 relating to OA, etc. It will be interesting to see how the EC evaluates observation of the OA rules of H2020 in the future, most likely in its report on evaluation of H2020 issued after the first half of H2020, and whether the programme that follows on from 2020 chooses a strict obligation to provide open access to research data without exception.

References

EUROPEAN COMMISSION. Directorate - General for Research & Innovation. *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020* [online]. European Commission, 2015 [cit. 2015-12-07]. Available from: **http://ec.europa.eu/research/participants/data/ref/h2020/mga/gga/h2020-mga-gga-multi_en.pdf**.

EUROPEAN COMMISSION. *H2020 General Model Grant Agreement — Mono: (H2020 General MGA — Mono)* [online]. European Commission, 2015 [cit. 2015-12-07]. Available from: **http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf**.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

PUBLICATION OF RESEARCH RESULTS: SELECTED LEGAL ASPECTS

Matěj Myška*

matej.myska@law.muni.cz

Faculty of Law, Masaryk University, Brno, Czech Republic

Abstract

This contribution focuses on the legal aspects of raw grey literature that is understood as yet to be published, interim or incomplete scientific results (e.g. raw primary data or drafts of scientific papers). Specifically the contribution deals with the limits of further re-use of such scientific results, when these were offered for publication in traditional publishing houses, and when these results should available under Open Access conditions.

Keywords

Copyright, Database Rights, Grey Literature, Open Scientific Data, Open Access

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

* Senior assistant professor, Institute of Law and Technology, Faculty of Law, Masaryk University
I thank JUDr. Jakub Harašta for critical revision of this paper. However, all mistakes are mine.

Introduction: A new scientific revolution

In 2006 Banks (Banks 2006) drew attention to erasing distinctions between standard published literature and grey literature. The significance of this finding has grown increasingly popular with the movement toward open access to the results of publicly funded research and development. [30] One of the ways to implement it lies precisely in the publication of "borderline" forms of scientific outputs - the so called "e-prints", i.e. electronic versions of scientific papers in various phases of publication process. The trend of Open Access also applies to the research data, [31] which the published scientific papers are based on. Only such a comprehensive disclosure can fulfil one of the basic ideals of scientific knowledge, namely its reproducibility. [32] This change in communicating research results significantly shatters the established practices. Achenbach talks directly about a "new scientific revolution"(Achenbach 2015).

For the area of grey literature and its exploration this change has a major impact on its very subject. [33] The traditional "New York" definition of grey literature is seen as "information produced on all levels of government, academic, business and industrial institutions, both electronically and in paper form, which have not undergone the standard publishing process, and which are not distributed in the standard sales network, i.e. they are issued by institutions whose main activity is not publishing" (Schöpfel 2011, p. 5).[34] In the case of the implementation of the green road to open access (Open Access)[35] and open research data[36] the traditional white and grey literature then blend together as they are published both officially, as well as "unofficially" in institutional or disciplinary repositories (auto-archiving), or repositories of grey literature.

To the complicated legal issues that need to be addressed in the case of making grey literature accessible, [37] others are added regarding the standard publishing process and the implementation of Open Access, [38] and new ones that relate to the interconnection of these two modes of publication are generated. Specifically, this brief paper deals with the possibilities and limits of dealing with such scientific results in a situation when they were offered for publication. In the first part the issue of the property rights to such results and the possibility of dealing with them are addressed, that is how to implement auto-archiving, without interference with the rights of the holder. Attention is then paid to the individual phases (stages) of the scientific paper publishing process and the related contractual arrangements in the publisher's license agreement. The theoretical description of the issue is complemented by

[30] A typical example of the trend towards openness is the Framework Programme for Research and Innovation Horizon 2020, which mandates Open Access for all supported results in the form of scientific papers. In selected calls Open Access to research data is then tested in a **pilot program**. In details see (European Commission 2014).
[31]The European Commission officially adhered to this trend in its Communication "Towards better access to scientific information: Boosting the benefits of public investments in research", Brussels, 17.7.2012 COM (2012) 401 final.
[32]Karl Popper stated that "*a random unique occurrence of phenomena has no meaning for science.*" (Popper 2002, s. 66)
[33] The Open Access movement is a challenge to the very understanding of science and its role in society and the manner of its communication to the public. Such considerations, however, go beyond the intent and scope of this paper. However, readers can be recommended the basic work of the "spiritual father" of this movement Peter Suber (Suber 2012). Copyright issues in Open Access are dealt with by Marc Scheufen (Scheufen 2015). The legal aspects of "opening scientific data" were comprehensively elaborated on by Guibault and Wiebe (Guibault and Wiebe 2013). I then solved selected aspects of Open Access in (Myška 2014).
[34] Translation according to: NRGL Definition of grey literature. Available from: **http://nusl.techlib.cz/index.php/Definice**.
[35]In the details of the green road to Open Access see (Suber 2012, s. 52–58).
[36]The research data was then called grey literature by e.g. Banks (Banks 2006, s. 9).
[37]In the Czech Republic, they were dealt with by e.g.: (Polčák a Šavelka 2009),(Polčák 2010).
[38]On the issue of the implementation of Open Access in the context of an effective copyright see (Scheufen 2015).

a practical analysis of the relevant agreement of the largest publisher Elsevier, which publishes more than 2,000 scientific journals.[39]

The results of research and development and the rights to them

The results of research and development are understood in the context of this paper as professional scientific papers designated for publication in professional scientific journals and the research data on which these papers are based. These are then protected by specific absolute rights - namely copyright and the sui rights of the database maker.[40]

Assessing whether the legal requirements for obtaining copyright protection for a scientific paper (or for its "developmental phase") were met is still a matter of national regulation since harmonization still has not occurred in this field.[41] In the Czech Republic the fulfilment of the character of a copyrighted work will then be assessed, i.e. whether pursuant to Section 2 para. 1 of the Copyright Act [42] it is a scientific work, a unique result of creative work and is expressed in an objectively perceivable manner. The provision of Sec. 2 para. 3 of the Copyright Act under which copyright applies to *"a completed work, its different developmental stages and parts"* is important. As stated by Telec a Tůma (Telec a Tůma 2007, p. 45), this protection applies to individual parts of a work, regardless of whether they may be *"used separately outside of the work as a whole or not."* However, these individual parts must separately satisfy the above-mentioned essential characteristics of the copyrighted work. In terms of further discussion, it is necessary to mention the extent of the rights that are granted to the author with a view to the publishing practice of publishers. In the context of the implementation of open access the basic property right is that of the author to use the work, in both the original and in other adapted forms, as well as the personal rights to the inviolability of the work. Another issue, whose solution, however, goes far beyond the scope of this paper is who, when and under what circumstances, is entitled to exercise these rights. [43] It must be particularly taken into account, whether one author or more authored the concerned work. It is also necessary to solve what kind of work it is[44] and whether the rights are contractually modified.[45] For example, in the case of employee's work a different regime from the statutory one can be negotiated or e.g. the employer may leave the employee the exercise of the rights by an internal directive, or a can representation of the employer (copyright holder) by the employee can be constructed. Finally, copyright protection is granted when the work is expressed in an objectively perceivable form (Sec. 9 of the Copyright Act). It is therefore

[39] Elsevier. At a Glance. Available from: **https://www.elsevier.com/about/at-a-glance**.

[40] In the details on the protection of scientific databases by intellectual property rights, see(Rieger 2010).

[41] As noted by Husovec, however, the European Court of Justice has already more or less harmonized the concept of copyright works in its decision-making (Husovec 2012).

[42] Act No. 121/2000 Coll., Copyright Act, as amended, hereinafter referred to as the "Copyright Act".

[43] It is the question of who is entitled to implement Open Access which is one of the most complex, especially because of the distinct national regulation and the plurality of entities. For details on the Czech legal environment see (Myška 2014, p. 613–614). A detailed comparative analysis is submitted by (Guibault 2011, p. 140–151)**.** In order to simplify the following analysis, we will assume that the author is only one natural person who is authorized to dispose of the copyrighted work - a scientific paper. Likewise, in the case of the existence of protective database regimes**,** we assume that these rights also indicate the author.

[44] Employee (Sec. 58 of the Copyright Act), collective (Sec. 59 of the Copyright Act), school (Sec. 60 of the Copyright Act) or upon request (Sec. 61 of the Copyright Act).

[45] In the details on these special types of works and deviation from the standard treatment of the handling of copyrighted works see the comments to relevant sections in (Telec a Tůma 2007).

irrelevant whether there has been a publication or not. This means that both the grey and white "version" of scientific papers are copyright protected.

The conditions for granting protection for a collection of research data (a database)[46] have already been harmonized on the EU level.[47] Directive 96/9/EC on the legal protection of databases (hereinafter referred to as the "Directive") lays down the conditions for the protection of databases by copyright and the sui generis right of the database maker. Copyright protection[48], according to Art. 3 para. 1 of the Directive applies to databases that "by the selection or arrangement of their contents constitute the author's own intellectual creation". The national legislator can then implement protection for the so-called "creative databases" that are for example protected in the Czech Republic as a collection in accordance with Sec. 2 para. 5 of the Copyright Act. In the case of research data, however, these modes will rarely be applied.[49] The selection and arrangement of content will be determined by "technical factors and or imperatives of accuracy and exhaustiveness." (Guibault a Wiebe 2013, p. 21) and thus ineligible for protection. Unoriginal databases can be protected by special rights (sui generis rights) pursuant to Art. 7, para. 1 of the Directive (Sections 88-94 of the Copyright Act). Requirement for obtaining the protection is a qualitatively or quantitatively substantial investment in the acquisition, verification or presentation of their contents. The Court of Justice of the European Union has clearly stated in its judgments that the protection does not arise when this investment (i.e. cost) relates only to the creation of elements (content) of the database.[50] Guibault and Wiebe rightly inquire whether it is actually possible to protect a database of scientific data with such rights, as in the case of scientific data and its collection into databases, most investments are linked precisely with its creation. This criterion is therefore quite uncertain and must be considered on a case-by-case basis. (Guibault a Wiebe 2013, p. 26). The findings of this ad hoc assessment are quite essential since the actual research data, not protected by any of the above-mentioned exclusive rights, is then not protected at all (cf. Sec. 2 para. 6 of the Copyright Act). In the context of the CJEU decision in the Ryanair case,[51] it should however be noted that the absence of any such protection does not mean that it would not be possible to regulate and limit the handling of databases contractually. As in the section dealing with copyright it should be noted that the rate of access to a database or the form of its publication, i.e. whether it is white or grey "literature" are not relevant for the granting of protection.

From grey to white

Grey and white literature intersect in a problematic manner precisely at the moment when the relevant right holder decides to publish the given result by one of the traditional publishers, or to "officially" publish the research data underlying such an outcome. Other areas of friction then arise when he or she would like to make such scientific results available under Open

[46] A database is then under Art. 3 of this Directive, '*a collection of works, data or other independent elements which are systematically or methodically arranged and individually accessible by electronic or other means. "*

[47]On databases in general (Derclaye 2008; Herr 2008; Connelly Kohutová 2013).

[48] Paper. 3 to 6 of the Directive.

[49]This statement, however, does not apply categorically in the area of qualitative research, where in the case of questionnaire surveys it would be possible to consider the protection of databases by copyright. I thank JUDr. Jakub Harašta for this idea.

[50] Judgments C-203/02 BHB point 31, 32, C-444/02 point 41, C-46/02, point 41 and C-338/02 Svenska Spelbod 24, 25. In the details of these regulations (Adamová 2011), (Davison a Hugenholtz 2005).

[51] The ruling of the Court of Justice of 15 January 2015. Ryanair Ltd v. PR Aviation BV, Case C-30 / 14.

Access conditions. Attention will be paid first to the possibility of implementing Open Access to copyrighted works (scientific papers) and subsequently to the set of research data.

The pre-condition of publication is usually the conclusion of a publishing (license) agreement. Such contract sets out, in accordance with the focus of the relevant publishers, different rules regarding the handling of scientific output - a paper protected by copyright - as well as the possibility for achieving open access for it. By default, these contracts require the maximum possible transfer[52] of rights to the publishers. The publishers then justify this process by the argument that they need to have legal certainty regarding the acquired rights so that they can safely realize the required substantial investment in the distribution of the relevant scientific result (Guibault 2011, p. 148). One cannot ignore the interpretive principle of the limitation of the scope of the license according to its purpose ("Zweckübertragungsgrundsatz"). The German courts have already stated that it also applies in the case of such "maximum" unlimited licenses, if it does not correspond to their actual purpose (Telec a Tůma 2007, p. 519).

The implementation of Open Access is contractually characterized by leaving the possibility to exercise certain property rights to the author. These allowed uses are then referred to as "Allowed uses", or "Retained Rights". The SHERPA/RoMEO portal provides general, but at the same time clear information on the allowed uses. Individual publishers[53] are then differentiated depending on which version of a scientific paper the author can use for realizing Open Access. The terms to describe each phase is then based on the traditional publishing process, which Guibault described in brief as follows (Guibault 2011, p. 149). The author sends the Editorial Board the manuscript in the final wording (the so-called "Last hand" version) (Telec a Tůma 2007, p. 482), which is known as "pre-print" or the "submitted version." [54] Subsequently, the manuscript is sent to the review process (peer-review). The comments of reviewers are then incorporated by the author and if the amended paper is accepted for publication, it is called "post-print" or also the "accepted version." [55] The final version after editing of the publisher and typesetting is then called the "publisher's edition" or the "final published version." Publishers who allow the auto-archiving of a paper in the last two stages are marked in green. If the publishing contract enables only the deposition of a version before the review process, such publishers are marked in yellow. White colour refers to publishers who are not in favor of Open Access and do not allow any form of auto-archiving. The exercise of retained rights may be bound by a time clause, i.e. the expiry of a certain period – the embargo. [56] This form of open access is then called "delayed open access".[57] The time of the acceptance of a paper for publication is then normally set as the moment of the effectiveness of an agreement, and therefore the application of the above possible restrictions. If a paper is

[52] Namely both translative and constitutive, if the translative transfer (transfer, assignment of rights) is not permitted, as e.g. under Czech copyright law (Sec. 26 of the Copyright Act).

[53] Resp. individual journals as it is not uncommon for the publishing policies to differ within the publishing house for the individual journals.

[54] It is this version of a scientific paper that is traditionally referred to as grey literature. See the definition of grey literature above.

[55]As noted by Scheufen (Scheufen 2015, s. 154) the only adequate substitute for a publishing version of an paper is solely the post-print version, which includes and incorporates the comments of reviewers and any suggested modifications. A paper without publication approval and thus the review process is not an adequate scientific output, but only self-publication ("vanity press").

[56] Typically six or twelve months.

[57]Another differentiation and a detailed insight into the issues is offered by (Guédon 2004).

not accepted for publication, as a result of the review process, the author can store grey literature in the repositories without risk because no publishing contract was concluded.

The crucial question for discussion in relation to grey literature and its publication (or deposition in repositories of grey literature) is the position of the pre-prints, which they are traditionally referred to as grey literature. Certain sources do in fact claim (ANON. undated) that the handling of pre-print is fully in the power of the author, even if a publishing contract is concluded and this issue is not specifically addressed when he or she concludes a publishing agreement. On this issue, however, I share the dissenting opinion of Carroll (Carroll 2006). During the whole process of traditional publishing, a scientific paper is in fact an identical work in all its stages in my opinion. Subsequent minor revisions made by the author are not fundamental enough to lead to the emergence of a new copyright protection on the basis of which the author could use the work other than as set out in the publishing agreement. For example the author cannot thereby grant any Creative Commons public licenses for pre-print, [58] if he or she has already transferred (whether constitutively or translatively) the rights to the publisher. This conclusion, in my opinion, remains unchanged notwithstanding the moment when the auto-archiving has been implemented, i.e. whether before or after the conclusion of a publishing agreement. The consequence of this assertion, among others, is also that the author must withdraw the pre-print version of the paper from the repositories of grey literature if the final paper is accepted for publication in the journal, whose policy does not allow the implementation of Open Access. By failing their obligation to withdraw the pre-print they would not meet the obligations of their agreement. Any version could then only be used within the applicable exceptions and limitations to copyright pursuant to the applicable law, such as a gratuitous legal license for citation. Such cases will not happen regularly, because the authors usually retain these rights.

In the case of research data we are still missing an adequate regulation how to deal with them. A well worked-out methodology as is applicable in publishing does not exist. The reason for this state is that traditional commercial publishers do not consider (for the time being) the "publication" of data as part of the publishing process. The regulation of the disposition with databases protected by any of the above-discussed rights will then be governed by specific contractual arrangements. By analogy, it is possible to apply the same basic conclusions as for scientific papers - if the copyright or sui generis right to the database is transferred (constitutively or translatively) to the publisher, the author of the database / its maker must not use, extract or re-use the database apart from the cases permitted by law (i.e. within the scope exceptions and limitations). A fairly interesting question to debate is whether a new right of the database maker may emerge. This would be the case when the author of a scientific paper transfers raw primary research data to the publisher, for which the publisher would subsequently make a substantial contribution in its verification or presentation. At that moment, the sui generis right of the database maker would be granted to the publisher, and the original author of the paper could not later claim any such rights to the database.

The contractual practice of the publisher Elsevier

The above theoretically discussed publishing agreement is almost always concluded as a contract of adhesion, in layman's terms under the motto "take-it-or-leave-it". The author

[58]For details and discussion of other errors when using public licenses see (Koščík and Šavelka 2013).

is the weaker contracting party and does not usually have the possibility to negotiate the individual arrangements. The purpose of the following section is to introduce contractual possibilities how to dispose with scientific results in terms of their opening and eventual storage in repositories if the result is published by the Elsevier publishing house. Attention is only paid to the free-of-charge implementation of the Open Access form of self-archiving, the so-called "green way". [59]

The publishing house Elsevier in its standard publishing agreement [60] requires a full transfer of rights (translative) [61] for any kind of use and for the duration of copyright. The author retains the right to use the work for personal use, internal institutional use and for scholarly sharing, no matter the phases of the paper.[62] At first glance, it is a very liberal policy as regards the implementation of Open Access but a closer look reveals significant limitations. Firstly, in the case of post-print and the final version, the retained right never include commercial use, which is understood as use for commercial gain (e.g. use in advertising) and use that directly substitutes the services of the publisher (e.g. distribution by e-mail).[63]

In the first case - personal use - the author can use the paper (versions thereof) in the classroom, distribute copies to colleagues for their personal use (including e-mail), include it in a compilation of the works of the author, include it in his or her thesis or expand it into book-length form.[64] Internal institutional use covers the use in the classroom and for internal needs, including reproduction distribution and communication to the public, as well as within e-learning platforms (but not the so-called Massive Open Online Courses) and the inclusion of a paper in grant applications. Scholarly sharing in the case of pre-print means sharing (communication to the public) on any sites and in repositories. After the can acceptance of the paper the author can (resp. it is recommended to) add the DOI and a link to the final version. In the case of post-prints Elsevier applies the delayed Open Access for scholarly sharing. Before the expiration of the embargo period the paper can be shared non-commercial on one's blog or homepage, in the framework of the institutional repository of his or her institution for internal institutional use or upon individual invitation. The paper can be further directly disseminated and reproduced to students and academic colleagues for their personal use; and for private scholarly sharing upon individual invitation to a commercial site, with which Elsevier has an agreement. After the expiration of the embargo the paper can also be shared publicly in the institutional repository and also within commercial sites with which Elsevier has the respective agreement. The final version of the paper can then be academically shared only by a reference to the DOI.

The fact Elsevier expressly recommends that the post-print versions should be licensed under the Creative Commons public licenses in the variant, which excludes commercial use, and modification of the work (i.e. CC BY-NC-ND) ads another level of complexity to the issues

[59] The attention is not paid to the various alternatives of the traditional publishers to gold journals (e.g. SpringerOpen), or hybrid models, where the traditional non-open journal offers the opportunity to "buy out" open access for specific papers (e.g. Springer Open Choice).

[60] Sample "Journal Publishing Agreement" available from: **http://www.elsevier.com/__data/assets/pdf_file/0006/98619/Sample-P-copyright.pdf**.

[61] Under the Czech Law it could thus constitute a gratuitous exclusive license, with the right to sublicense, for all types of use and for the duration of copyright at most.

[62] Elsevier then specifically uses the terms "Preprint", "Accepted Manuscript 'and' Published Journal Paper."

[63]However, the paper can be expanded into a book or included in the compilation of works of the author.

[64]The author may also part of his or her work in other works - but this is covered by the above-mentioned citation licenses (here in concreto according to Sec. 31 para. 1 letter a) and b) of the Copyright Act).

examined.[65] This can lead to rather paradoxical consequences. If, for example an author publishes his or her paper under this public license still in the embargo period on his or her blog, the selected license entitles his or her colleagues to deposition of the work in any institutional or departmental repository. Thus, if the author does not want to violate his or her contractual obligations, he or she should never publish the work in such a way, even if Elsevier recommends it.

For the experimental primary data the publishing agreement uses the term "Supplemental material". If in addition to this the author of the paper uses the services offered by the publishing house, he or she also grants the publisher a nonexclusive license to publish, post, reformat, index, archive, make available, change format and link to such data including the right to sublicense it to this extent. The author, however, may store the data in one of the repositories of grey literature. Somewhat cryptically the publishing agreement stipulates that the above permission to use the data also applies to it even if it is made public by a reference in the paper. It is therefore evident that the agreement does not address in detail issues related to other property rights. The publishing agreement shall be thus interpreted in accordance with the above principle of the limitation of the scope of the license. The purpose of such contract is to entitle the publisher with the right to use the data in the above-mentioned foreseen manner. If the data collections handed over in this way are protected by any of the above-mentioned modes of protection the corresponding license would be also be granted to them (Telec a Tůma 2007, s. 519). Finally, in the case that the paper is not published the rights transferred to the publisher under the agreement will revert back to the author including any rights to supplementary material. It should be noted in conclusion that the publishing house Elsevier is generally regarded as a "green" publisher, however the respective journals have specific individual Open Access policies.

Conclusion

The gradual convergence of grey and white literature, as anticipated by Banks (Banks 2006), and the gradual move towards the openness of science and research brings with it new legal challenges. As this paper attempted to demonstrate in brief, the fundamental problem is the plurality of modes of protection of the results, as well as various contractual practices of the respective publishers and the complexity of the issue of the implementation of Open Access. The quest for an intuitive solution of the legal problems, as demonstrated on the issue of auto-archiving of pre-prints, can quite easily lead to a breach of the relevant publishing agreement and liability for damages caused. In the case of scientific papers there is a quite progressive proposal on [66] how to overcome these problems by introducing the inalienable right to secondary publication, which, without any contractual arrangements, would ensure the author the ex lege right to auto-archive the post-print version.[67] For effective functioning such an exception would have to have a mandatory character for all EU members. Such an effective solution could then be used also for publishing of research data in the event that the publishing agreement also dealt with any transfer of rights to the compilation of research data. Without further arrangements, the question of the publication of research data in particular remains a subject matter for contractual regulation (with respect to the CJEU judgment in the Ryanair

[65]This is however not an obligation – the verb "should" is used and this indicates a suggestion.

[66] The relatively extreme solutions such as the abolition of copyright protection for scientific works in General, as proposed by (Shavell 2010) are left aside.

[67]In the details of this law, see the resources in the footnote. fn. 7 in (Scheufen 2015, p. 144).

case). A question for further research and discussion is to what extent such a new right (exception) would further blur the differences between grey and white literature.

References

ADAMOVÁ, Zuzana, 2011. Rozpisy futbalových zápasov a konských dostihov a právna ochrana databáz. *Revue pro právo a technologie*. **2**(3): 20–22. ISSN 1804-5383. ISSN 1805-2797.

ACHENBACH, Joel, 2015. The new scientific revolution: Reproducibility at last. *The Washington Post* [online]. [cit. 2015-09-30]. ISSN 0190-8286. Available from: **http://www.washingtonpost.com/national/health-science/the-new-scientific-revolution-reproducibility-at-last/2015/01/27/ed5f2076-9546-11e4-927a-4fa2638cd1b0_story.html**.

ANON, undated. Self-Archiving FAQ. *eprints* [online] [cit. 2015-09-30]. Available from: **http://www.eprints.org/openaccess/self-faq/#publisher-forbids**.

BANKS, Marcus, 2006. Towards a Continuum of Scholarship: The Eventual Collapse of the Distinction between Grey and Non-Grey Literature. *Publishing Research Quarterly*. Springer US, **22**(1), p. 4-11. ISSN 10538801.

CARROLL, Michael W., 2006. Copyright in „Pre-Prints" and „Post-Prints". *Carrollogos* [online]. [cit. 2015-09-30]. Available from: **http://carrollogos.blogspot.cz/2006/05/copyright-in-pre-prints-and-post.html**.

CONNELLY KOHUTOVÁ, Radka, 2013. *Databáze ve věku informační společnosti a jejich právní ochrana*. Praha: C.H. Beck. Právní instituty. ISBN 978-80-7400-493-3.

DAVISON, Mark J. a P. Bernt HUGENHOLTZ, 2005. Football fixtures, horse races and spin-offs: the ECJ domesticates the database right. *European Intellectual Property Review*. (3): 113–118. ISSN: 0142-0461.

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

EUROPEAN COMMISSION, 2014. *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020* [online]. European Commission. Available from: **http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf**.

GUÉDON, Jean-Claude, 2004. The “Green” and “Gold” Roads to Open Access: The Case for Mixing and Matching. *Serials Review* [online]. 30(4): 315–328. ISSN 0098-7913. Available from: doi:10.1016/j.serrev.2004.09.005.

GUIBAULT, Lucie M. C. R., 2011. Owning the Right to Open Up Access to Scientific Publications. In: Lucie M. C. R GUIBAULT a Christina ANGELOPOULOS, ed. *Open Content Licensing: From Theory to Practice* [online]. Amsterdam: Amsterdam University Press, p. 137–167. ISBN 978-90-485-1408-3. Available from: **http://www.oapen.org/download?type=document&docid=389501**.

GUIBAULT, Lucie M. C. R a Andreas WIEBE, ed., 2013. *Safe to be open study on the protection of research data and recommendations for access and usage* [online]. Göttingen, Germany: Universitätsverlag Göttingen [cit. 2015-09-30]. ISBN 978-3-86395-147-4. Available from: **http://webdoc.sub.gwdg.de/univerlag/2013/legalstudy.pdf**.

HERR, Robin Elizabeth, 2008. *Is the Sui Generis Right a Failed Experiment: A Legal and Theoretical Exploration of How to Regulate Unoriginal Database Contents and Possible Suggestions for Reform*. Copenhagen: DJØF Pub. ISBN 978-87-574-1881-1.

HUSOVEC, Martin, 2012. Judikatórna harmonizácia pojmu autorského diela v únijnom práve. *Bulletin slovenskej advokácie*. Bratislava: Slovenská advokátska komora, vol. 18, n. 12, p. 16–20.

KOŠČÍK, Michal a Jaromír ŠAVELKA, 2013. Dangers of Over-Enthusiasm in Licensing under Creative Commons. *Masaryk University Journal of Law and Technology* [online]. Brno: Masarykova univerzita, **7**(2) [cit. 2015-09-30]. ISSN 1802-5951. Available from: **https://journals.muni.cz/mujlt/paper/view/2633**.

MYŠKA, Matěj, 2014. Vybrané právní aspekty otevřeného přístupu k vědeckým publikacím. *Právní rozhledy*. **22**(18): 611–619. ISSN 1210-6410.

POLČÁK, Radim, 2010. Legal Aspects of Grey Literature. In: Petra PEJŠOVÁ, ed. *Grey Literature Repositories* [online]. Zlín: VeRBuM, p. 67–89 [cit. 2015-09-30]. ISBN 978-80-904273-6-5. Available from: **http://invenio.nusl.cz/record/97129/files/idr-285_1.pdf**.

POLČÁK, Radim a Jaromír ŠAVELKA, 2009. *Digitální zpracování tzv. šedé literatury pro Národní úložiště šedé literatury* [online]. Brno: Masarykova Univerzita, Právnická fakulta, 2009. Available from: **http://repozitar.techlib.cz/record/284/files/idr-284_1.pdf**.

POPPER, Karl R., 2002. *The Logic of Scientific Discovery*. London; New York: Routledge. ISBN 0-203-99462-0.

RIEGER, Sören, 2010. *Der rechtliche Schutz wissenschaftlicher Datenbanken*. Tübingen: Mohr Siebeck. ISBN 978-3-16-150377-1.

SHAVELL, Steven, 2010. Should Copyright of Academic Works be Abolished. *Journal of Legal Analysis*. No. 1, p. 301. ISSN 1946-5319.

SCHEUFEN, Marc, 2015. *Copyright Versus Open Access* [online]. Cham: Springer International Publishing [cit. 2015-09-30]. ISBN 978-3-319-12738-5. Available from: **http://link.springer.com/10.1007/978-3-319-12739-2**.

SCHÖPFEL, Joachim, 2011. Towards a Prague Definition of Grey Literature. *Grey Journal (TGJ)*. Vol. 7, no. 1, p. 5–18. ISSN 15741796.

SUBER, Peter, 2012. *Open Access*. Cambridge: MIT Press. ISBN 978-0-262-51763-8.

TELEC, Ivo a Pavel TŮMA, 2007. *Autorský zákon: komentář*. 1. ed. Praha: C.H. Beck. Velké komentáře. ISBN 978-80-7179-608-4.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

FROM AN INSTITUTIONAL REPOSITORY TO THE BASE OF KNOWLEDGE - CASE STUDY

Weronika Kubrak

W.Kubrak@bg.pw.edu.pl

Warsaw University of Technology - Main Library, POLAND

Abstract

The Warsaw University of Technology Base of Knowledge it is not only a institutional repository but also good place to promote the scientific activities of the university staff. Unquestionable feature of this Base is that it allows present not only published papers but also collect f.e. patents and projects documentations, professional activity of our staff and students dissertations. The paper presents the advantages of system which combines functions of a repository and the Base of Knowledge functions.

Keywords

Repositories, Base of Knowledge, University Repository, Scientific Output

Introduction

The main objective of the Warsaw University Base of Knowledge (WUT Base of Knowledge), based on the OMEGA-PSIR[68] system, is to present and disseminate research achievements of the University faculty and students, both in Poland and all over the world. It provides access to academic and scientific publications, reports, dissertations and theses, as well as to information on the current research at the Technical University of Warsaw. Thus, the Base of Knowledge contributes to widening contacts and developing cooperation between the WUT academic staff and scientific and business communities in Poland and abroad.

One of the most important parts of the system is a repository which archives metadata and digitized full-text documents, such as monographs, journal articles, book chapters, reports and papers, such as dissertations and theses which are necessary for awarding university degrees and academic titles. The data in the repository support the needs of the university research, promote the university achievements and are also used for internal and external reporting.

Effects of implementation and usefulness of the system

During the implementation of the OMEGA-PSIR system and its later development, its creators and editors[69] responsible for data entry encountered a number of problems which turned out to be a barrier to build a joint University repository. The most important and most common problem was how to convince the authorities of some faculties to stop using their own databases on scientific achievements of the WUT researchers - some, faculties use their own systems, often incompatible with the new one. Therefore, the priority was to convince the authorities of each faculty to stop using their databases and to transfer all data to the Base of Knowledge. A decisive argument for such change was the Resolution of the WUT Senate No. XLVIII / 2012 of November 21, 2012 concerning the rules for creating a central system for recording and archiving of the writing, publishing and teaching achievements of the University staff, doctoral student and students, all University units and the WUT Repository. Then followed resolutions to specify the activities and the level of responsibilities of particular groups responsible for creating the WUT Base of Knowledge.

Thanks to the support of the University authorities, the Base of Knowledge became the main source of information to serve both recording and presenting the scientific achievements of the WUT academic staff and students. The Base is also used for reporting and for promoting science and research.

Capabilities of the OMEGA-PSIR system

The main purpose of the software authors was to create an integrating system, accessible to everyone, which would not only include scientific publications and other documents but

[68] OMEGA-PSIT it's software designed by a team from the Warsaw University of Technology Faculty of Electronics and Information Technology and is used by the University repository. The same team is involved in developing the research Base of Knowledge.

[69] Each University faculty has to appoint a faculty editor – a person responsible for entering the data on publications submitted by authors. Currently, there are over 120 editors. There are between several and over a dozen editors in a faculty, depending on the faculty size and the number of publications.

would also enable using the data for reporting purposes. The system functionality can therefore be divided into several types:

1. The repository functions related to recording achievements of the university scientists, archiving the achievements (in accordance with the copyright protection) and provision of the following:
    - Books and chapters
    - Journal papers
    - Engineer's and Bachelor's theses
    - PhD theses
    - Researchers projects
2. Presentation of documentation on the WUT projects and patents
3. Database on presented papers and published conference proceedings
4. Reporting on internal needs of the University, including generating reports for the academic staff assessment
5. Reporting for the needs of external units which evaluate the University activities (e.g.: Pol-on[70], MNiSW[71], PBN[72]) and communication with external systems (e.g. Google Scholar)
6. Presentation of the University scientific achievements and transfer of knowledge within the university and outside
7. Presentation of experts from various fields of science
8. Building a tag cloud presenting areas of research throughout the University, its individual units and each staff member.

Base Of Knowledge as a source of information about the university units and scientific researchers

The faculty "editors", appointed by the faculty authorities, are responsible for entering the data from their respective faculties into the repository, which is and integral part of the University Base of Knowledge. The University academic staff are obliged to submit the data to the Base of Knowledge for the purpose of updating it. It is also the only source of obtaining the data necessary for assessing the staff and the University unit, which is in turn necessary for the national system of assessing the functioning of the Polish universities (Pol-on, PBN). The lack of up-date data on the units means that it will not be taken into account in the reports.

[70] POL-on is an integrated information system for higher education, supporting the work of other Polish systems

[71] MNiSW - Ministry of Science and Higher Education

[72] PBN - Polish Scholarly Bibliography is a portal of the Polish Ministry of Science and Higher Education, collecting information about publications of Polish scientists and Polish and foreign scholarly journals.

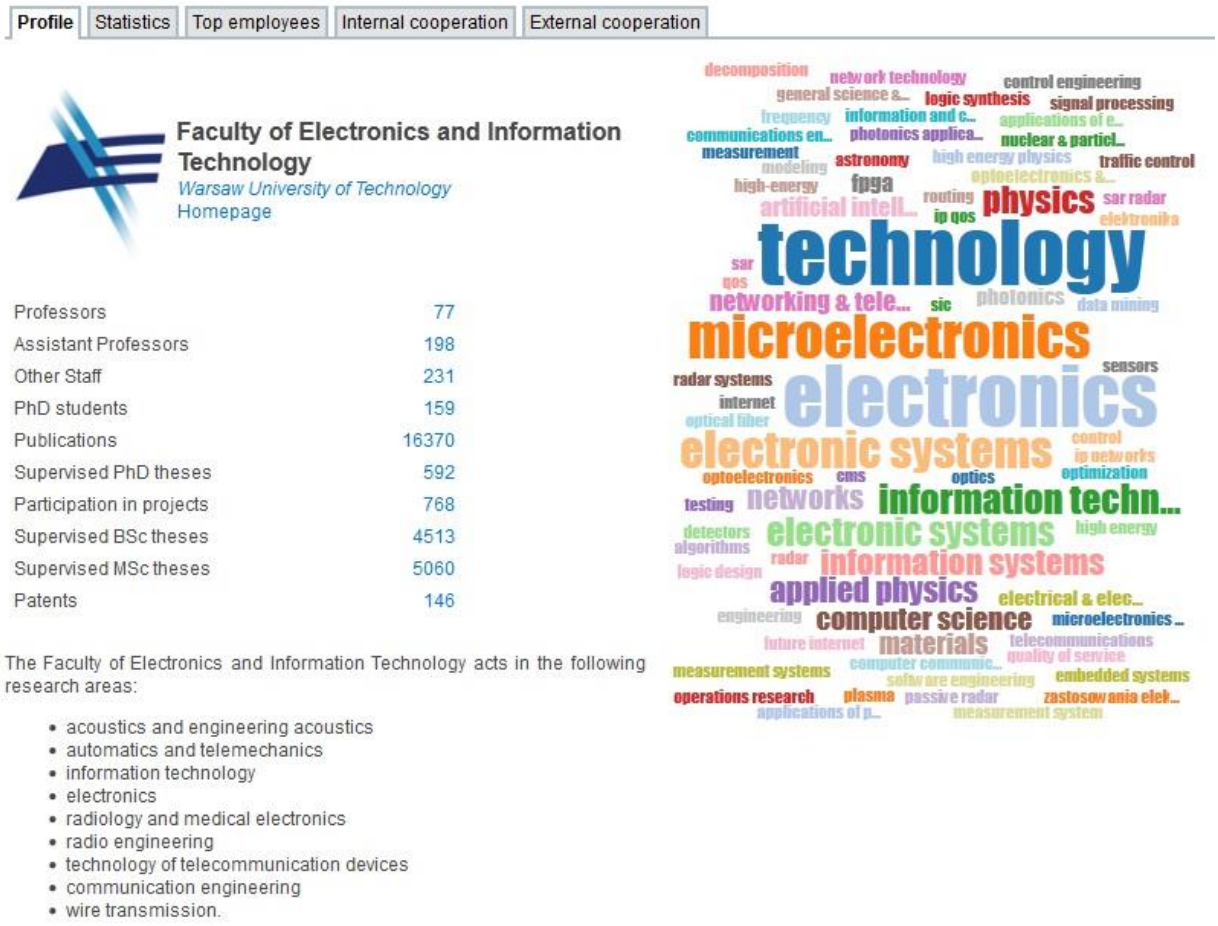

Figure 1 Profile presenting the characteristics of the Faculty

The figure shows an example of the faculty profile page where you can find information on not only the characteristics of the Faculty activities and the list of publications, but also:

- Statistics – graph demonstrating the increasing number of publications of the unit's authors over the years and the points awarded for them in accordance with the Ministry's regulations
- Top employees – graph presenting researchers with the highest number of publications in the unit
- Internal cooperation – schemata presenting all internal WUT units with which the unit cooperates
- External cooperation – refers to external units cooperating with the faculty
- BSc and MSc – list of BSc and MSc degree students thesis
- PhD – list of PhD students of the faculty
- Projects and Patents – list of projects implemented in the faculty and patents owned by WUT researchers.

The Cloud of Tags in profiles, generated at irregular frequency on the basis of key words added to the description of publications. Each keyword from the cloud of tags refers to publications linked to the given word.

The home page of the University researcher looks similar.

Figure 2 Researcher's profile

Similar to the faculty profile, also in this case, all information concerning the scientific activities of the researcher are divided into groups and provide the following information:

- Profiles – provides contact information and the characteristics of scientific activity
- Publications – list of researcher's publications
- PhD – list of doctoral theses promoted by the scientist
- Projects – list of metadata and attached documents, presenting scientific projects in which the researcher participated
- Activities - information on memberships, e.g. in organization/s, participation in editorial committees, committees organizing conferences or seminars
- Citation – statistics related to the author's most frequently cited publication. In addition, on the basis of this data, the Hirsch index is calculated - approximate calculation obtained in the Repository based on the scientist's publications (including auto citations) in the Repository and the Internet information analysis. The value is close to the value obtained with the Publish or Perish system. In general, it is higher than the value given by the Scopus or Web of Science sites.
- Statistics – the same as in the case of the faculty profile, the tab presents a graph showing increasing number of publications and the Ministry points awarded over the years

- Cooperation – graph shows scientific co-authors of the researcher's publications and researchers cooperating in projects implementation and patents development

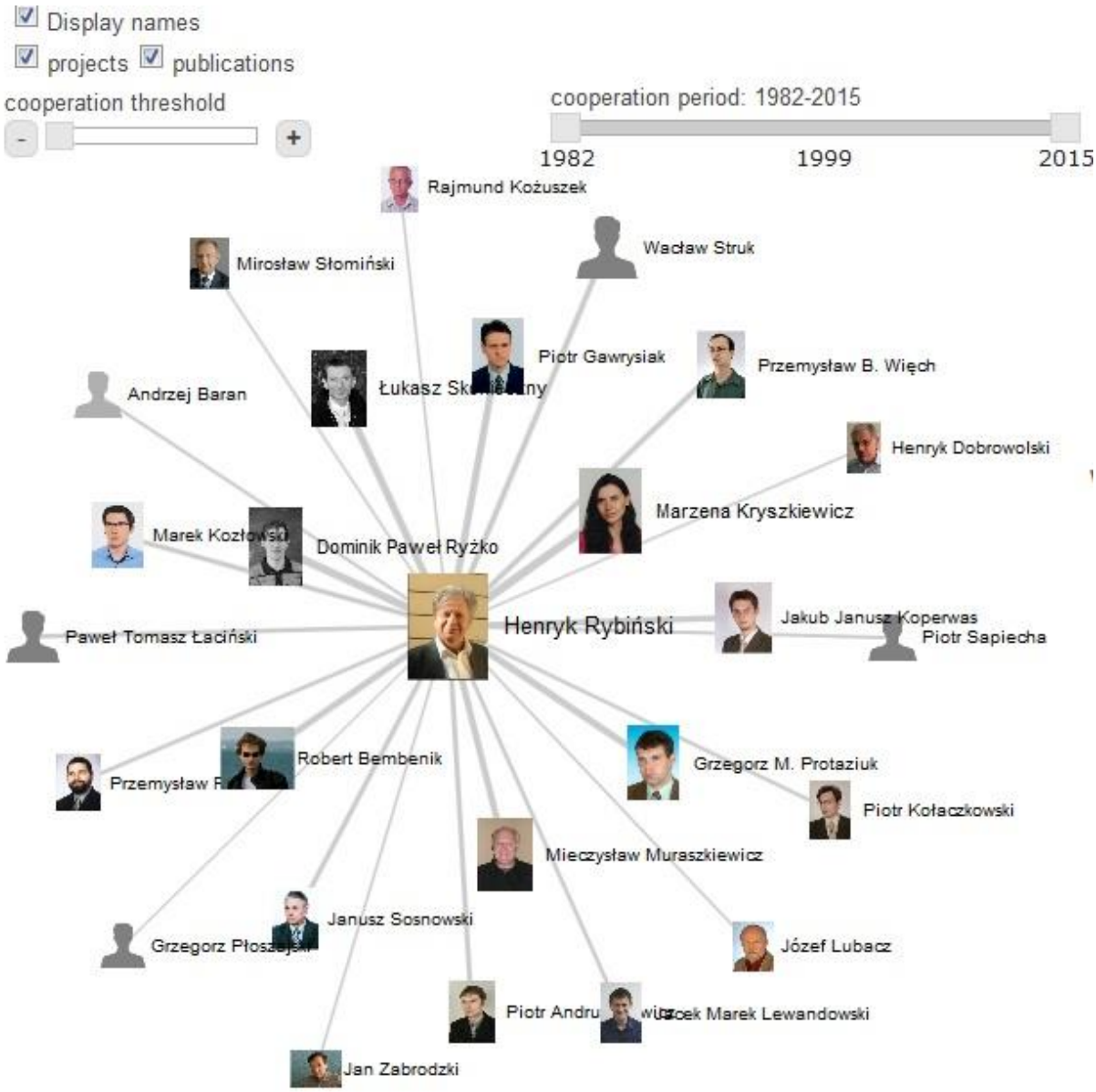

Figure 3 Cooperation graph in researcher's profile

Base of Knowledge used for reporting

The final function of the Base of Knowledge discussed in this paper is the capacity of generating various types of reports. The first and the simplest one is a report on the overall scientific activities of the researcher. Such report includes the University logo and can be attached to e.g. documents needed to assess the researcher while applying for professional

advancement. This document is generated as a PDF[73] file and it can be also used as part of the researcher's resume. In addition, the repository can generate reports necessary for preparing the unit annual report or if needed by the faculty dean. Depending on the type of report, it can be generated as an Excel file, CSV[74], Bibtext[75] or as a bibliography presented on the Base of Knowledge website for internal use. The most comprehensive report is a survey of the unit and is prepared every four years. The report is generated in the form of an Excel file and its fields it can be, if necessary, edited and sent to the national system of assessing the Polish universities, assuring that the transmitted data are complete and do not require additional adjustments.

show level 3

The Repository Statistics

WUT - Warsaw University of Technology

13-09-2015 / 14:36

General statistics	Publications all	Publications 2009-2014	Publications 2014	Publications (Report) (Art+Roz+Ksi)	Publications (obce języki)	Publications (jęz. polski)	PhD, M.SC, B.Sc. (BSc+MSc+PhD)	Patents (Zgł+Pat)
Warsaw University of Technology	WUT	WUT	WUT	WUT	WUT	WUT	WUT	WUT
WUT Central Administration	5	3	1	0+1+0	0	1	0+0+0	0+0
Main Library of Warsaw University of Technology	107	70	19	0+0+0	0	19	0+0+0	0+0
WUT Information Technology Centre	2	0	0	0+0+0	0	0	0+0+0	0+0
College of Economics and Social Sciences	195	186	66	31+17+4	15	51	11+0+0	0+0

Figure 4 Repository statistics - several types of choice e.g. all publications in repository

Conclusion

Each OMEGA-PSIR function listed in this paper will surely facilitate reporting of the University faculties. The Base of Knowledge has been operating for not long, but we can already say that the implementation of this joint University system was a right decision. The process of preparing reports has been facilitated and standardized. The function of promoting knowledge and the scientific potential of the University and its staff is also a great advantage. All metadata entered into the Base are accessible through the Internet without any limitations. Some data are available together with full-text files that have been included in the Base in compliance with the Polish copyright law.

Undoubtedly, despite the advantages of our Base, there still exist several obstacles: the principles for assessing research units often change, new bases for assessment are being created, the authors are often reluctant to make their work pubic. Therefore, the creators of the Base of Knowledge and people who oversee its merits have to make sure that the Base is being advanced on ongoing basis, that it is compatible with such external systems as POL-on or PBN. They also have to broaden their knowledge on the changing legislature and be up-to-date with constantly developing Open Access in Poland and all over the world.

[73] PDF - Portable Document Format

[74] CSV - data storage format in text files.

[75] Bibtext - tool for formatting bibliographies.

References

KOPERWAS, Jakub Janusz, RYBIŃSKI, Henryk and Łukasz Skonieczny. *Projekt i implementacja pilotowego systemu repozytorium dla prac dyplomowych (inżynierskich, magisterskich i doktorskich) oraz publikacji pracowników Politechniki Warszawskiej*. Warsaw: the Warsaw University of Technology, Faculty of Electronics and Information Technology, dec. 2010 [cit. 2015-12-07].

STEPNIAK, Jolanta. *Ewidencja dorobku i repozytorium Uczelniane* [online]. Warsaw: Biblioteka Główna Politechniki Warszawskiej, 2012 [cit. 2015-12-07]. Available from: **http://bcpw.bg.pw.edu.pl/dlibra/docmetadata?id=3869**.

NIEZGÓDKA, Marek. *Wdrożenie i promocja otwartego dostępu do treści naukowych i edukacyjnych: Praktyki światowe a specyfika polska. Przewidywane koszty, narzędzia, zalety i wady.* [online]. Warszawa: Ministerstwo Nauki i Szkolnictwa Wyższego, 2011, 288 p. [cit. 2015-12-07]. Available from: **http://depot.ceon.pl/handle/123456789/1545?show=full**.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

THE ROLE OF THE ACADEMIC LIBRARY IN DISSEMINATING GREY LITERATURE – ADAM MICKIEWICZ UNIVERSITY REPOSITORY AS A CASE STUDY

Małgorzata Rychlik

rychlik@amu.edu.pl

Poznań Uniersity Library, Poland

Abstract

One of the main and crucial strategic projects that any university, or an institution of higher learning, has been recently involved in is the creation and development of an institutional repository. The AMUR repository is an open web-based archive of scholarly material that originated in 2010 and has aggregated a collection of more than 12,000 diverse digital objects. This paper presents the types of grey literature that have been archived in the repository. The following key elements concerning grey digital objects are presented: policy of the repository, legal issues, standards for metadata and the language coverage. In addition, usage data are provided.

Keywords

Adam Mickiewicz University Repository, AMUR, Grey Literature, Open Access, Usage Statistics, Scientific Communication

Introduction

The libraries of universities and research institutions all over the world contribute to development of scientific communication by launching OA repositories, among others. As a result, the intellectual output of an institution can be widely visible in the world. OA repositories collect, preserve and disseminate different types of digital objects. These objects include those distributed by traditional channels, such as articles and books published by traditional publishing houses. However, part of digital content archived in repositories can be described as the so-called underground literature (Boukacem-Zeghmouri, Schöpfel, 2005). Until recently, that type of content has not been fully indexed and could be encountered usually in the "Invisible Web" area (Derfert-Wolf, 2007). Thanks to OA repositories that comply with the Open Access Initiative Protocol for Metadata Harvesting and provide persistent URLs, it is now possible to search and retrieve materials defined as grey literature. For the purpose of this article I will use currently most frequently used New York definition (Myška, Šavelka, 2013) that characterizes grey literature as follows: "[Grey literature is] that which is produced on all levels of government, academics, business and industry in print and electronic formats, but which is not controlled by commercial publishers i.e., where publishing is not the primary activity of the producing body" (Schöpfel, 2010).

An analysis of the content types in OpenDOAR[76] shows that grey literature is by no means a marginal part of the content. Among five most frequent content types of OA repositories, three of them belong to a group defined as grey literature. At present, OpenDOAR indexes 2,973 OA repositories. Electronic Theses and Dissertations (abbrev. ETD) are collected by 1,643 OA archives, unpublished reports by 1,078 and conference and workshop papers by 1,063 OA repositories (Figure 1).

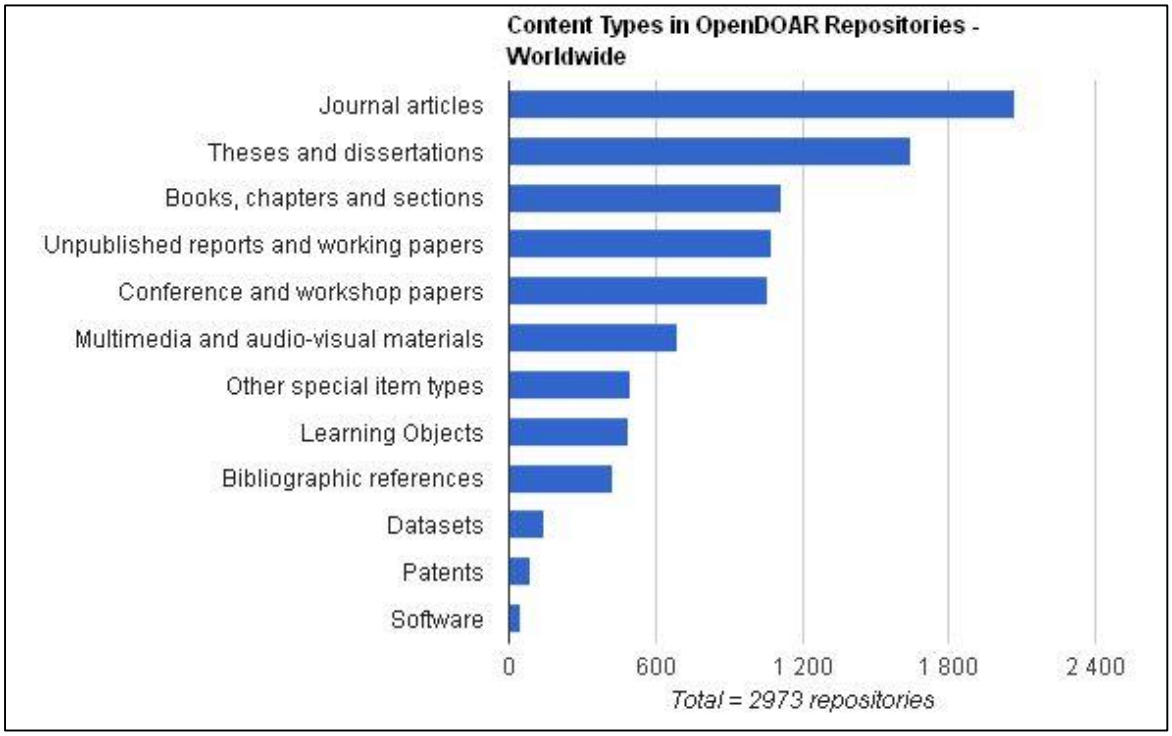

Figure 1 Content types in 2,973 OA repositories

[76] **http://www.opendoar.org/index.html**

Adam Mickiewicz University Repository (AMUR)[77]

The first institutional repository in Poland was established at the Adam Mickiewicz University in Poznań in 2010. The University is one of the finest Polish universities. More than 40,000 students and 1,300 PhD students attend the University. There are over 3,000 scholars employed. The AMUR repository is a big institutional archive with more than 12,000 digital objects. It ranks 155th among 2,275 indexed repositories in the Ranking Web of Repositories (July 2015 edition). The ranking places the AMUR platform in the first place among Polish repositories. This repository is based on DSpace software.

Types of grey literature in the AMUR repository

Communities and collection in the AMUR repository have a hierarchical structure. The basic framework of the repository is composed of 15 communities of the university faculties plus the community of journals published at the university and the community of ETD. About 89% of digital objects collected in the repository accounts for white literature (articles, books and book chapters). The remaining 11% belongs to grey literature. ETD outnumbers the remaining collections that archive several dozen digital items each (Figure 2). The total number of grey items in the AMUR repository is 1,411.

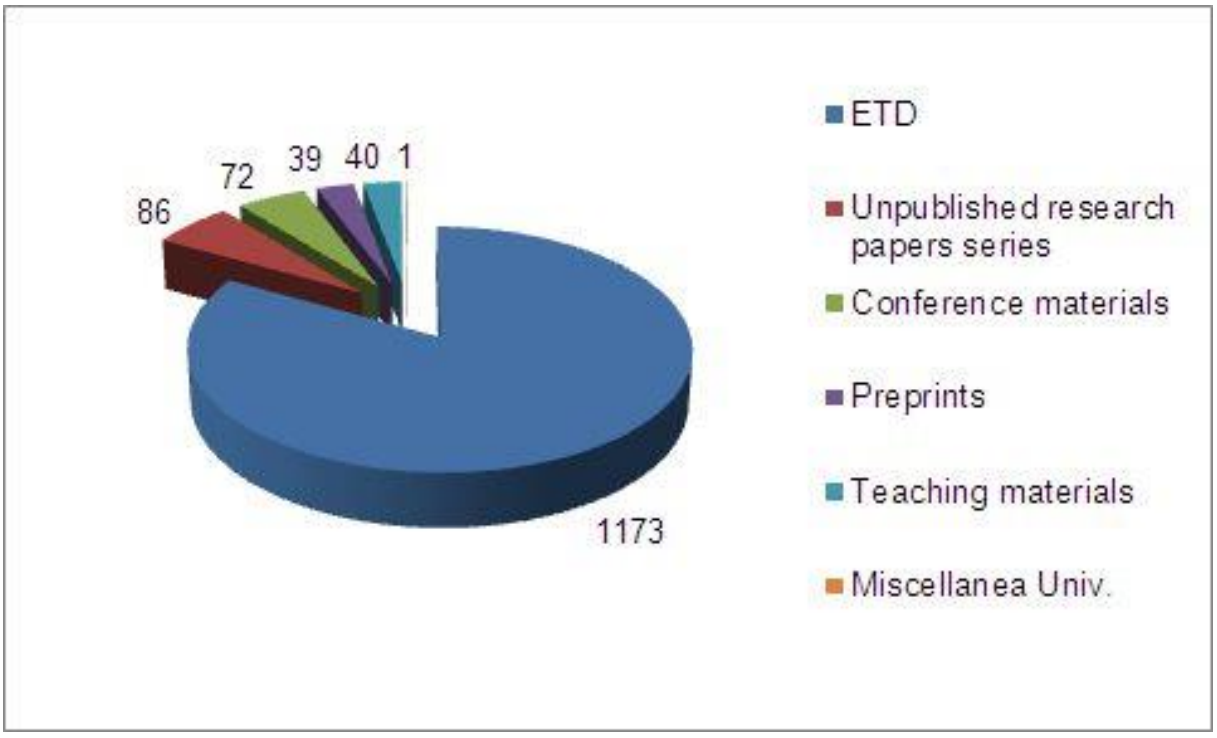

Figure 2 The share of different types of grey literature in the AMUR repository

[77] http://repozytorium.amu.edu.pl

Policy

One of the main aims of OA repositories is to disseminate the scientific and scholarly output of an institution. And one of the primary ways to improve content acquisition is to implement the open access mandate. This requires researchers to make their works open access by self-archiving. So far, more than 500[78] research organizations around the world have obliged scholars to deposit their works in OA repositories. It is essential to notice that these OA mandates are not always fully open and are subject to internal regulations of an institution. In the AMUR repository, there is a regulation implemented by the Rector of Adam Mickiewicz University which says that all the PhD theses defended at the University since 2010 have to be deposited in the repository. Each doctoral candidate signs a non-exclusive license and grants either the right to use PhD thesis in open access or limits the use of PhD thesis to just on-campus-only access. Hence, about 7% of the content of the repository is not OA.

In the AMUR repository all white content is open access, but as regards grey literature the percentage of PhD theses with restricted access is quite high. As many as 820 of ETD are not open due to on-campus restrictions. It gives 59% of grey literature as non-OA, whereas 41% remains open access.

Metadata

Many repositories implement the Dublin Core scheme composed of 15 classic metadata set. The scheme provides the ability to expand by adding qualifiers (subattributes to specify the main attribute). This allows the order of the attributes to be changed. Selected data can be modified, which may be useful for particular local applications. However, it is essential to realize that such local solutions often bring information noise. In the AMUR repository, we use a set of 15 elements (title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage and rights). One specific element is created as regards grey literature - dc: thesis supervisor. This element consists of the name of a PhD thesis supervisor.

Legal issues

Generally, digital content of repositories is protected by copyright. In the Open Science area there is a distinction between two kinds of OA: gratis OA and libre OA. The one is defined as free, perpetual, open access which grants the right to use the work in the frame of fair use (i.e. for personal use and didactic purposes) to all users (Siewicz, 2012). Gratis OA occurs widely in OA repositories, although better and more efficient for the users would be the other type, i.e. libre OA that removes not only price barriers (as gratis OA does), but also at least some permission barriers (Suber, 2012).

For users, it would be much easier to have access to a work which is offered under a free license. Hence, it is really essential to make scholars aware of the necessity of possessing corresponding rights enabling them the use of adequate free licenses. As it happens, an author who does not possess such rights, offers his/her work under the CC license (Polčák, 2010). Many authors while self-archiving their works in OA repositories are in the possession

[78] **http://roarmap.eprints.org/**

of publisher's agreement for depositing the work only in the repository (gratis OA). Therefore, gratis OA is much more common than libre OA in OA repositories.

In the AMUR repository, approximately 10% of the whole content is offered under CC licenses. In the case of grey literature only 41 digital objects are under the CC licenses.

In our repository, each PhD candidate signs a non-exclusive license (see Policy). This results from the lack of regulations concerning PhD theses both in the Act of 4 February 1994 on Copyright and Related Rights as well as in the Act of 27 July 2005 on Law of Higher Education.

Usage data [79]

Usage data provide information about the use and access to works deposited in repositories. The data enable the use of repositories content to be measured and help in promoting the content. Usage statistics are gathered at the level of individual objects as well as they are available for the whole content within its different aspects.

While analysing usage statistics of grey literature in the AMUR repository we can see that both ETD and teaching materials are downloaded at similar levels - their average number of downloads is around 1,500 per item. Conference materials and preprints have got much lower number of downloads per item. Both, Miscellanea Universitatis and the collection of unpublished research papers series represent the extreme results - the highest and the lowest (Figure 3).

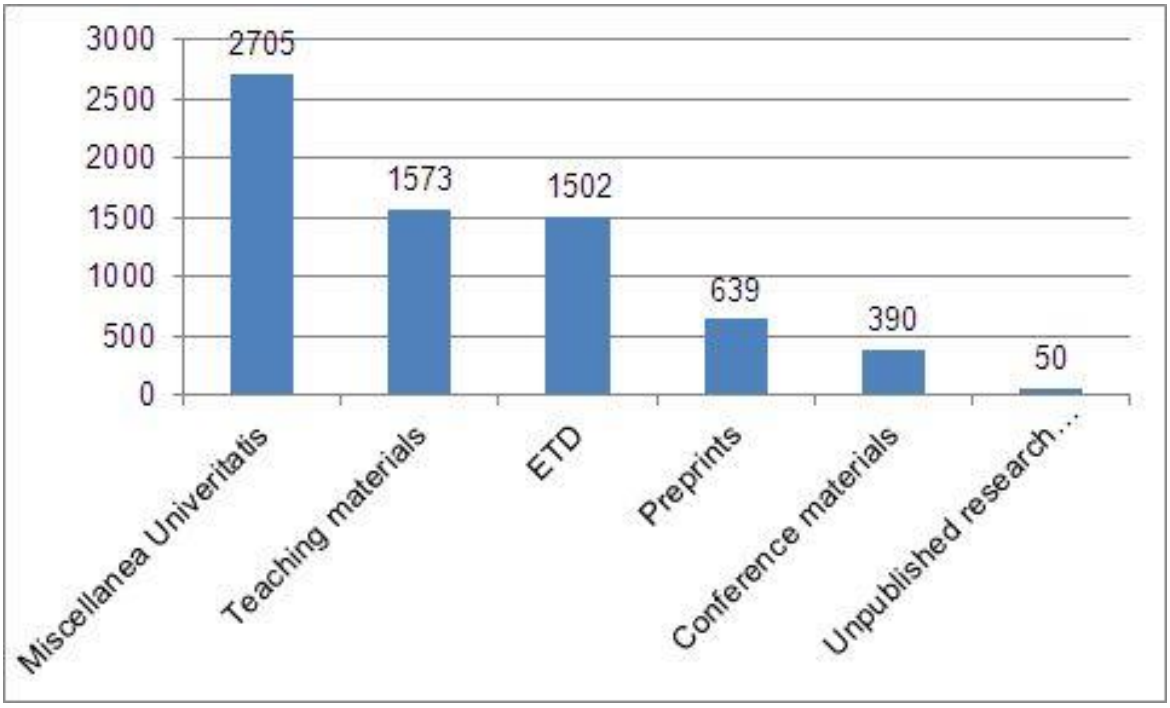

Figure 3 Average number of downloads per item of grey literature in the AMUR repository

[79] Usage data are provided only for grey OA items in AMUR (820 non-OA PhD theses are excluded)

While comparing usage statistics of white and grey literature we can observe that grey items are downloaded more frequently than white, which is well visible (Figure 4). For a comparison, I used three most frequently downloaded types of the white content (articles, books and book chapters), and three most frequently downloaded types of the grey content (teaching materials, ETD and preprints). I excluded Miscellanea Universitatis as this collection contains so far only one item and therefore is incomparable. The average number of downloads for grey literature is 1,238, whereas for white it is 579,7.

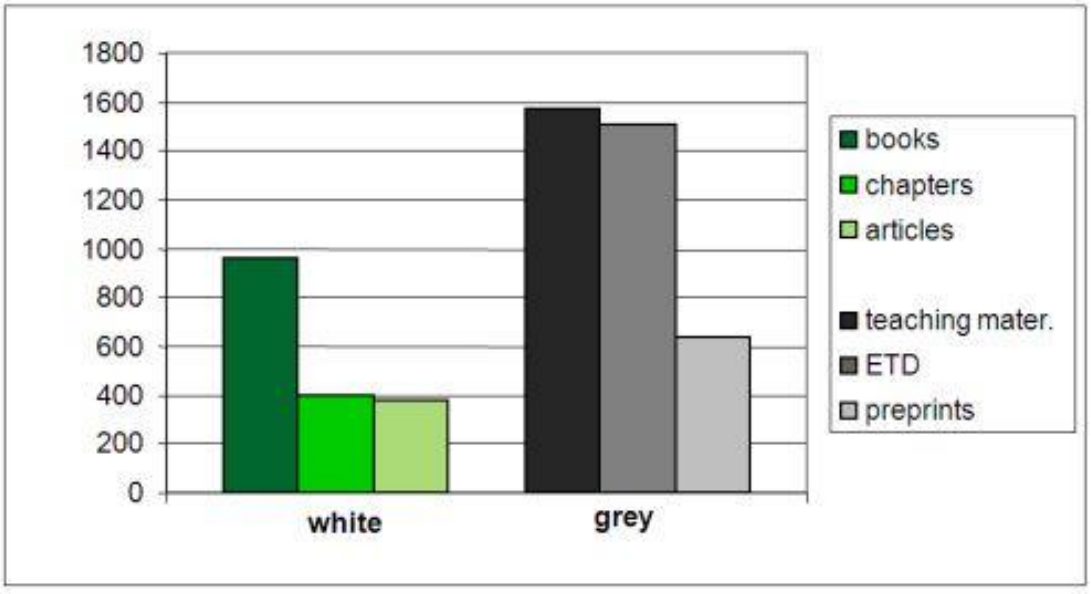

Figure 4 Comparison of average number of downloads between grey and white literature

Our usage statistics can help determine our users. We can trace their IP addresses and, based on download statistics of grey literature, are able to say that 50% of our "grey" users are from Poland, 17% from the States and the remaining 33% come from other countries (all continents are represented). Download origins are related to language coverage. The digital objects in the primary language of the repository (and most often the country) are more frequently downloaded from IP addresses located in that country than from other addresses. As to the language coverage in our repository, grey items written in Polish (70%) outnumber those written in English or other languages (Figure 5).

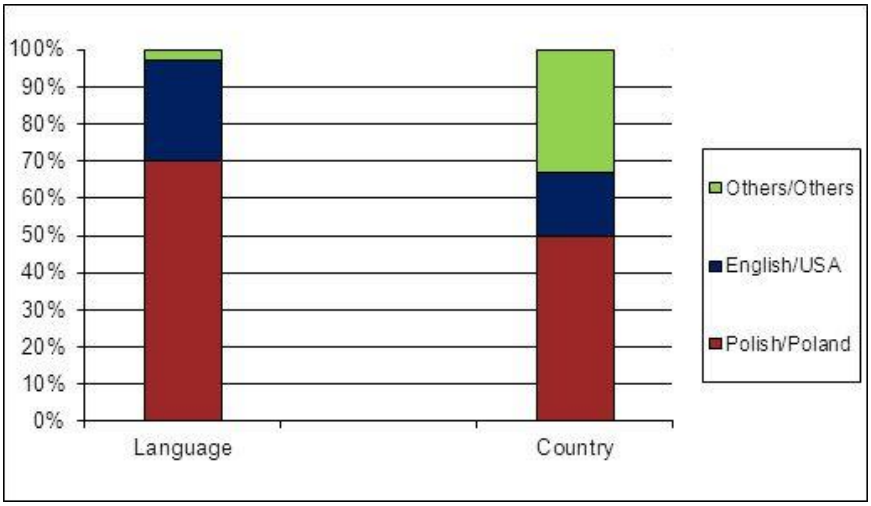

Figure 5 Language coverage of grey literature in the AMUR repository and download origins (by countries)

Conclusions

A determination of the core issues concerning grey content in an OA repository is a vital step towards better understanding of the role of grey literature.

- In the case of AMUR, the ratio of the total amount of digital items to grey literature ratio is almost 10 to 1. As we see in Figure 4, the average number of downloads of grey literature is much higher than that representing white literature, therefore we should implement a promotion strategy which would be helpful in gaining more grey literature into the repository.
- Undoubtedly, legal issues are the ones of most frequent barriers relating to self-archiving. The awareness of scholars in the context of granting free licenses needs to be raised. The lack of access to 820 PhD theses is an immense disadvantage for our users. Unfortunately, this situation is not going not change considerably for some time until a new law concerning theses and dissertations has been enacted in Poland.
- There are some decisive elements concerning metadata scheme that we should work on. Unquestionably, new elements regarding conference materials should be created.

A robust and efficient repository would be a project based on a cooperation between librarians and researchers. It is important to strengthen a synergy between these two groups. Librarians should remain responsive to new developments in the Open Science area. They should implement new tools and standards which can be useful in disseminating knowledge. While the task of scholars is to actively build a repository content by self-archiving their works (both white and grey). A growing number of scholars believes that such an activity can bring tangible benefits to both researchers and their institutions in terms of greater visibility and citation advantage.

References

BOUKACEM-ZEGHMOURI, Chérifa and Joachim SCHÖPFEL. *Access and document supply: a comparative study of grey literature*. 7th International Conference on Grey Literature, Dec 2005, Nancy [online]. [cit. 2014-09-03]. Available from: **http://archivesic.ccsd.cnrs.fr/sic_00091843/document**.

DERFERT-WOLF, Lidia. Odkrywanie niewidzialnych zasobów sieci. In: *II seminarium z cyklu Infobroker: Wytwarzanie i przetwarzanie cyfrowych informacji, Warszawa 17 kwietnia 2007*. [online]. [cit. 2014-09-14]. Available from: **http://eprints.rclis.org/8862/1/derfert_CPI.pdf**.

MYŠKA, Matěj, ŠAVELKA, Jaromír. *A model framework for publishing grey literature in Open Access* [online]. 2013, vol. 4, no. 2, p. 104-115 [cit. 2011-09-05]. ISSN: 2190-3387. Available from: **http://www.jipitec.eu/issues/jipitec-4-2-2013/3744/myska-savelka.pdf**.

POLČÁK, Radim. Legal Aspects of Grey Literature. In: P. Pejšová ed.. *Grey literature repositories* [online]. Zlín: VeRBuM, 2010 [cit. 2014-09-05] Available from: **http://eprints.rclis.org/8862/1/derfert_CPI.pdf**.

SCHÖPFEL, Joachim. *Towards a Prague Definition of Grey Literature Twelfth International Conference on Grey Literature: Transparency in Grey Literature,. Prague, 6-7 December 2010, Czech Republic* [online]. 2010 [cit 2014-09-03]. Available from: **http://archivesic.ccsd.cnrs.fr/sic_00581570/document**.

SIEWICZ, Krzysztof. Otwarty dostęp do publikacji naukowych. Kwestie prawne [online]. Warszawa: Wydawnictwa Uniwersytetu Warszawskiego, 2012 [cit. 2014-09-03]. Available from: **http://repozytorium.ceon.pl/bitstream/handle/123456789/335/K_Siewicz_Otwarty_dostep_do_publikacji_naukowych.pdf?sequence=4**.

SUBER, Peter. *Open Acces* [online]. Cambridge: The MIT Press, 2012 [cit. 2014-09-14]. The MIT Press Essential Knowledge Series. ISBN 978-0-262-51763-8. Available from: **https://mitpress.mit.edu/sites/default/files/9780262517638_Open_Access_PDF_Version.pdf**.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

10 YEARS WITH GREY LITERATURE AT TOMAS BATA UNIVERSITY IN ZLÍN

Lukáš Budínský

budinsky@knihovna.utb.cz

Tomas Bata University in Zlín, Library, Czech Republic

Abstract

The Library of Tomas Bata University in Zlín has been working on accessioning academic grey literature systematically for the last 10 years. During this time, one of the first information systems storing theses and dissertations changed from a closed, manually updated system into a fully automated and fully open repository with a number of integrations with both internal and external systems. New types of documents were added and several services were introduced for working with grey literature effectively. This contribution will describe its current state and outline the hurdles encountered getting there.

Keywords

Grey literature, Theses, DSpace, Fulltext, License Agreement

Introduction

Tomas Bata University in Zlin (TBU) is a dynamically developing university whose six faculties provide a wide range of studies in the humanities, natural sciences and technical and artistic disciplines. It is a centre of top-class science and research in the country and, in many aspects, on an international scale. With around 10,300 students, TBU is a medium-sized university in the Czech Republic.

The library at TBU is known for its active approach to new challenges that affect it either directly or indirectly. It uses these opportunities to develop the range of services it offers students, academic staff and the public.

One of the challenges to which it has responded, as one of the first institutions in the Czech Republic to do so, is that of making dissertations and theses available in electronic format.

The origins of electronic dissertations and theses at TBU

It was already clear back in 2004 that the situation in taking in and making dissertations and theses accessible did not respect the possibilities and needs of all sides involved. Students were forced to print out 3 copies of their electronic documents, which piled up at institutes and were subsequently handed over to the library to be processed. Each paper was put into a catalogue and placed in the archive by hand. The library at that time was housed in a building with limited space and the ever-growing number of printed papers could no longer be stored. Students only had the opportunity to search them using basic metadata, before collecting them from the archive.

All these negative factors could be eliminated by making dissertations and theses available in electronic format. The document was created electronically, meaning there was no reason not to leave it as such. The library seized the moment in updating the "Uniform formal regulation of university dissertations and theses, their storage and accessibility" guidelines, in which electronic access to dissertations and theses was mentioned for the first time, specifically in the following points of Guidelines of the Rector No. 41/2004[80]:

(3) All papers shall be kept on record in the electronic catalogue of the TBU University Library and in the Digital Library of University Dissertations and Theses (hereinafter referred to as the ""UDT Digital Library"), where they shall also be stored.

(4) After having been processed, papers which have been defended shall be accessible within the UDT Digital Library in accordance with Licence Agreement.

It must be recognised that pushing through the e-UDT was a step into the unknown because no UDT digital library even existed at the time the guidelines came into force. The Study Information System (IS STAG) was modified during the winter so as to be able to link one file with full text to the submission of a paper. At the same time, an electronic licence agreement form was prepared for students to fill in when handing in their work and in doing so restrict the possibility of accessing the full text of the paper. Metadata was exported from IS STAG and a MARC 21 record was machine-created for import to the library catalogue.

[80] UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ. Jednotná formální úprava vysokoškolských kvalifikačních prací, jejich uložení a zpřístupnění. Univerzita Tomáše Bati ve Zlíně, 1.11.2004. Článek 3, Uložení a zpřístupnění. Guidelines of the Rector No. 41/2004. Available from: **http://www.utb.cz/file/12033/**

Let the problems begin

The switch to electronic format demanded a number of changes in processing UDT that did not pass without any teething troubles. The year 2005 was therefore considered transitional and the obligation to submit UDT in electronic format was anchored in the updated Guidelines of the Rector No. 5/2006[81]. Most of the shortcomings were gradually, but successfully resolved. Unfortunately, some are related to the system itself and persist to this day. What exactly were the problems?

- **Formatting papers and subsequent conversion to PDF**

Given the accessibility of full texts in electronic format to the general public, students were required to stick to uniform rules relating to the formal layout of their work. Templates were created for individual faculties in which students formatted their work. Unfortunately, poor computer literacy among students, combined with insufficient support for the use of templates, led to the situation in which library employees spent hours on end helping with the formatting of papers. Even then, the storage site contains a number of incomplete (no title page) or wrongly formatted e-UDT.

Back in 2005, of course, there was no comfortable tool available to facilitate conversion of documents from MS Word format to PDF file. Students used various tools that added undesirable watermarks in the resulting texts or that required a password for printing or even reading. There were no rules for checking legibility when adding a file. A problem also arose during subsequent full-text indexing, when certain programmes managed to properly convert the text visually, but Czech accents were entirely missing when extracting the text.

Solution: both problems essentially solved themselves over time as a result of students becoming better informed. Comments were made on templates and these were simplified so much that students no longer had any problem properly formatting the resulting text. Removing the licence agreement meant that they no longer had to integrate it with such complication in the resulting text.

A tool for uniform conversion to PDF documents (and instructions) was installed and pre-set on all computers in the library's study rooms. Over time, quality conversion to PDF documents became a standard feature of the main word processors.

- **Students adding information about their work**

Although efforts are made to ensure that information about a UDT is taken from the study system to as great an extent as possible, one integral part of submitting a paper is the provision of information by the students themselves, particularly the abstract and key words. Key words are still something that many students do not understand to this day. Whole sentences appear in the records, frequent misspellings and phrases such as "there are none" and "don't know". The fact is that key words are important when searching documents and for their indexing, particularly in a library system.

[81] UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ. Jednotná formální úprava vysokoškolských kvalifikačních prací, jejich uložení a zpřístupnění. Univerzita Tomáše Bati ve Zlíně, 17.4.2006. Guidelines of the Rector No. 5/2006. Available from: **http://www.utb.cz/file/11946/**

Solution: two alternative solutions were tested. First of all, the library tried to increase awareness among students at dissertation seminars and then in the form of hints by the box with key words. Unfortunately, neither method had any significant effect on the quality of key words provided. The other route was to correct key words before import to the library system. Nonsense was first removed by machine. Then several librarians divided up thousands of words among them and checked for misspellings. This was manageable in the first year given that they no longer had to catalogue all papers by hand. In the second year, correcting around 60,000 key words met with clear opposition and was eventually dropped.

The eventual solution came with the application of the DSpace system, which indexes documents full-text, meaning that key words became less important to the end user.

- **Importing e-UDT to the library catalogue**

Importing to the library catalogue is closely related to the previous problem. Picking complete metadata from the study system meant that the next step was to end the cataloguing of UDT by hand and import machine-generated records. Unfortunately, this plan met with strong opposition from the cataloguers, who considered the machine record to be imperfect and unsuitable for import. It was able to add papers to the catalogue for several years, but a change of concept eventually became a necessity as a result of the key words mentioned above.

Solution: there are several appropriate ways that can be chosen according to preference.

1. Continuing with the import of machine records and their subsequent modification. It is not a problem to create your own base for UDT in the catalogue, one that will not be mixed in indexes with hand-generated records of books acquired. It is now nothing special to import a collection of tens of thousands of electronic books records directly to a catalogue. It would be possible to leave out hand-generated and error key words.
2. Putting a separate system for providing access to e-UDT into operation. Informative education helps users fast become accustomed to looking for a certain type of document elsewhere than in the catalogue. The situation now has been simplified thanks to the existence of discovery systems that know how to integrate search results across sources.

- **Full texts and access to them**

Access to the full texts of e-UDT mainly underwent development in the first 5 years after being launched. Students had the opportunity to restrict access to their work back in 2006 using a licence agreement as follows: (1) prohibit its publication entirely; (2) a time embargo on publication of 1, 3 or 5 years; (3) restriction to the internal TBU network, and; (4) free publication. The opportunity to entirely prohibit publication was replaced in 2007 by an embargo for up to 30 years. These possibilities led to a frequent embargo on papers or restriction for an extremely long time.

Solution: first of all, the embargo was lowered to 10 years in 2008, before the licence agreement was abolished entirely in 2010. All papers are published in the following regime to this day:

1. all papers are mandatorily published without restriction;
2. in special cases, access to a paper may be restricted to the TBU internal network at the request of the dean of the faculty;
3. students have the chance to keep companies anonymous in the case of sensitive data.

Other documents we could class as grey literature have individual accessibility settings according to the requirements of faculties, institutions and authors.

- **The unsuitability of the study system for long-term archiving and providing access to documents**

Storing e-UDT and making these accessible in the study system was understood to be provisional from the very beginning. It was important to begin gathering data in any way, but making it permanently accessible was more of a marathon than a sprint.

Why is the study system unsuitable for permanent accessibility to e-UDT when all metadata is produced therein and full texts are stored there? The reasons are primarily as follows:

1. Technical:

The study system is one of the key systems in running the university and deals with all operative activities. The data therein must be safely stored and quickly available. In comparison with this, e-UDT are more like archive material that must be accessible over the long-term. After several years of storing full texts, the study system began placing huge demands on disk arrays, backing-up and the IT infrastructure as a whole.

2. Library-related:

Entirely different requirements are made of a system for the permanent accessibility of documents from the librarian's perspective. The system must be robust, secure and scalable. Nonetheless, it should also be open and comply with standards for cooperation with the surrounding world. The aim is to integrate data about e-UDT in a continually changing environment. The form of the library catalogue has already changed in the 10 years of its existence -> repository -> meta-searching -> discovery system.

Solution: several possible methods were considered. One that was quickly rejected was modification of the study system for the needs of long-term archiving and making documents accessible. The requirements of the two systems differ so greatly that it would have been technically demanding and in all likelihood unsustainable over the long-term.

The second option was to use the Aleph library system, for which the library has an ADAM module for providing access to electronic documents. As was shown in applicability tests, however, this module was created only as a quick stand-in for a fully-fledged system and was not developed any further. It did not, therefore, meet the functional demands of storing e-UDT.

Acquiring a fully-fledged commercial system did not even come under consideration in light of budget restrictions and so the library turned its attention to open-source projects. There were

rapid developments in this field in 2009 and the range on offer became sufficiently diverse. The DSpace system was eventually chosen, mainly for its active Czech user group[82].

DSpace at TBU – from grey literature to publication activity

Open-source systems have an undoubted advantage in terms of their acquisition costs, although the costs of implementation and of their operation must also be taken into consideration. These can be covered by the user itself or in the form of outside services. In the case of the library at TBU, these costs were spread over time. Implementation was entrusted to a contractor. A tool for importing records of e-UDT from IS STAG was developed to order. Ing. Ivan Masár from the library's IT department then took charge of the subsequent operation and development of the system and his active approach saw him become an official DSpace contributor.

What did the library and users gain from the application of DSpace[83]?

- a robust and expandable system capable of long-term archiving of e-UDT;
- automatic import of papers after defence, with reports on results;
- persistent identifier of records (Handle);
- a uniform, logically-structured metadata storage site;
- the storage of full texts and other documents relating to records (appraisals, defence);
- the opportunity to store non-text appendices and display them (photographs, videos);
- full-text searching and indexing of documents;
- the opportunity to restrict access to full texts according to licencing agreements;
- standardised protocols for communication with external systems;
- access to the source code and the option to broaden the functionality of the system.

The library was able to concentrate on broadening the type of documents stored after putting a stable system into operation. Faculties were first offered the electronisation of published scripts which had not been sold at the planned level of circulation for some considerable time. Although students unanimously voted in favour of electronisation of scripts in a survey, conservative faculties resisted for several years, insisting on loss-making printing. Only the Faculty of Applied Informatics and the Faculty of Humanities took up the offer and now publish scripts and documents in electronic format, with access restricted to within the university. Other activities that promoted electronic scripts at other faculties were discontinued due to lack of interest.

The DSpace platform saw most development at TBU in the form of the institutional repository. The path that was chosen, for many reasons, was complete separation of the digital library of e-UDT and the repository. Although the repository of publication activity is a unique application of the DSpace system, it deals with the issue of grey literature only peripherally

[82] **http://dspace.cz/**
[83] **http://digilib.k.utb.cz**

(automatic archiving of published articles, conference proceedings etc.). For more detailed information on this issue, we recommend older pieces written by the author [84] and [85].

Opening up grey literature to the world

The application of the DSpace system allowed grey literature to be accessed within the university and at the same time opened it up to the world. The first stage of integration, however, had come earlier with the use of services of the Theses system. Given that plagiarism is part of the study agenda, the development of this interconnection continues in the STAG system. It now functions on an entirely automatic basis – full texts which are submitted are immediately recorded on Theses servers and checked, with the result of this check being displayed in the study system.

However, the connection of TBU to the network of repositories through the OAI is of greater benefit to the public, in that they offer fast and user-friendly access to grey literature. Among the many are the National Repository of Grey Literature[86], the Bielefeld Academic Search Engine[87] and the regional Zlín Libraries project[88].

Conclusion

The TBU digital library for providing access to e-UDT, scripts and other documents has been under construction for 10 years now. It experienced a number of teething problems and has now reached productive age. The services which it offers continue to grow, although not at such a rapid speed as with other projects at the TBU library. Its most valuable benefit for the Library itself is not the space saved in the fund or the decrease in unnecessary work with cataloguing, but the experience of the implementation team. The team is still dedicated to developing library services and can avoid certain mistakes … so that it has the opportunity to make new ones. The TBU digital library is available at this address: **http://digilib.k.utb.cz/**.

[84] BUDÍNSKÝ, Lukáš a Ondřej FABIÁN. Institucionální repositář jako nástroj podpory VaV. In: *INFORUM 2012: 18th Annual Conference on Professional Information Sources, Prague 22-24 May 2012* [online]. Praha: Albertina icome Praha, 2012 [cit. 2015-09-30]. ISSN 1801–2213. Available from: **http://www.inforum.cz/sbornik/2012/12/**.

[85] BUDÍNSKÝ, Lukáš. Cesta je ten nejdůležitější cíl – hledejme tu správnou v podpoře vědy a výzkumu. In: *INFORUM 2015: 21st Annual Conference on Professional Information Sources, Prague 26-27 May 2015* [online]. Praha: Albertina icome Praha, 2015 [cit. 2015-09-30]. ISSN 1801–2213. Available from: **http://www.inforum.cz/sbornik/2015/40**.

[86] **http://nusl.cz**

[87] **http://www.base-search.net/**

[88] **http://knihovnyzlin.cz/**

References

BUDÍNSKÝ, Lukáš a Ondřej FABIÁN. Institucionální repositář jako nástroj podpory VaV. In: *INFORUM 2012: 18. ročník konference o profesionálních informačních zdrojích, Praha 22.-24. května 2012* [online]. Praha: Albertina icome Praha, 2012 [cit. 2015-09-30]. ISSN 1801–2213. Available from**: http://www.inforum.cz/sbornik/2012/12/**.

BUDÍNSKÝ, Lukáš. Cesta je ten nejdůležitější cíl – hledejme tu správnou v podpoře vědy a výzkumu. In: *INFORUM 2015: 21. ročník konference o profesionálních informačních zdrojích, Praha 26.-27. května 2015* [online]. Praha: Albertina icome Praha, 2015 [cit. 2015-09-30]. ISSN 1801–2213. Available from: **http://www.inforum.cz/sbornik/2015/40**.

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ. *Jednotná formální úprava vysokoškolských kvalifikačních prací, jejich uložení a zpřístupnění*. Zlín: Univerzita Tomáše Bati ve Zlíně, 1.11.2004. Guidelines of the Rector No. 41/2004. Available from: **http://www.utb.cz/file/12033/**.

UNIVERZITA TOMÁŠE BATI VE ZLÍNĚ. *Jednotná formální úprava vysokoškolských kvalifikačních prací, jejich uložení a zpřístupnění*. Zlín: Univerzita Tomáše Bati ve Zlíně, 17.4.2006. Guidelines of the Rector No. 5/2006. Available from: **http://www.utb.cz/file/11946/**.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

SHARING LIABILITY FOR A REPOSITORY BETWEEN EMPLOYER AND EMPLOYEE

Michal Koščík

koscik@mail.muni.cz

Faculty of Law, Faculty of Medicine, Masaryk University, Brno, Czech Republic

Abstract

Each research institution that operates a repository has to make a decision whether to allow all its employees to upload their own works (the liberal model) or whether to create a special organizational unit that will review and approve each file shared via repository. This dilemma is accompanied with another important decision that an institution needs to make. Should the institution let its researchers to freely license their works to whichever publisher they choose, or should it apply a managing approach to publication activities? The post will outline legal challenges of both approaches and will formulate practical recommendations, how to formulate internal institutional norms that regulate the institutional repository.

Keywords

Open Access, Repository, Research publications, Research data

This paper was suported by the project "Právni rámec sběru, zpracování, uchovávání a užívání výzkumných dat (GA15-20763S)" of Czech science foundation.

Introduction – The motivation among institutes to set up a repository

More and more Czech public-sector institutions are setting up their own institutional repositories in order to share the results of their scientific and research activities. One of the main reasons for this wave is the adoption of the "Berlin Declaration" on open access to knowledge in natural sciences and humanities, to which the three largest producers of science publications in the Czech Republic have signed up in recent years. The aim of the Berlin Declaration is not to set out rigid, legally-binding and judicially-enforceable legal obligations. Instead the signatories sign up to the idea of publishing "Complete versions" of academic works and "all additional materials", together with a "public licence to use the work" in the repository, which is "supported and maintained by an academic institution, scientific company, government agency or other established organisation".

It must be said that the model of setting up one's own repositories is far more popular among research institutes than placing publications in central publication platforms. The motivation for establishing institutional repositories is to enhance the prestige of individual institutions. It must be said that grant policy in the Czech Republic also generates favourable conditions for the establishment of institutional repositories at a local level.

A dual approach to filling repositories

The way in which institutions fill the relevant institutional repository primarily differs in two parameters at individual levels:

1. Whether the storage of employees' works in the repository is compulsory or voluntary.

2. Whether the administration of individual pieces of output is entrusted to individuals or to a specialised department, for example an institutional library – this is related to the question of whether the author or the institution decides on publication within the open access regime.

Re. Parameter 1: Basic legal foundation concerning the (in)voluntary nature of filling the repository

Digital repositories which are filled on a voluntary basis and which rely on the spontaneous coaction of employees usually only contain an insignificant fragment of the scientific output of an institution.

The current trend is therefore an intensive motivation of employees to submit their results, by imposing the employees' obligations under labour law. The employer relies on the fact that it has the right within the Czech legal environment to exercise "copyright" to the works created by its employees and indeed has the right to impose on employees the obligation to strive for publication in selected periodicals or to keep the results of their work on record in institutional repositories. However, the rights of the employer specified above are not usually unlimited. In practice, the right of the employer might be limited by two circumstances in particular:

a) not all authors of a work are employees of the institution.

Research publications are often created by way of cooperation between research teams from several institutions. Doctorate students without any working relations to the institution become co-authors, or indeed people who are not employed at any research institution (typically, for example, a doctor involved in a clinical study). The employer clearly has no influence over these people and often relies that its own employees will obtain consent from outside co-authors. Mutual communication between co-authors regarding the publication of selected publications conceals many legal pitfalls, which will be considered later.

b) the employer has waived the right to the work or transferred this right to another.

The people that decide to publish in open-access regime and the authors themselves are often unaware of the many different ways that their employer can cede publication rights to third parties. For example, rights to output can even be waived much earlier than the researcher actually begins doing the work. One typical example is the situation in which the employer waives the right to publish results separately from other joint solvers or joint recipients before the start of a specific grant project. Several years usually pass from signing the grant agreement to the publication of results, whereby the wording of the agreement is not usually available to those that administer the institutional repository.

The employer usually transfers the right to exercise property rights to the publisher of the periodical in the interests of publishing results in prestigious magazines or simply waives the exercising of property rights. Certain institutions subsequently endeavour to get the rights back from the institution or try to negotiate conditions with the publisher under which the work can be accessed in open-access[89] regime, often using time embargos. The principle shortcoming of the open-access policy of certain research institutions is that they rest too much on procedures following the publication of an article, whereas it would be more productive to deal with questions of open access before the creation of an actual publication.

A somewhat extreme example of the opposite approach might be, for example, the open-access policy at Harvard University, which leaves the authors in no doubt that their work will be randomly published in the open-access regime if they do not seek a waiver from this policy before the publication of an article[90]. This also applies in the case that members of the academic community are merely co-authors of the chosen texts. Without this waiver, the authors have no right to grant the publisher exclusive rights. It must be said that a prestigious university can afford such an approach without any significant harm to its publication output. Research teams from less prestigious institutions, however, could be practically deprived of the possibility to publish in prestigious publications founded on the principle of subscription according to the rule of "Quod licet Iovi, non licet bovi", if this policy were to be strictly applied. It must be taken into consideration that the interest of research teams in the Czech Republic, for example, in publishing their output in foreign magazines is generally higher than the interest shown by foreign periodicals in publishing output from Czech institutions. These teams cannot therefore dictate the conditions under which their articles are published.

[89] Compare MYŠKA, Matěj, 2014. Vybrané právní aspekty otevřeného přístupu k vědeckým publikacím. *Právní rozhledy*. Vol. 22, issue 18, pp. 611–619. ISSN 1210-6410.
[90] Compare **http://hls.harvard.edu/library/for-faculty/open-access-and-scholarly-publishing/**

Re. Parameter 2: Is the administration of individual output entrusted to individuals or to a specialised department?

The question of eventual responsibility for filling the repository is dealt with differently by individual research institutions. Masaryk University, for example regularly leaves it up to the author of a paper whether to record the work in the repository and whether to publish it in the open-access regime. The university administration has the authority to moderate the manner of publication in the case that an employee publishes content illegally.

By contrast, the Academy of Sciences of the Czech Republic entrusts a specialised employee (an "administrator") from university administration with gathering data and subsequently publication. The author of a text may enter it in the repository himself and indicate his/her wish regarding the manner of publication and the processor confirms the correctness of the settings of the parameters for saving the full text ("If the author saves the file, a trained processor checks the setting of the file and only after this check is the file made accessible[91]").

Both approaches have their pitfalls. The first approach almost unwisely relies on the initiative of the authors and transfers to the authors, almost like passing the buck, the risk of possible breach of the rights of third parties, in particular if such publication is accompanied by a public licence[92]. On the other hand, the centralised approach might be inflexible in the case of co-authors from other institutions, when the repository administrator reaches agreement with co-authors with far greater difficulty than an active member of a research team. A system based on the initiative of the authors must deal with the problem that the person who decides on the publication of the work need not always be a person authorised to have disposal of such a piece of work. On the other hand, a system based on the initiative of the administrator must deal with unclear communication between the authors and the repository administrator.

The most common errors when publishing work

One of the most serious mistakes is, of course, publishing work to which third persons have rights. The explanation below points to errors which individual researchers or administrators might make when publishing work within the regime of open access.

1. Co-authors are not authorised to grant consent to publication in the repository

Plenty has already been written about this problem and research and administrative workers have become far more legally aware in the past five years. Institutions make considerable efforts to prevent legal disputes by negotiating with publishers as the organisations they are most willing to litigate. The aim of this chapter is to point to a number of facts that might be regularly overlooked in practice, even in an effort to proceed in the correct way with regard to the rights of third parties.

[91] See. Maunal of myASEP (repository of Academy of sciences) **https://www.lib.cas.cz/asep/repozitar-asep/vytvoreni-uctu-myasep/** . cit. 19.10.2015

[92] The author considers this in more detail in other articles: KOSCIK, Michal; SAVELKA, Jaromir. Dangers of over-Enthusiasm in Licensing under Creative Commons. Masaryk UJL & Tech., 2013, 7: 201.; KOSCIK, Michal. Creative Commons Will It Do Good in the Czech Republic. Masaryk UJL & Tech., 2008, 2: 61.

In the event that a publication is created by the research teams of several institutions, it is common to ask for the consent of the co-authors before publication in the repository[93]. Here, however, it is important to be aware that there is the considerable likelihood that the co-authors will not be authorised to grant consent because their work on the publication is also work for an employer, to whom the relevant rights truly pertain.

In some cases, the fact that the author (or one of the authors) is an employee of several institutions at the same time might also cause complications, in that such an employee need not actually be fully aware for which of them he/she carried out the specific scientific task, seeing the publication as that of the knowledge he/she acquired in the past.

2. Different understanding of "academic co-authorship" and co-authorship according to the Copyright Act

It is important to be aware that people who take part in research, for example by coming up with an innovative idea, carrying out a large number of measurements, collecting a large sample of patients etc., are generally presented as co-authors. These people, however, are not always the authors of the text itself, which is generally written and edited by younger colleagues. Copyright protects an expression and not the idea itself and a detailed analysis might show that a person correctly stated as a "co-author of the result" in the sense of citation ethics is not a co-author in terms of copyright. If it is not the work of a fictitious author in the sense of copyright regulations, then not even an institution may exercise rights to it as the work of an employee.

3. The copyright problem with Ph.D. students

Consideration is often made of the fact that students are not necessarily employees and that there is no transfer of authorisation to exercise property rights. Students are not employees and their work can only be used "for teaching or for the internal needs" of the institution. This statutory licence is insufficient for the purposes of open-access publication and the consent of students must be treated within a special regime.

4. Copyright associated with typesetting and the issue of preprints

When publishing books and articles, copyright does not relate solely to the words, but to the typesetting, graphics or illustrations, which need not be the intellectual property of the same author. Such graphic elements are usually created by independent specialists working according to the wishes and at the expense of the publisher. An institution cannot therefore automatically share a graphic copy of works already having been published in the open-access regime.

Sharing preprints, meaning the manuscripts of the author without any modifications, presents itself as a solution to this. Nonetheless, it is important to be wary of the myth that preprints do not enjoy legal protection. Only the question of copyright to graphics is really resolved by

[93] For example, the methodology of the AS CR published at: **https://www.lib.cas.cz/asep/repozitar-asep/** "Co-authors may be asked for consent to their work being made accessible within the regime of open access".

publishing preprints. However, the publication of preprints could be at odds with the law in the case that an exclusive licence exists for the publisher to the text in question.

Conclusion

The conclusion can be reached that there are two fundamental risks for institutional repositories. The first risk is the author the second is the institution. The latter might get it wrong, even in efforts to act in good faith and respect the rights of the authors.

If the aim of the Berlin Declaration is to see a shift change in making scientific results accessible, the author would also take the liberty of considering a possible shift change in adding to existing repositories.

Instead of institutions trying to fill repositories by negotiating with licencing houses and forcing their employees to record specific works in local repositories, institutions should concentrate on motivating their employees towards primary publication in open-access sources. It is essentially unimportant in open-access publication whether the publication is actually stored in the repository of the publisher or of the author's institution. In places in which institutional repositories already exist, these would concentrate on obtaining copies of texts already having been published under free licences without the coaction of the actual authors of the publication. Negotiation of time embargos and the publication of domestic authors within the regime of "green open access" would be transferred to a central institution (library), which would agree with the publishing houses on open access to selected publications by domestic authors for IP addresses in the Czech Republic.

The system outlined above would minimise the legal risks caused by action by the authors of a publication which is frequently unpredictable, it would reduce the costs of the institution of administering intellectual property rights in the repository and it would, moreover, achieve considerable savings in scope in light of the central purchase of rights for "green open access".

References

BUREŠOVÁ, Iva. Otevřený přístup (Open Access). In: Pavla KOVÁŘOVÁ ed. *Trendy v informačním vzdělávání*. 1. ed. Zlín: VeRBuM, 2012, p. 51-60. ISBN 978-80-87500-18-7.

BUREŠOVÁ, Iva. Otevřený přístup (Open Access) v Akademii věd. *ITlib* [online]. Bratislava: CVTI SR, 2013, vol. 17, no. 3, p. 10-15 [cit. 2015-12-07]. ISSN 1335-793X. Available from: **http://itlib.cvtisr.sk/buxus/docs/10_OTEVRENY%20PRISTUP.pdf.**

AKADEMIE VĚD ČR. *Dotace pro podporu publikování formou Open Access. Knihovna AV ČR, v. v. i.* [online]. Praha, 2013 [2013-05-21]. Available from: **http://www.lib.cas.cz/openaccess.**

AKADEMIE VĚD ČR. Politika otevřeného přístupu AV ČR. *Akademie věd České republiky* [online]. Praha: Akademie věd České republiky, 2013 [cit. 2015-12-07]. Available from: **http://www.cas.cz/o_avcr/zakladni_informace/dokumenty/politika-otevreneho-pristupu.html**.

KOŠČÍK, Michal a Jaromír ŠAVELKA. Dangers of Over-Enthusiasm in Licensing under Creative Commons. *Masaryk University Journal of Law and Technology* [online]. Brno: Masarykova univerzita, 2013, **7**(2): 201-227 [cit. 2015-12-07]. Available from: **https://journals.muni.cz/mujlt/article/view/2633/2197**.

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

MYŠKA, Matěj, Libor KYNCL, Radim POLČÁK a Jaromír ŠAVELKA. *Veřejné licence v České republice*. Brno: Tribun, 2012. ISBN 978-80-263-0343-5.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

MAKING DATA IN PHD DISSERTATIONS REUSABLE FOR RESEARCH

Joachim Schöpfel

joachim.schopfel@univ-lille3.fr

GERiiCO laboratory, University of Lille 3, France

Hélène Prost

helene.prost@inist.fr

INIST (CNRS), France

Cécile Malleret

cecile.malleret@univ-lille3.fr

Academic library, University of Lille 3, France

Abstract

How can an academic library contribute to make data submitted together with PhD dissertations useful for further research? Our paper provides some recommendations for information professionals, based on a review of studies and projects and on empirical evidence from a content analysis of data sources and types from 300 print and digital dissertations in social sciences and humanities (1987-2013) and a survey on data management conducted with the scientists and PhD students of the University of Lille 3 in April and May 2015.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

Keywords

PhD Dissertations, Research Data, Data Management, Open Access, Digital Humanities, Social Sciences and Humanities

Dissertations and data

Open access to PhD dissertations[94] is on the agenda of academic libraries. Today, the rapid development of data-driven research (e-Science) and the debate on open data and re-use of research results has led us to discover another challenge in the field of PhD dissertations, beyond the debate on open access and embargo, i.e. the existence of large amounts of data produced by the PhD candidate and partly submitted together with the text of the dissertation. How can these data be made available in the context of open access and open data policies, what are the potential barriers for dissemination and reuse, and how can academic libraries contribute to this challenge?

Research results produced by PhD students could contribute to data-intensive scientific discovery (Schöpfel et al. 2014). The PhD dissertations could become "windows for the scientist" not only to understand but also to reproduce and extend scientific results (Lynch 2009), in so far as they could integrate research data that could be enriched, updated, extracted, shared, aggregated and manipulated (McMahon 2010). They could become live documents. However, there are two barriers.

First, dissertations must be freely available in open access, deposited in institutional or other repositories and disseminated with sufficient user rights to allow reuse. So far, the reality is mitigated. On the one hand, more than half of all open repositories contain theses and dissertations and the technical and political environment globally supports open access to academic works. On the other hand, a significant portion of the digital dissertations are not online, not open, not freely available but embargoed or under restricted access (Schöpfel et al. 2015).

The second barrier is the fact that research data related to PhD dissertations are largely "dark data", i.e. "data that is not easily found by potential users (…), unpublished data (and) research findings and raw data that lie behind published works which are also difficult or impossible to access as time progresses" (Heidorn 2008, pp.281 and 285). They are, in other terms, "hidden treasures".

These data, defined as reusable research results, collected, observed, or created for purposes of analysis to produce original research results, are produced in a large variety of formats, sources and types. Research results may be presented as tables, graphs, etc. in the paper or as additional material (appendix). In the past, print theses and dissertations have regularly been submitted together with supplementary material and data, in various formats and on different supports (print appendices, punched cards, floppy disks, audiotapes, slides, CD-ROMs). In the new ETD infrastructures, such material can be processed together with the text

[94] In the following we shall use the term "PhD dissertation" to designate the document submitted in support of candidature for the academic degree of doctorate, as synonym for "PhD or doctoral thesis".

files or as supplementary files in different formats, depending on disciplines, research fields and methods. But as the ETDplus project at the Educopia Institute at Atlanta, Georgia, states, these "complex digital objects (e.g., software, multimedia files, digital art, and other material that sometimes is integral to the thesis or dissertation itself…)" are often not collected or preserved[95]. Sometimes the data are available on a distant server. And too often the data are simply not available; or data, methodology, tools, primary sources are mingled, not or badly indexed, or not linked with the text.

Description and preservation of digital objects are part of the work of traditional academic libraries. For this reason, they generally consider research data curation and management as a new challenge, a kind of new frontier for the development of their campus services, either on a local level or as part of a scientific network (CLIR 2013). For the same reason, we started to work on the topic from 2013 on. Empirical results and recommendations are based on our research at the University of Lille 3, a large social sciences and humanities campus in the Northern part of France, with 19,000 students and nearly 500 PhD candidates in three graduate schools and 55 doctoral degrees. The project is going on.

Sources and types of data

In order to find out more about the data deposited by PhD students, we conducted a survey on 283 dissertations from 1987 to 2013 from the University of Lille 3, covering nearly the whole range of all disciplines on the campus[96] and representing about 30% of all dissertations of that period. 88 were digital (31%) and 195 print dissertations (69%). 188 dissertations contain one or more appendices, i.e. documents attached to the end of a dissertation, with some kind of research data (66%)[97]. The length of these appendices varies widely, from 5 to 829 pages, with a median of 81 pages, and totalling more than 25,000 pages. All disciplines have appendices with data but some disciplines such as History of Art, Education, Archaeology and Egyptology, "produce" rather large appendices, while others, like Psychology or Philosophy, often contain shorter appendices (Figure 1).

[95] **http://educopia.org/research/grants/etdplus**

[96] In our sample, History, Psychology, Philosophy, Foreign Languages and Literature (English and American, Spanish, Slavonic, Hebrew…), Information and Communication Sciences (including Library Sciences), History of Art, Linguistics, Archaeology and Egyptology were the most represented disciplines.

[97] NB: some pages contain empty questionnaires or survey forms, experimental procedures, bibliographies etc. which cannot be considered as data.

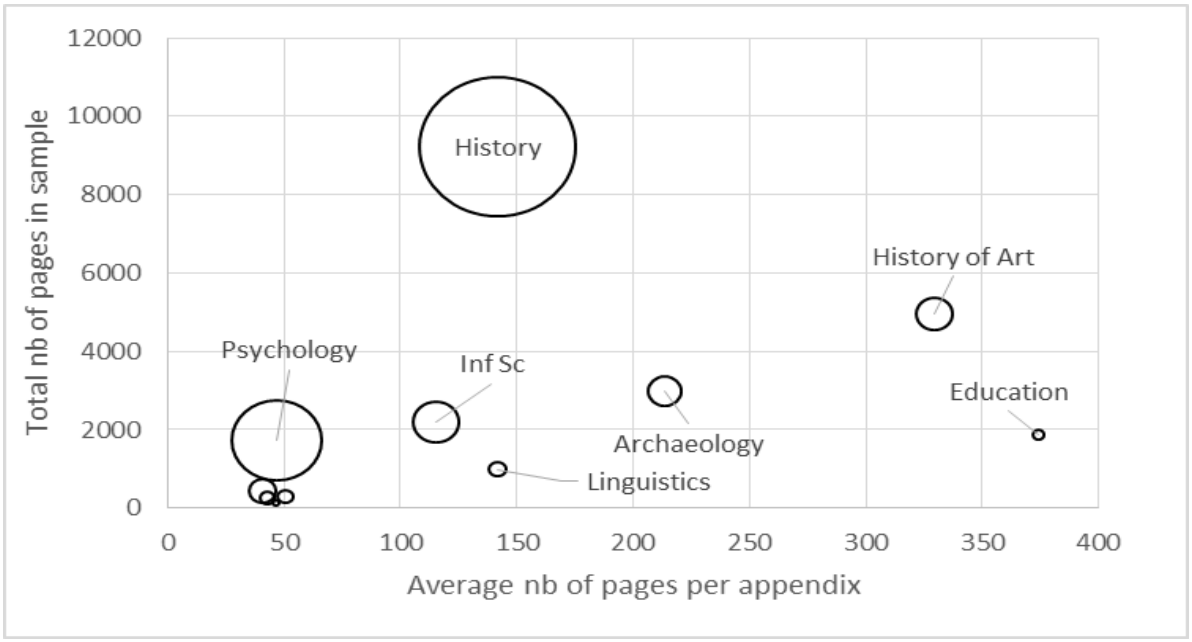

Figure 1 Size of appendices (circle size = number of dissertations, N=188)

For our survey, we tried to distinguish between data sources and research results. This is not always simple, as the concept of research data depends on scientific fields and methods and not clearly discernible from data sources. For instance, are photographs of archaeological inventories primary sources students used for their analysis, or results of their research, or both? Our approach was to identify and describe types of research data that are potentially reusable which means that they may become, together with the dissertation, sources of further research and future research data.

The PhD students used a wide variety of sources for their scientific work. We identified three major data sources, i.e. archives, surveys and interviews and text samples (Figure 2).

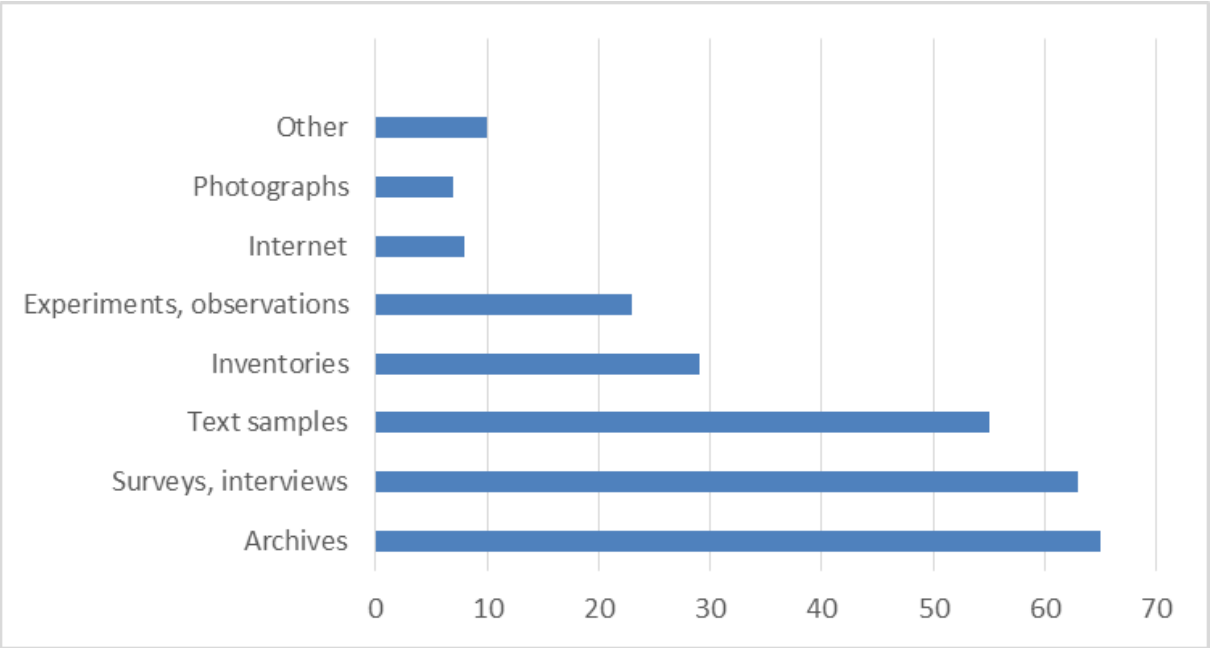

Figure 2 Data sources per dissertations (N=188)

Other less exploited sources are inventories, experiments and observations, the Internet and photographs. The distribution of data sources is to some extent specific for each discipline. Here are some examples of heavily used sources:

History: archives, text samples
Psychology: surveys, experiments
Information and Communication Sciences: surveys, text samples, the Internet
Archaeology and Egyptology: inventories, photographs

These are typical research data sources for the social sciences and humanities. Compared to other surveys, data sources like observations, simulations, statistics, reference data or log files (usage data) are unusual or missing.

As for the research data present in the appendices, our survey reveals several different and heterogeneous data types (Figure 3).

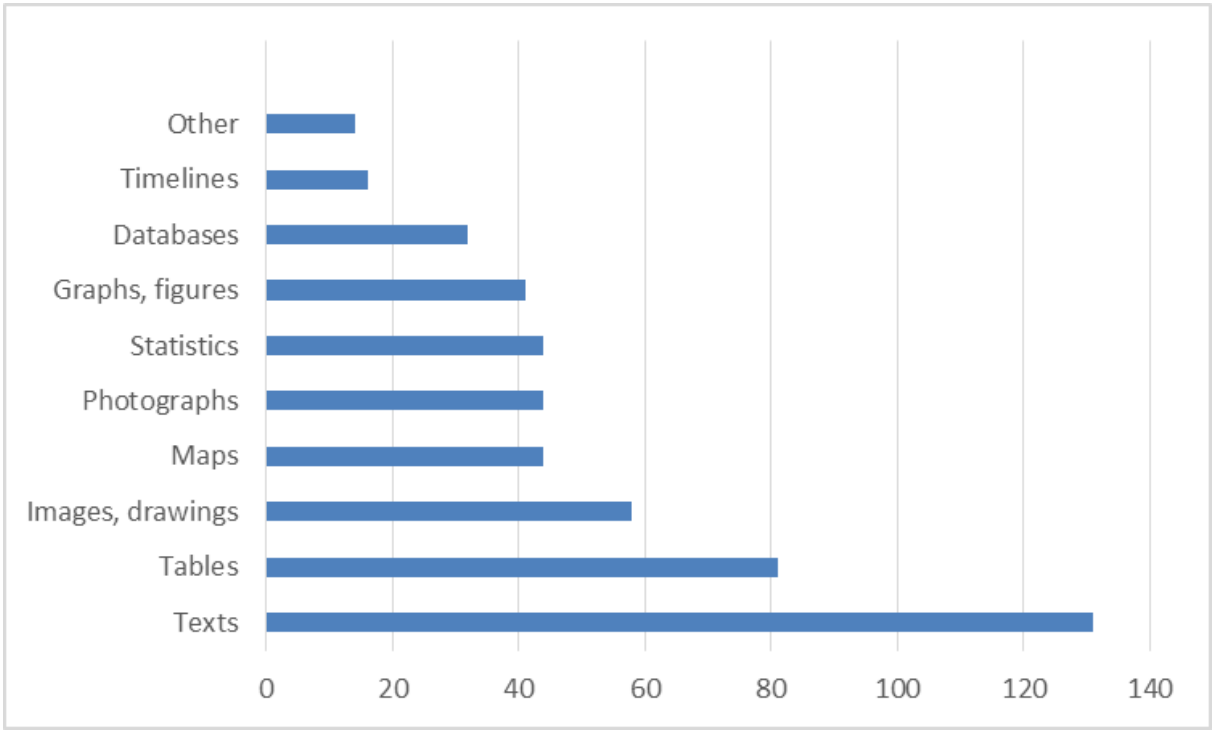

Figure 3 Data types, per dissertations (N=188)

Again, text samples are the most important data type, followed by spreadsheets, images and drawings, maps, photographs, statistics, graphs, databases and timelines (chronologies). We only found one dissertation with audio-visual media (recorded interviews) and we have not found any dissertation with geolocation data. In some disciplines, one or two data types are predominant. This is the case in Philosophy, Linguistics and Foreign Languages and Literature where text samples represent more than half of the data. Other disciplines are characterized by a wide number of different types of research data. Some examples:

History: ten different data types, including text (21%), tables (15%) and images (14%).
Information and Communication Sciences: ten different data types, including text (33%), tables (17%) and graphs (13%).

Psychology: nine different data types, including tables (29%) and statistics (28%).
Archaeology and Egyptology: nine different data types, including photographs (21%), maps (17%) and images (17%).

Some data types are present in all disciplines, like text samples, images, tables or graphs and figures. Others, in particular inventories or audio-visual material, are at least in our sample specific for one or two disciplines.

Medium and format of data

While 30% of the print dissertations clearly separate text and appendices in different volumes, digital dissertations do not often separate text and appendices but glue them together into the same file (52%). Also, some dissertations have poor or no table of contents for their appendices. All files of digital PhD dissertations must be deposited with the text, and the French national computer centre for Higher Education[98] maintains a list of accepted file formats for long term preservation[99]. However, nearly all files in our survey are in PDF (image or text), and other formats are very rare. For instance, in our sample only one dissertation has been submitted with video and audio files on CD-ROM. Other data are available on a distant web site or deposited on compact disc, DVD or USB flash drive. Dissertations in History, especially for studies on historical social groups, sometimes contain detailed and well-structured biographical information, presented like a database. One example for this "prosopographical" approach: a dissertation on the Renaissance elite of the old Flemish town of Douai[100] with biographical records of 423 aldermen, with structured information about, among others, place and date of birth, date of death, mandate period, noble titles and occupation.

Data management

Many, if not all of these data could be of real value for further research. These data could be used to create image databases, digital maps collections or digital libraries with manuscripts, archival material and other text samples open for text mining tools. Results from experiments and surveys could be published in a way that allows for reuse, data mining and automatic meta-analysis on different datasets. Research results could thus become new data sources and generate further research. However, this potential reuse requires data management and curation to remain accessible and interpretable over time, including metadata and long-term preservation (Neuroth et al. 2013). For young scientists and PhD students, learning how to design and implement a data management plan (DMP) is even more important in so far as more and more funding bodies evaluate the existence and quality of DMPs in research project proposals. Our empirical data do not tell us if the PhD students conducted a data management plan. But only few dissertations demonstrate a real effort of data management and curation. In particular, our study reveals three barriers to open data:

Incomplete, inadequate or missing description of the whole datasets and/or individual data.

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

[99] **https://www.cines.fr/archivage/des-expertises/expertise-formats/liste-des-formats-archivables/**

[100] Duquenne, F. (2011). *Un tout petit monde : les notables de la ville de Douai du règne de Philippe II à la conquête française (milieu du XVIe siecle-1667) : pouvoir, réseaux et reproduction sociale*. Université de Lille 3.

> Missing organisation. Research data are presented without any structuration or organisation, often together with other, not reusable material in a kind of information mash-up not suitable for further research.
> Inadequate format. Data and text are glued together in a PDF file instead of being separated and published in adequate file formats.

In a second survey on research data management and sharing at the University of Lille 3 (Prost & Schöpfel 2015), PhD students represented 33% of the whole sample of 270 scientists. Compared to professors, senior lecturers etc., they have less experience with data management. They all store their data on the hard disks of their personal computers, sometimes also on a computer at the research laboratory or department, with back-ups on an external device like hard drive, USB flash drive or DVD, and sometimes even in the cloud (Dropbox). This is more or less personal knowledge management, good enough for personal research work and small projects but not compatible with larger research projects, such as the European H2020 programme. Also, they do not delegate this management. The Lille PhD students are not really different compared to other universities, as other survey results show[101] - many PhD students are interested in data management and to some extent in support of sharing at least some data but have little or no experience at all.

Data sharing

Our survey on research data at the University of Lille 3 confirms that PhD students have less experience with data sharing, which is not surprising as they are at the very beginning of their scientific career. More than other scientists, they often simply do not know options and opportunities for the deposit and sharing of their research results. Yet, 30% of them declare that other persons of their research team have access to their own data. This is a basic way of data sharing, not on the Internet but on their computers or via flash drives, Dropbox, the University Intranet etc. Also, they are more interested in reuse of data from other researchers than other categories.

Nearly one third (28%) of the students do not want to make their data available in the future or at least hesitate, which is the same part as for other scholars and researchers. Yet, they show a significantly higher motivation to deposit their research results in a data repository (63% compared to 43%), even in a local repository (laboratory, department) while the other scientists clearly prefer international and domain-specific sites. When asked which kind of service they would need, they ask for technical advice and help for data management plans for the publishing of their results.

More than the elder staff, they also ask for assistance in ethical and legal issues. As a matter of fact, privacy issues and third party copyright are two serious legal problems that need awareness. Our survey on PhD theses reveals two issues:

> Some appendices contain personal data, about living or dead people, historical persons or unknown (anonymous) people. These may be survey data, experiments, interviews, biographies etc. In so far as the information allows identifying individual persons, they need special processing and careful handling.

[101] See for instance Simukovic et al. (2014) and a recent, unpublished survey from the University of Strasbourg.

> Some dissertations contain material that is protected by copyright and cannot be reproduced or disseminated without authorization, even by fair use or copyright exceptions (short citation, research…). These may be text samples, maps, photographs, copies from books etc. – material not created by the PhD student him/herself.

These problems should be addressed as a part of PhD education on data management, well ahead of decisions on preservation and dissemination.

Recommendations

Advice and assistance will be necessary for PhD students to prepare their data in an adequate way. Adequate means at least:

Clear separation of text and data. Digital research data must be submitted in different and separate files.

Structuration of the research data, with a detailed and organized tagging (markup) of the datasets.

Metadata of good quality. The data must be described in a standard language and format, with sufficient detail for retrieval and data mining.

Deposit in original format. Data should be submitted in their original and if possible, open format (and not in PDF), to facilitate long-term preservation and reuse.

Clearing of privacy and copyright issues.

The empirical evidence of this study suggests that assistance and advice for PhD students to help them manage their research data must go beyond general rules and recommendations. Not all doctoral projects produce research data. Not all data are submitted with the dissertation to back up the research in the dissertation or to further explain and clarify the matter. Not all data can be reused especially, but not only, for legal reasons. And finally, even if our sample is not representative, it seems obvious that many characteristics of data sources and types have strong relationships with disciplinary methods, topics and approaches.

Following the work of Reznik-Zellen et al. (2012) at the University of Massachusetts Amherst, we develop three tiers of research data support services for PhD students on our campus, including education, consultation and infrastructure (Figure 4).

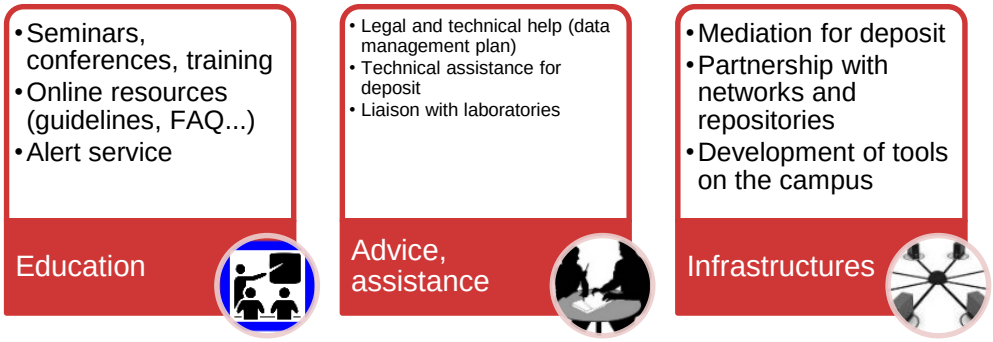

Figure4 Research data support services

Education: We already organized three events on research data especially designed for PhD students in social sciences and humanities, and we will launch a first doctoral seminar in October 2015 on data management and sharing. In the future, we will edit or adopt guidelines and make them available for the PhD students, together with frequently asked questions and updates on data management, open data etc.

Advice and assistance: Probably as a part of the future Learning Centre of the University of Lille 3, we will develop personalized help and assistance for PhD students, able to provide answers and advice to their specific questions and problems.

Infrastructures: Our approach is based on intermediation, not on research and development. Even if we will probably develop some basic tools for temporary storage and metadata on our campus, our main idea is to partnership with existing data networks and repositories, including agreements if necessary and delegation of the deposit.

These three tiers of research data support services will be launched progressively between 2015 and 2018. Their development will follow five guiding principles (Figure 5).

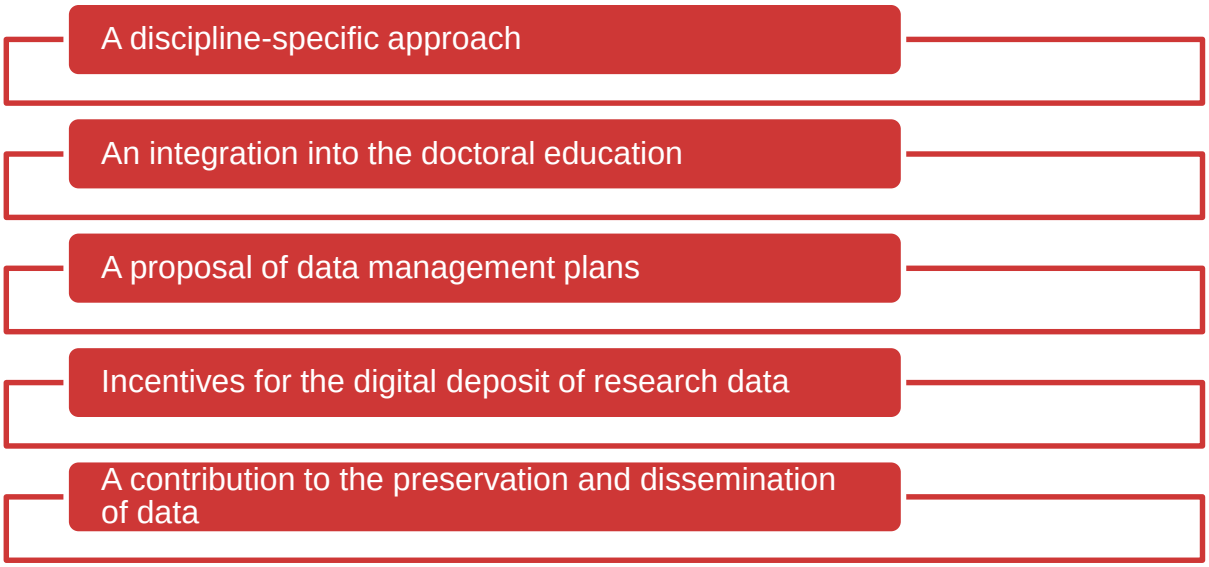

Figure 5 Five leading principles for the implementation of research data support services

One size does not fit all. Research data support services must be flexible and adjusted to the scientific disciplines and domains of the PhD research. This means a very good knowledge of research methodologies, data types, formats etc. but also a good cooperation with the research teams, large projects and laboratories.

Data management and sharing must become part of the mandatory doctoral education syllabus, such as project management, scientific writing or data analysis.

The University of Lille 3 will develop its own templates for data management plans, in line with social sciences and humanities and be compatible with the criteria of the European research projects.

Deposit of research data along with PhD dissertations should become near to mandatory. At least, there should be strong incentives to submit those data for temporary storage and long term preservation.

Finally, as mentioned above, our University will contribute to the preservation and dissemination of these research data – not necessarily with campus-based infrastructures (they are not excluded, though) but rather through partnerships and networking with local or national providers. We are already doing so in the field of open access, with good success, as our institutional repository is hosted by the Lyon-based CCSD[102] and part of the national open repository HAL[103].

The academic library, already present and engaged in ETD management and open access, will be a leading partner for these new research data support services, in cooperation with the graduate school and the research laboratories. Nevertheless, this leading position must become legitimate and accepted by the scientific community and the PhD students. So far, following our campus survey on data management and sharing, scientists and students have not identified the academic library as a potentially useful structure for their data. In other words, the implementation of the new services must be accompanied by communication about the role and usefulness of each partner, and by the acquisition of new skills and knowledge by the information professionals.

References

CLIR (2013). *Research Data Management: Principles, Practices, and Prospects*. Report, Council on Library and Information Resources, Washington D.C. Available from: **http://www.clir.org/pubs/reports/pub160**.

HEIDORN, P. B. (2008). Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*. Baltimore: Johns Hopkins University Press, **57** (2), p. 280-299. Available from: **https://www.ideals.illinois.edu/bitstream/handle/2142/10672/heidorn.pdf?sequence=2.**

[102] **https://www.ccsd.cnrs.fr/**
[103] **http://hal.univ-lille3.fr/**

LYNCH, C. (2009). Jim Gray's Fourth Paradigm and the Construction of the Scientific Record. In: HEY, T., S. TANSLEY, & Tolle, K. (Eds.). *The Fourth Paradigm. Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Corporation, 2009, pp. 177-183. Available from: **http://research.microsoft.com/en-us/collaboration/fourthparadigm/.**

MCMAHON, B. (2010). Interactive Publications and the Record of Science. *Information Services and Use* [online]. IOS Press, **30** (1), p. 1-16. ISSN 0167-5265. ISSN 1875-8789. Available from: **http://iospress.metapress.com/content/f4th457822023783/fulltext.pdf.**

NEUROTH, H., S. STRAHMANN, A. OSSWALD, and J. LUDWIG (Eds.) (2013). *Digital Curation of Research Data. Experiences of a Baseline Study in Germany*. Glückstadt: Verlag Werner Hülsbusch. ISBN 978-3-86488-054-4. Available from: **http://www.nestor.sub.uni-goettingen.de/bestandsaufnahme/Digital_Curation.pdf.**

PROST, H. & J. SCHÖPFEL (2015). *Les données de la recherche en SHS*. Une enquête à l'Université de Lille 3. Rapport final. Université de Lille 3, Villeneuve d'Ascq.

REZNIK-ZELLEN, R., J. ADAMICK, & S. MCGINTY (2012). Tiers of research data support services. *Journal of eScience Librarianship*. **1** (1), p. 27-35. ISSN: 2161-3974. Available from: **http://dx.doi.org/10.7191/jeslib.2012.1002.**

SCHÖPFEL, J., S. CHAUDIRON, B. JACQUEMIN, H. PROST, M. SEVERO, & F. THIAULT (2014). Open Access to Research Data in Electronic Theses and Dissertations: An Overview. *Library Hi Tech.* Emerald, **32** (4), p. 612-627. ISSN 0737-8831. Available from: **http://www.emeraldinsight.com/doi/abs/10.1108/LHT-06-2014-0058.**

SCHÖPFEL, J., H. PROST, M. PIOTROWSKI, E. R. HILF, T. SEVERIENS & P. GRABBE (2015). A French-German survey of electronic theses and dissertations: Access and restrictions. *D-Lib Magazine*, **21** (3/4). ISSN 1082-9873. Available from: **http://www.dlib.org/dlib/march15/schopfel/03schopfel.html.**

SIMUKOVIC, E., M. KINDLING, & P. SCHIRMBACHER. (2014). Unveiling Research Data Stocks: A Case of Humboldt-Universität zu Berlin. In: *iConference, 4-7 March 2014, Berlin.* P. 742-748. Available from: **http://hdl.handle.net/2142/47259**.

The paper is a shortened and updated version of the following article: Prost, H., C. Malleret, J. Schöpfel, 2015. Hidden treasures. Opening data in PhD dissertations in social sciences and humanities. *Journal of Librarianship and Scholarly Communication.* Pacific University Libraries, **3** (2), eP1230. E-ISSN 2162-3309.

All websites were accessed in August 2015.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

CHALLENGES IN PROVIDING UNPUBLISHED RESEARCH DATA IN BIOMEDICINE TO GREY LITERATURE REPOSITORIES

Pavla Francová[1], Stephanie Krueger[1,2]

pavla.francova@techlib.cz, stephanie.krueger@techlib.cz

[1] National Library of Technology, Prague, Czech Republic

[2] University of Chemistry and Technology, Prague, Czech Republic; Humboldt-Universität zu Berlin

Abstract

Regardless of the scientific field or focus, every researcher produces a multitude of unpublished research data during his or her career, including project diaries, project proposals, laboratory notes, and also the outputs of collaboration between researchers – for example, images, video, and measurements. Such "ephemeral" data can be crucial in inspiring new research directions and perspectives, and is often currently not shared in open repositories, although making such data more accessible undoubtedly has value for researchers. While not "literature" according to the traditional definition, such contextual materials and data – at least in bioengineering and biophysics – are the de facto bases for any grey literature produced in such fields and are directly relevant when discussing the utility of a grey literature repository in relation to such research endeavors. In this paper, the authors describe the difficulties posed in searching for grey literature and data on a specific bioengineering topic, Magnetic Resonance Imaging (MRI) of lung structures using the Magnetization Transfer Contrast

method, and also provide examples of unindexed grey literature and data produced by scholars in this field.

Keywords

Grey Literature, Data Repository, Dark Data, Research Data, Magnetic Resonance Imaging, Magnetization Transfer Contrast, Biomedical Engineering, Biomedicine, Lung

---

Introduction

In this paper, the authors describe the difficulties posed in searching for grey literature on a specific bioengineering topic, Magnetic Resonance Imaging (MRI) of lung structures using the Magnetization Transfer Contrast (MTC) method. They also provide examples of unindexed grey literature and data produced by scholars in this field. Via this method, the authors combine (in case study format) the perspectives of information scientist and active researcher in order to address the broader questions about the current status of the accessibility of scholarly outputs in bioengineering based on a real-life example of information use and retrieval in this field, together with a description of scholarly outputs invisible to the outside world prior to the creation of what might be considered grey literature (e.g., conference paper or pre-print).

The authors identify a lack of accessibility to grey literature and data in this specific field which pose challenges for conducting research into this specific bioengineering topic and provide a list of possible areas in which libraries and/or scientists themselves might enhance accessibility to research outputs in the future.

The Bioengineer's Perspective: Searching for Literature and Data (including Grey Materials)

In this section, the authors describe a "real-life" search for literature and data as an example of information retrieval from the perspective of a researcher in the field of bioengineering utilizing the topic "Magnetic Resonance Imaging (MRI) of lung structures using the Magnetization Transfer Contrast method." The description of the development of a search query and the information retrieved represents an actual research session conducted by one of the contributing authors in October 2015 during a recent extension of her research project which was accepted at the University of Würzburg's Pulmonary Imaging Network (PINET) research group under European Union grant FP7-ITN and Marie Skłodowska-Curie Actions (MSCA) grant PITN-GA-2010-264864. The description of this session provides a starting point for future research into the accessibility of grey literature for scholars conducting research in this topical area. It is intended to address, by means of a real example, the paucity of understanding of the information-related behaviors in the sciences by describing, from the emic perspective of an actual working scientist, the steps involved in conducting research in her field. [A]

The following information retrieval steps (A-D) represent, according to this bioengineer, the way in which she typically conducts a search for literature and data, in which she begins her search using published scholarly outputs and – only in Steps C and D – opens up her search to include the outputs of pre-prints, grey literature, and dark data [B], the so-called "long tail of science" [C, D].

- **Pre-Step**: Develop search query
- **Step A**: Search fully-indexed bibliographic databases for relevant information and data
- **Step B**: Search for full-text in fully-indexed databases for relevant information and data
- **Step C**: Search institutional repositories (fully- or partially-indexed; this varies) for relevant information and data
- **Step D**: Search other grey and dark data repositories and resources.

In the following sections, the authors describe how the researcher typically creates a query set and the results of applying query variants to different kinds of data sources (i.e., steps A-D). For each step, the authors then comment on the availability of grey literature and data, from the perspective of the researcher.

Developing a Search Query Set

In medicine and biomedical engineering, researchers in this researcher's field are – in her opinion – not usually familiar with the MeSH (Medical Subject Headings) thesaurus, a controlled vocabulary for indexing articles utilized by PubMed (1[104]) (the free search engine for accessing the MEDLINE database provided by the United States National Library of Medicine and the US National Institutes of Health). However, the authors decided in this case study to use the MeSH controlled vocabulary to make their search queries more precise.

For this example, the researcher mapped her initial keyword set (Lung structure, Magnetic Resonance Imaging, and Magnetization Transfer Contrast Imaging) to the MeSH subject headings: "**Magnetic Resonance Imaging**", "**Lung,**" and "**Magnetization Transfer Contrast Imaging**" as well as several variants of the latter. The authors then utilized these subject headings for subsequent searches in different data sources. The variants of "**Magnetization Transfer Contrast Imaging**" allowed the authors to analyze the effect of phrase variants on the number of results for each resource.

[104] Available from **http://www.ncbi.nlm.nih.gov/pubmed**

A	"**Magnetic Resonance Imaging**" AND "**Lung**"
B	"**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**"
C	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**"
V1	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**"
V2	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**" OR "**Magnetization Transfer Imaging**"
V3	"**Magnetic Resonance Imaging**" AND "**Lung**" AND "**Magnetization Transfer**" OR "**Magnetization Transfer Contrast**" OR "**Magnetization Transfer Imaging**" OR "**Magnetization Transfer Contrast Imaging**"

Table 1 Search terms developed using MeSH for the topic "Magnetic Resonance Imaging", "Lung," and several variations of "Magnetization Transfer Contrast"

Step A: Bibliographic Databases

Using these keywords, the authors retrieved results in the following bibliographic databases which are licensed by the National Library of Technology for its registered patrons (as is the case for the full-text databases mentioned in Step B below).

Resource Used	Query A	Query B	Query C	Query V1	Query V2	Query V3
PubMed	8 275	2 213	17 [1-3]	225	617	617
SCOPUS	25 231	3 002	12 [1-4]	12	12	12
Web of Science - title	351	1 398	1 [3]	147	395	395
Web of Science - topic	3 293	3 462	10 [1-3,5]	458	1 161	1 161

Table 2 Number of results for queries A-V3 from Table 1 in key bibliographic databases. Phrase C provided the most relevant found articles. Search conducted October 2015.

Phrase C provided the most relevant articles according to the researcher, though the authors deemed only five as being directly relevant to the research topic. Unfortunately, no other kind of grey literature except the conference papers was included in these sources.

[1] ARNOLD, J.F.T. , M. KOTAS, R.W. PYZALSKI, E. D. PRACHT, M. FLENTJE, and P. M. JAKOB. Potential of magnetization transfer MRI for target volume definition in patients with non-small-cell lung cancer. *Journal of Magnetic Resonance Imaging* [online]. 2008, 28(6): 1417-1424 [cit. 2015-10-19]. DOI: 10.1002/jmri.21436.

[2] JAKOB, P.M., T. WANG, G. SCHULTZ, H. HEBESTREIT, A. HEBESTREIT, M. ELFEBER, D. HAHN, and A. HAASE. Magnetization transfer short inversion time inversion recovery enhanced1H MRI of the human lung. *Magma: Magnetic Resonance Materials in Physics, Biology, and Medicine* [online]. 2002, 15(1-3): 10-17 [cit. 2015-10-19]. DOI: 10.1007/bf02693839.

[3] KUZO, R.S., M.J. KORMANO, and M.J. LIPTON. Magnetization Transfer Magnetic Resonance Imaging of Parenchymal Lung Disease. *Investigative Radiology* [online]. 1995, 30(2): 118-122 [cit. 2015-10-19]. DOI: 10.1097/00004424-199502000-00011.

[4] NIEMI, P.T., M.E.S. KOMU, and S.K. KOSKINEN. Tissue specificity of low-field-strength magnetization transfer contrast imaging. *Journal of Magnetic Resonance Imaging* [online]. 1992, 2(2): 197-201 [cit. 2015-10-19]. DOI: 10.1002/jmri.1880020213.

[5] ARNOLD, J.F., M. KOTAS, D. PRACHT, M. FLENTJE, and P.M. JAKOB. Could Functional MRI Improve Radiation Therapy Planning in Non-Small Cell Lung Cancer? *International Journal of Radiation Oncology*Biology*Physics* [online]. 2005, 63: S224-S225 [cit. 2015-10-19]. DOI: 10.1016/j.ijrobp.2005.07.384.

Step B: Full-Text Databases

The authors then conducted a search across full-text subscription databases (see Table 3) utilizing the queries defined in Table 1.

Resource Used	Query A	Query B	Query C	Query V1	Query V2	Query V3
EBSCOhost	2 235	834	4	44	208	67 371
ScienceDirect	42 465	5 860	300	300	300	300
SpringerLink **Biomedical Sciences**	5 872	815	83	156	322	322
SpringerLink Medicine	26 778	1 989	387	681	1 016	1 016
SpringerLink Public Health	1 215	138	45	65	76	76
Wiley Online Library	26 109	5 796	711	1 489	2 073	2 073
ProQuest Dissertations & Theses	9 483	6 681	282	855	2 317	2 317
ProQuest Health and Medicine	43 772	6 681	282	855	2 317	2 317

Table 3 Number of results for the query defined in Table 1 in selected full-text databases

Here, relevant results rarely included grey literature (i.e., project or technical protocols, or conference materials) or data. In the researcher's opinion, SpringerLink provided the most useful and relevant information in terms of grey literature because it indexes conference materials from field-related events organized by the European MRI society (ESMRMB), which enables access to reports about ongoing research projects [F] as well as conference abstracts and posters. In contrast, conference materials from another important scholarly society in this field, the International MRI society (ISMRM), are not indexed in these databases and therefore are only accessible to ISMRM members via login at the ISMRM websites and are not (as of October 2015) available to researchers at large who are not ISMRM members.

Step C: Institutional Repositories

The authors then conducted a search across selected institutional repositories, those with renowned research groups in this area of research (Table 4 below). Relevant grey literature results include conference materials (mostly posters and abstracts), dissertations, and occasionally project summaries or reports. Links to all repositories in Table 4 are available at The Directory of Open Access Repositories OpenDOAR (2).

(2) Available from **http://www.opendoar.org/countrylist.php**

Resource Used	Query A	Query B	Query C	Query V1	Query V2	Query V3
Universität Würzburg	143	18	14	143	143	143
Friedrich-Alexander-Universität Erlangen-Nürnberg	88	7	3	88	88	88
Eberhard-Karls-Universität Tübingen	643	509	710	755	755	755
Forschungszentrums Jülich	4	3	0	0	0	0
Ruprecht-Karls-Universität, Heidelberg	161	2	23	0	0	0
Health Services Research Projects in Progress	8	0	0	0	0	0

Table 4 Number of results in selected institutional repositories using query defined in Table 1

Step D: Dedicated Grey Literature Repositories

Finally, the authors conducted a search across dedicated grey literature repositories. Most useful for this particular area of inquiry, from the researcher's perspective, are the **Public Health Grey Literature Sources (3)** (listed in Table 5) provided by the OPHLA (Ontario Public Libraries Association), which covers American, Canadian, and international grey data repositories; and the **Data Sharing Repositories** (4) provided by US National Library of Medicine.

International European Repositories	A	B	C
Electronic Theses Online Service (ETHOS) | British Library	22	2	0
Center for Research Libraries Foreign Dissertation	537	1	538
DART-Europe E-theses Portal	30	18	30
National Institute for Health and Clinical Excellence (NICE)	24	0	0
Public Health England	1	0	0
UK Department of Health	22	172	95
Nature Precedings	15	1	0
World Health Organization	93	0	0

Table 5 Number of results for query from Table 1 in dedicated grey literature repositories.

(3) Available from **http://www.ophla.ca/pdf/Public%20Health%20Grey%20Literature%20Sources.pdf**
(4) Available from **https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html**

Because of the narrow nature of this research query, the authors found no relevant results in GreyNet International, European Commission (EUROPA) Public Health or European Union Open Data Portal.

In sum, the grey literature and data results from all four sample searches yielded little relevant information and results were primarily conference proceedings and related materials. Raw data of other ephemeral materials such as laboratory notes relevant to this project was not available from any resource.

In the following section, the authors describe the particular difficulties faced by researchers in this field regarding data and related ephemeral materials, which is – as seen in the examples above – currently quite inaccessible to scholars in traditional databases as well as in repositories.

The Torturous Path from Dark to Grey Data

Any research conducted in this particular area of bioengineering research results in various data outputs. For this researcher, data she commonly works with for one project can be divided into the following groups (Table 6), with related storage requirements indicated in the far-right column.

Data Type	**Size**
Single data set (i.e., an individual MRI image received using a measuring protocol)	**3 - 4 MB**
RAW data sets (total images per scientific project)	**80 - 100 GB**
Laboratory notes and diaries (evaluation of single data sets, parameters)	**MB**
Summaries and statistics (comparison of data sets per chosen parameter)	**MB**
Conference materials (posters, presentation, supportive materials)	**2 - 3 GB**
Supportive materials for peer-reviewed outcomes (images, tables, graphs)	**MB**
Programming files (measurement and evaluation files, necessary .exe programs)	**35 - 60 GB**
Research-related literature and data (including full-text results of literature search, text variants, images etc)	**2 GB**
Total size of all the related project materials	**100 - 170 GB**

Table 6 List of different scientific data outputs and their file sizes (based on the realized scientific project); in several cases, only list as megabits (MB) because it is impossible to calculate their exact size in comparison with other data types within the context of this paper.

Other valuable and even essential data sources include records of development for the project's working hypotheses, false measurements (human versus technical errors), and RAW data with unusual artifacts.

If these kinds of data (and similar data for related projects) were more broadly available to researchers in this field, direct benefits for researchers would include: detailed information about the chronology and methodology of a particular project, how the project was evaluated,

possible pitfalls future researchers might avoid, and areas in which future research is not yet possible because of "dead ends" (e.g., inability of current technologies to address particular research questions). In particular, a deeper understanding of the challenges of a chosen method and verification of hypotheses against previous experiments might be particularly useful if such data were more readily available.

Further benefits of accessible RAW data images would include, for MRI programmers, the ability to test data sets in order to improve evaluation software (e.g., to compare MRI images against previous results). Also, researchers would be able to additionally compare results from other students according to selected parameters even if the original author did not evaluate them (e.g., influence of sex or age, ventilated gas, breath or cardiac phase, etc.). This detailed data would enable more thorough and reliable statistics regarding parameters, data, and even artifacts.

Laboratory diaries and notes offer a unique complex perspective into each particular scientific project. Especially when conducting basic science, even small notes about what was manageable or what failed can spare other researchers tremendous amounts of time. However, many small side experiments might have a great value and can contribute to deeper understanding topics, yet they are currently mostly stored only in researchers' heads and officially are not made public.

Of course, the strong experiment-dependence of the Magnetization Transfer Contrast methods might make comparison with the results of different experiments impossible. But with detailed knowledge of ALL the measurement parameters which might be possible to observe using MTC techniques, similarities and regularities not obvious at first (or even third) glance might be actually comparable for a carefully-chosen parameter.

Many of the challenges of publishing such dark data, on the other hand, are not so obvious. When you have terabits of dark data, where might one (publicly and accessibly) store them? And how should one properly index them to make them accessible? And – particularly relevant in biomedicine and biomedical engineering – how might one address the ethical questions regarding human (i.e., volunteer or patient) data? No potential data repository in this narrow field currently provides satisfactory answers to these questions.

Another complication is the manner in which one might present individual data set results, laboratory notes, etc. In this field, there are not currently any standardized formats or platforms which might broaden access to these materials. Because of this, every researcher currently must find a compromise between the presentation of original data (e.g., a written lab diary) and its public presentation (experiment's records or technical protocols). This extra work, combined with a need for highly structured texts and outcomes, discourage many scientists from sharing such data. There is currently no universal platform or repository for sharing such information in this field.

Illustrations of Typically Unindexed Information (DARK DATA)

As a brief introduction for non-scientists, the authors would like to provide in this section examples of real project outcomes: MRI images (full single data set, images with artifacts; Fig. 1 and 2) and sample scientific laboratory notes (Fig. 3). Fig. 4 provides an overview of the data summary.

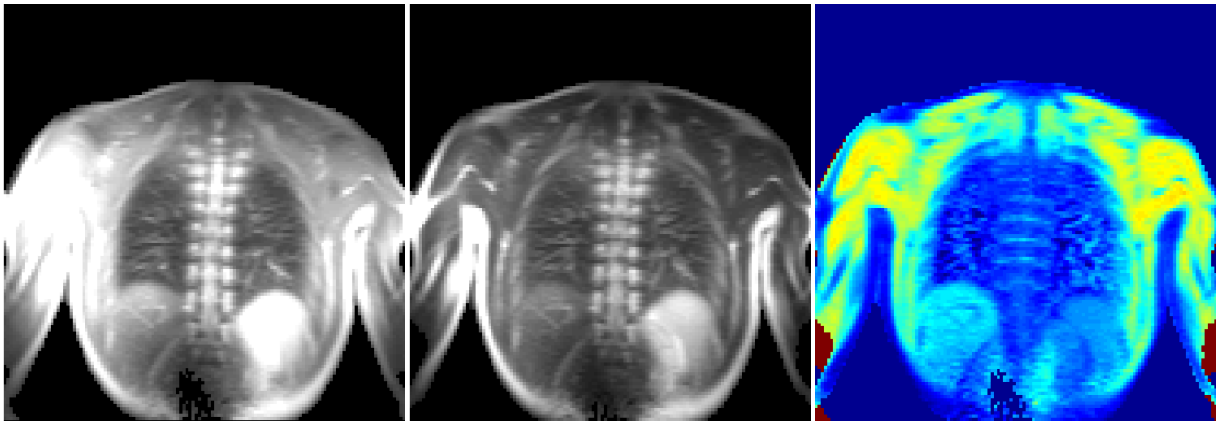

Figure 1 Typical example of MTC-MRI evaluated single data set - MRI images without and with MTC preparation and the final calculated (colorful) contrast image (useful data set).

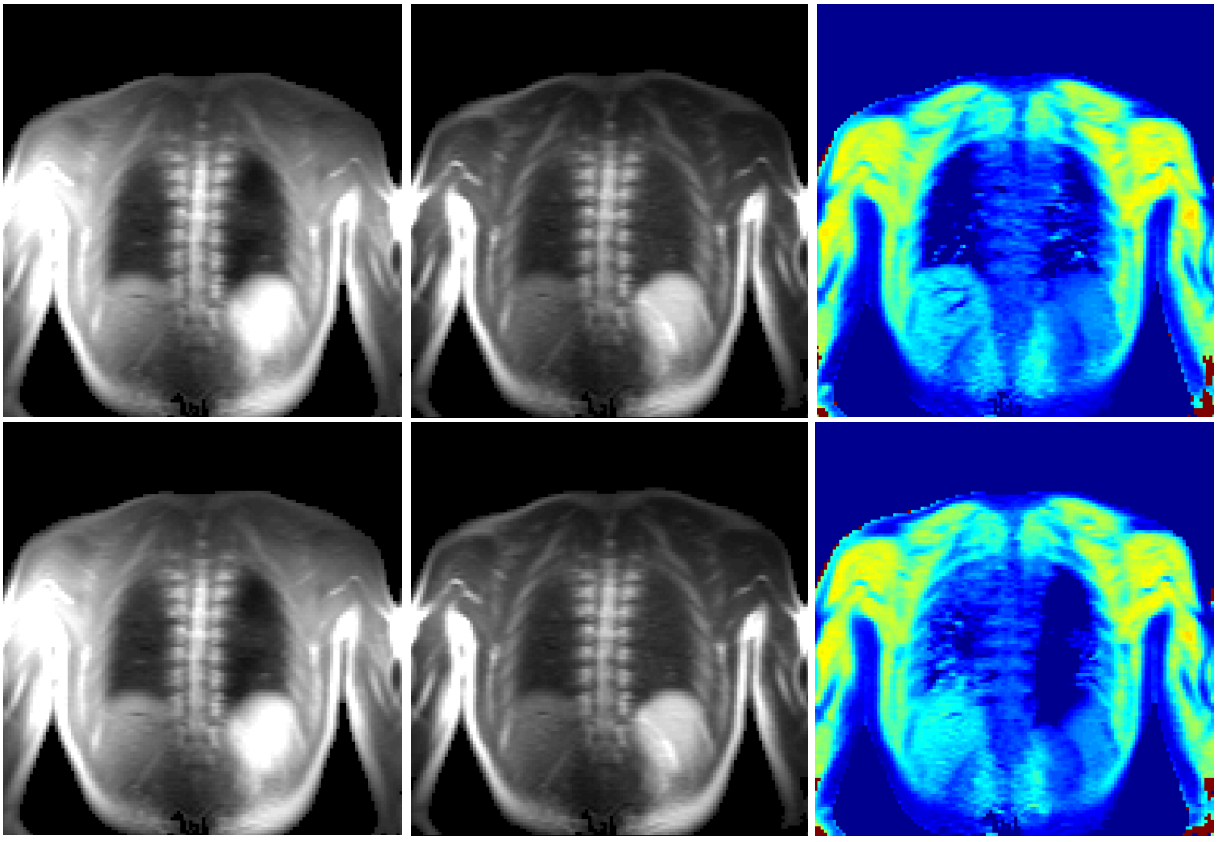

Figure 2 Two of typical examples of a severe artefacts in the final calculated MRI image (useless data set) - first set according to comparison from two images with different cardiac phase, second data set due to motion artefact.

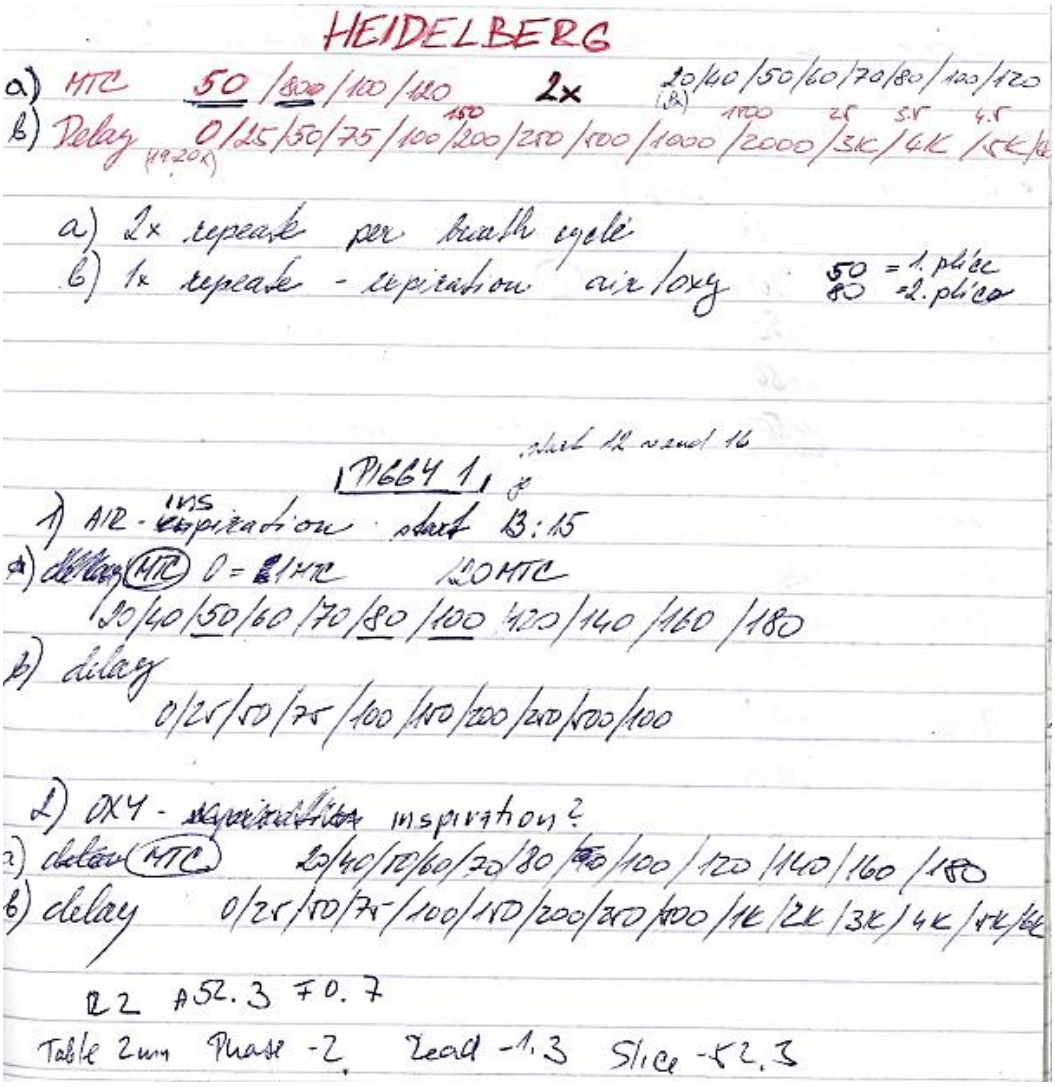

Figure 3 Illustrational scan of the laboratory diary with notes (hand-written, in current state unpublishable)

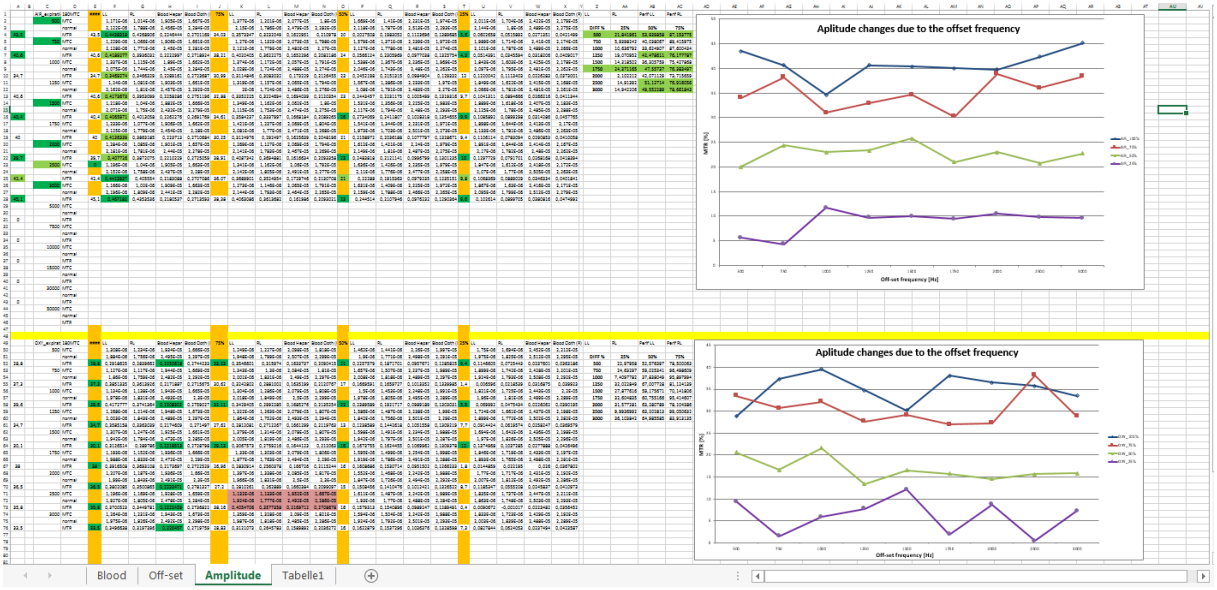

Figure 4 Illustrational scan of evaluated and roughly visualized individual data set's results (no personal data included) - shown results represent the measurement on the lung phantom with ex-vivo porcine lungs, ventilated with room air and pure oxygen.

Conclusion and Recommendations

In this case study, the authors provided examples of the availability and accessibility of grey literature and data in a new, very narrow research topic in medical diagnostics/bioengineering modeled on the real-life, completed project "Magnetization Resonance Imaging of the Lung Structure using Magnetization Transfer Contrast Method." All images and examples used belong to the authors and have not yet been previously published.

Grey literature identified in four sample searches across different resource types include conference materials (abstracts, posters and e-posters, presentations) and dissertations, but rarely project reports. For national institutional and other grey data repositories, language barriers potentially create great challenges; this paper focused on search results using English-language queries only and additional language queries would need to be tested in future studies, as they were beyond the scope of this paper.

In terms of grey literature and grey/dark data, while materials from conferences or scientific meetings as well as dissertations and projects reports from universities are available and can be found, other data types such as laboratory diaries, measurement reports, and case studies are – in this emerging field – completely inaccessible. Due to aforementioned challenges such as the overall storage requirements for dark data (especially RAW data), no standardized indexing formats, ethical guidelines, and lack of universal storage platforms in this field, only a few researchers are willing to share their dark data in grey literature repositories. The authors did identify a few bioinformatics pioneers who are interested in opening the doors of access to bioengineering data in the future and who envision grey data repositories which might include more than conference materials. Examples of recent projects in this area include XTENS [G], BIRN (4), and OpenScienceLink (OSL, (5)) [I]; the open source ***Biomedical Data Journal*** provides a forum for recent research in bioinformatics.

To improve the accessibility of grey literature and grey/dark data in this field, the authors recommend the following: first, it is necessary to determine which data sources could/should be stored in grey data repositories and prepare a universal format/platform in order to properly structure and describe data for future research. Second, ethical questions regarding the handling of personal information regarding human subjects (i.e., volunteers and patients) including preventing misuse or appearance to the public (non-medical) audience must be addressed. Until these basic questions are defined, it is unlikely that researchers in this particular field would be willing to share dark data more broadly.

(4) Available from **http://www.birncommunity.org/about/birn-video-intro/**
(5) Available from **https://www.gopubmed.org/web/oslplatform/**

References

[A] KRUEGER, S. *Beyond the Paywall: A Multi-Sited Ethnographic Examination of the Information-Related Behaviors of Six Scientists*. Berlin: Humboldt-Universität zu Berlin. Forthcoming dissertation to be published 2016.

[B] YOUNG, J. M. *An epidemiology of big data*. Syracuse: Syracuse University, 2014. Dissertation. ISBN 9781303909979. Available from ProQuest Dissertations & Theses Full Text: The Sciences and Engineering Collection.

[C] HEIDORN, P.B. Shedding Light on the Dark Data in the Long Tail of Science. *Library Trends*. 2008, vol. 57, no. 2. pp. 280-299. ISSN 00242594. Available from ProQuest SciTech Collection.

[D] LANDER, H. and A. RAJASEKAR. *DataBridge: Creating Bridges to find Dark Data* [online]. Chapel Hill: RENCI (University of North Carolina), 2015 [cit. 2015-10-18]. RENCI White papers series, Vol. 3, No.5. DOI: 10.7921/G0MS3QNF.

[E] JIROTKA, M., C.P. LEE, and G.M. OLSON. Supporting Scientific Collaboration: Methods, Tools and Concepts. *Computer Supported Cooperative Work (CSCW)* [online]. 2013, **22**(4-6): 667-715 [cit. 2015-10-18]. DOI: 10.1007/s10606-012-9184-0.

[F] *ESMRMB 2015: 32nd Annual Scientific Meeting, Edinburgh, UK, 1-3 October: Abstracts, Thursday* [online]. Springer Berlin Heidelberg, 2015 [cit. 2015-10-18]. ISSN 1352-5243. DOI: 10.1007/s10334-015-0487-2.

[G] IZZO, M., G. ARNULFO, M.C. PIASTRA, V. TEDONE, L. VARESIO, and M. M. FATO. XTENS - A JSON-based digital repository for biomedical data management (2015). In: F. ORTUNO and I. ROJAS, ed. *Bioinformatics and Biomedical Engineering: Third International Conference, IWBBIO 2015, Granada, Spain, April 15-17, 2015. Proceedings, Part II.* Springer International Publishing, 2015, p. 123-130. Lecture Notes in Computer Science, vol. 9044. ISBN 978-3-319-16479-3. DOI 10.1007/978-3-319-16480-9

[H] EISINGER, D., G. TSATSARONIS, A. PETROVA, E. KARANASTASIS, V. ANDRONIKOU, and E. CHONDROGIANNIS. OSL Platform: A Link to Open-access Scientific Information and Structured Data. *Biomedical Data Journal* [online]. 2015, **01**(1): 52-54 [cit. 2015-10-18]. DOI: 10.11610/bmdj.01109.

8th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2015 [cit. 2015-12-15]. Available from: **http://nrgl.techlib.cz/index.php/Proceedings**. ISSN 2336-5021.

PARLIAMENTARY INSTITUTE

Stanislav Caletka

caletka@psp.cz

Office of the Chamber of Deputies of the Parliament of the Czech Republic

Abstract

Parliamentary Institute is one of the departments of the Office of the Chamber of Deputies of the Parliament of the Czech Republic. The Institute performs tasks of parliamentary scientific, informative and educational center. Parliamentary institute has been divided into three departments – Department of General Analysis, Department of EU Affairs and Department of Communication and Education. The paper will provide information on the activities of particular Departments. Emphasis will be placed especially on the processing of various documents, background materials for MPs, analyses or studies.

Keywords

Parliament of the Czech Republic, Parliamentary Institute, National Repository of Grey Literature

Introduction – presenting the Parliamentary Institute[105]

The Parliamentary Institute (also referred to as "PI") is one of departments of the Office of the Chamber of Deputies of the Parliament of the Czech Republic. The origins of the PI stretch back to 1990, when an initiative of Deputies was established in 1990 within the scope of the Federal Assembly of that time to create a workplace whose aim it was to provide expert services to Deputies (mainly economic and legal analyses). Most employees of the Office of the Federal Assembly were released following the demise of the Federal Assembly at the end of 1992, but employees of the Parliamentary Institute, the Parliamentary Library and certain other departments were taken on in April 1993 by the newly-created Office

[105] The text of this chapter is taken from the statistical yearbook of the Parliamentary Institute: Parlamentní institut, Parlament České republiky, Kancelář Poslanecké sněmovny, XII. vydání, Praha: [Kancelář Poslanecké sněmovny Parlamentu ČR], 2014, 293 str., page 7 and 29 – 35. The yearbook is available from the Secretariat of the Parliamentary Institute or at the Chamber of Deputies' information centre.

of the Chamber of Deputies. It became an independent department within the Office of the Chamber of Deputies in September 1995 following the adoption of a new organisation order.

The competencies of the PI are defined in the organisation order, according to which the Parliamentary Institute discharges tasks of a scientific and informative nature and those of a training centre for the Chamber of Deputies, its bodies, Deputies and the Office of the Chamber of Deputies. It also carries out such tasks for the Senate, its bodies, officials, Senators and the Office of the Senate. The Parliamentary Institute has within it three departments: The Department of General Analysis mainly compiles answers to questions and requests submitted by members of both parliamentary chambers and provides a general service for the bodies of the Chamber of Deputies or information from the sphere of foreign policy. The Department of EU Affairs mainly functions as an expert base for the Committee on European Affairs of the Chamber of Deputies. It processes the database of documents arriving from European Union bodies and other matters to concern the EU. The Department of Communication and Education provides services to the general public (such as guided tours, information about the work of the Chamber of Deputies etc.) and operates the Information Centre of the Chamber of Deputies. The individual departments of the Parliamentary Institute are in close contact with each other, jointly providing (in particular) expert training and seminars for members of both chambers of parliament and for the general public. Authors from all departments share in larger joint projects and studies.

How the Parliamentary Institute works – the basic principles[106]

The Parliamentary Institute discharges tasks of a scientific, informative and educational nature and, within the scope of its activity, answers questions posed by members and bodies of both chambers of parliament and by the Office of the Chamber of Deputies and the Office of the Senate. It mainly deals with legal and economic issues and issues of political science and with other questions.

Requests for the Parliamentary Institute to provide an answer to a question can be made by telephone, by e-mail or using the form available at the Chamber of Deputies' website.[107] Deputies or Senators can use the answers or studies compiled, for example, for their own legislative activity, for criticism or consideration of adopted laws, for decision-making on adopted laws and for other needs of the performance of a mandate, for example to reply to voters, for making speeches etc. The PI, however, does not compile studies to provide information for assistants or other people.

The Parliamentary Institute places considerable emphasis on providing objective information irrespective of who submits the request to compile a study or analysis. The submitter of the question (i.e. Deputy, Senator, body of a parliamentary chamber etc.) remains anonymous and his/her/its name is not published.

[106] The text of this chapter is taken from the statistical yearbook of the Parliamentary Institute: Parlamentní institut, Parlament České republiky, Kancelář Poslanecké sněmovny, page 26 – 27 and 40 - 41.
[107] **http://www.psp.cz/sqw/hp.sqw?k=59**[cit. 2015-10-07]

After presenting the study compiled, the Parliamentary Institute reserves two weeks for the submitter of the question (according to the principle of exclusivity). During this time the study is not made available to any other interested parties. It may then be made available to other Deputies and Senators after two weeks. If such a study has a broader scope of use, it is filed in the relevant database of selected works, which are subsequently published on the Chamber of Deputies' website in the Parliamentary Institute section.[108] Answers to questions and other work that is/are not timeless in nature are archived at the Parliamentary Institute for further internal use or as base material for the creation of new studies or analyses. The PI compiles some studies in advance in relation to presented and planned legislation. In such case, studies are immediately made available to all Deputies and Senators in distributed newspapers and are also published on websites.

The Parliamentary Institute administers in its internal databases all materials compiled, including approximately 2,800 shorter analyses and other specialised works sent to the relevant submitters in the past as part of the answer to a question.

Studies are made available to the public on the Chamber of Deputies' website (in the Parliamentary Institute section) that are also regularly entered in the database of the National Repository of Grey Literature, administered by the National Library of Technology (NLT), based on a contract between the NLT and the Office of the Chamber of Deputies. These are studies compiled by the Parliamentary Institute from a whole range of areas of interest (for example, law, international policy, public administration, economics, security etc.).

Furthermore, the Parliamentary Institute publishes information on its website that provides an overview of economic events in the EU, selected economic, currency and social indicators (monthly) and selected topics (i.e. anticipated topics that are to come up for debate at the Chamber of Deputies and the Senate). The Parliamentary Institute is currently preparing new material regarding the issue of migration, which it will regularly post on the Internet.

The Parliamentary Institute compiles internal materials for Deputies and Senators to concern the Common Foreign and Security Policy, statements regarding the printing of non-governmental bills, foreign political documents and reports by the Permanent Representative of the Chamber of Deputies at the European Parliament and compiles a selection of EU documents.

The work of the Parliamentary Institute[109]

Requests for answers to questions sent to the Parliamentary Institute are mainly processed by the Department of General Analysis. This department therefore prepares various types of materials, which can be divided into "work on request" and "regular overviews and core topics".

Answer to a question provides basic information on an issue, answering the questions asked by Deputies and Senators in an operative way. These materials need not contain a link to the source of information from which they were compiled. Informative materials provide more

[108] **http://www.psp.cz/sqw/ppi.sqw?d=1**[cit. 2015-10-07]

[109] The text of this chapter is taken from the statistical yearbook of the Parliamentary Institute: Parlamentní institut, Parlament České republiky, Kancelář Poslanecké sněmovny, page 8 – 11.

detailed answers to questions. Comparative studies are a special form of material, providing an overview of the way of dealing with certain areas in different countries. Comparison, however, concentrates mainly on European Union Member States. The final group is studies, which are based on expert analyses of a certain topic. Employees at the Department of General Analysis also provide interested Deputies and Senators with personal consultation, the content of which depends on mutual agreement.

One special type of document is the regular overview of documents regarding Common Foreign and Security Policy, which is compiled for the Committee on European Affairs and the Committee on Foreign Affairs at the Chamber of Deputies, or indeed other interested parties. Informative materials follow on from this overview, elaborating in more detail on information about areas of the Common Foreign and Security Policy which the committees mentioned have decided to consider in more detail and over the long-term. In terms of foreign policy, the Department also processes documents for foreign-policy discussion involving members and bodies of the Chamber of Deputies. Due to developments in the eurozone, the Parliamentary Institute also compiles an Overview of Economic Events in the EU (published twice a month). Material on the topic of Selected Economic and Social Indicators is prepared and published once a month. A series of Selected topics is also published. This series mainly considers up-and-coming topics that are to come under consideration at the Chamber of Deputies and the Senate. The results of elections and the latest political developments, in particular in Member and Candidate EU States, are also presented in this series, as is economic development within the countries of the eurozone. The aim of the series is to provide members of Parliament with information at the time the relevant topic reaches the agenda. Employees also process core topics within their own specialisations, topics which the Chamber of Deputies deals with over the long term or regularly returns to. Such topics include election systems, the performance of the mandate of members of representative unions and the institute of immunity, the system of electing a head of state, the issue of referenda and their role in a system of representative democracy, the system of municipal democracy etc.

The Department of EU Affairs compiles information on, analyses and comprehensive studies of European Union political documents, legislation and policies. It also compiles regular weekly overviews of European Union documents that are mainly drawn up for discussion at the Committee on European Affairs. The department also prepares background materials for checking government procedure in European Union matters by the Chamber of Deputies. These materials offer a starting point for the proposed resolutions of the Committee on European Affairs. The department compiles materials for this committee in relation to regular consultation on legislative proposals of the government from the perspective of their compatibility with EU law. One important activity is the processing of statements on the compatibility of laws with European Union law in relation to non-governmental bills.

The Department of Communication and Education produces, among other, a regularly updated, special-purpose series of printed informative materials about the activities of the Chamber of Deputies. Materials are freely accessible from the information centre. This department also deals with preparing answers to questions posed by foreign parliaments within the inter-parliamentary ECPRD network (European Centre for Parliamentary Research and Documentation).

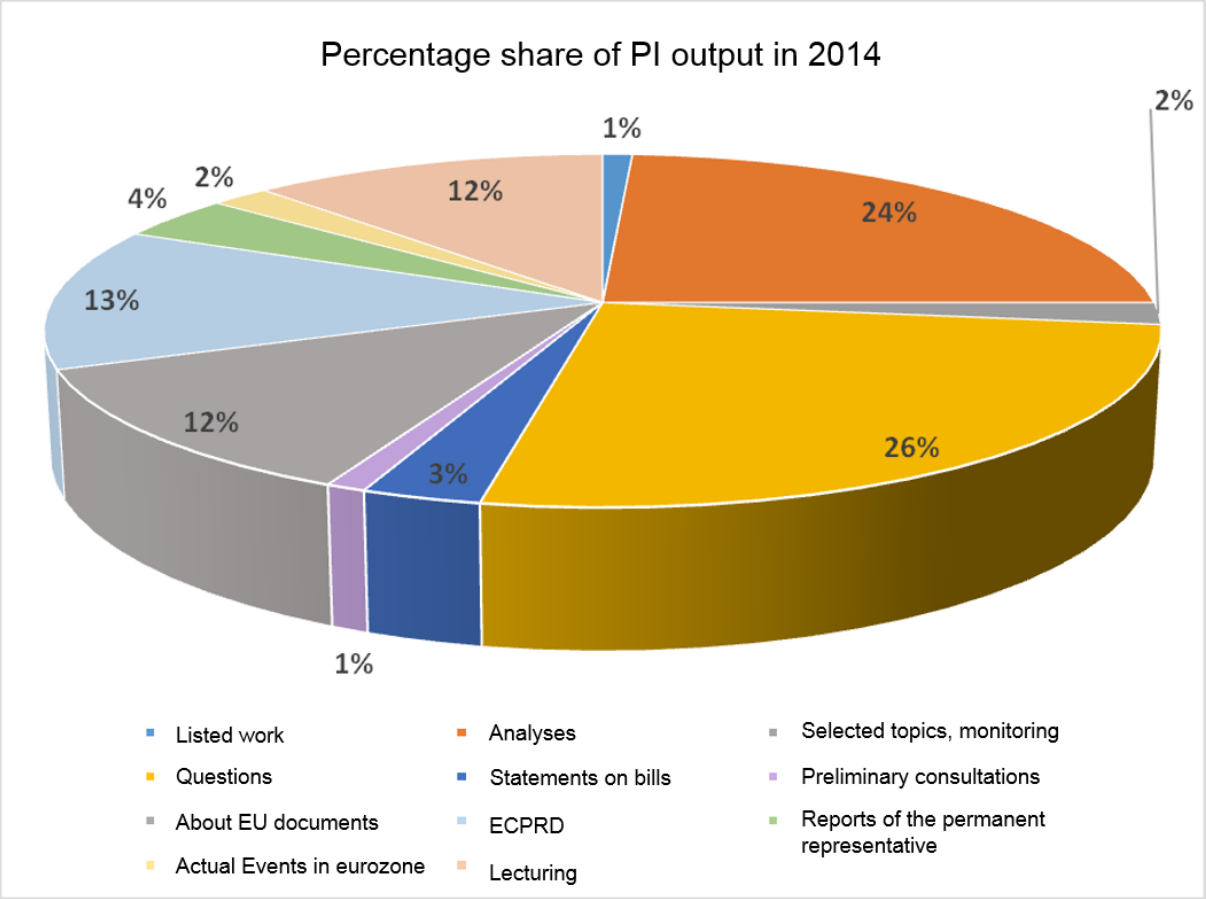

Figure 1 Percentage of outputs of PI in 2014

Percentage share of PI output in 2014		
Listed work	Analyses and other specialised work	Selected topics, monitoring
Questions: e-mail, fax, telephone	Statements on bills	Preliminary consultation
Information regarding EU documents	ECPRD	Reports of the permanent representative
Current development in the eurozone	Lecturing	

Table 1 Percentage share of PI output in 2014 (Source: internal statistics of the Parliamentary Institute)

Statistics regarding work produced at the Parliamentary Institute for the year 2014		
(from 1.1.2014 to 31.12. 2014/	Quantity of work produced	
	number	%
Total quantity of work produced	1016	100%
of which:		
presented in the List of Selected Work ("Listed Work")	14	1%
Analyses and other specialised work	248	24%
selected topics and monitoring	20	2%
answers to questions (letter, fax, e-mail)	261	26%
statements on bills	33	3%
preliminary consultation	4	1%
information regarding EU documents (source materials, overviews)	122	12%
The European Centre for Parliamentary Research and Documentation (ECPRD)	133	13%
reports and materials of the permanent representative to the European Parliament	41	4%
current development in the eurozone	17	2%
lecturing activity	123	12%
Work produced:	1016	100%
on request (total)	931	92%
for the Chamber of Deputies	695	75%
for the Senate	49	5%
for MEPs	1	0%
foreign parliaments	158	17%
other	28	3%
at the PI's own initiative	85	8%
Work ordered by the Chamber of Deputies or the Senate:	744	100%
by a Chamber of Deputies or Senate committee or commission	307	41%
by a Deputy or Senator him/herself	437	59%

Work according to political affiliation of the submitting deputy or senator	437	100%
ČSSD (Czech Social Democratic Party)	66	15%
ODS (Civic Democratic Party)	16	4%
TOP09+S	106	24%
KSČM (Communist Party of Bohemia and Moravia)	35	8%
SZ (Green Party)	3	1%
ANO 2011	69	15%
KDU-ČSL (Christian and Democratic Union – Czechoslovak People's Party)	52	12%
Dawn - National Coalition	84	19%
Independent	6	2%

Table 2 Statistics regarding work produced at the Parliamentary Institute for the year 2014 (Source: internal statistics of the Parliamentary Institute)

The Parliamentary Institute and the National Repository of Grey Literature

The origins of cooperation between the National Library of Technology and the Office of the Chamber of Deputies of the Parliament of the Czech Republic in relation to cooperation with the National Repository of Grey Literature date back to the second half of 2011, when the director of the National Library of Technology contacted the director of the Parliamentary Library in a letter. The Parliamentary Institute was contacted, given the nature of the work and the documents produced, to become the main contributor to the grey literature database for the Chamber of Deputies. Several months of exchanging information were followed at the end of 2012 by signing an agreement on the terms and conditions of using original works. The Parliamentary Institute progressively enters major studies in the National Repository of Grey Literature database, in that these are also available at the Parliamentary Institute website.[110] A total of 120 such papers were stored in the National Repository of Grey Literature database on 28th September 2015, dating back to the years 1992 – 2013. According to the available statistics for 2013, the Parliamentary Institute entered a total of 84 documents in the database for the relevant year, the highest number of all cooperating institutes.

Conclusion

The aim of this brief paper was to present the work of the Parliamentary Institute and emphasise the material produced, some of which is also stored in the database of grey literature. The specialised activity of the Parliamentary Institute helps members of both parliamentary chambers access up-to-date, objective and apolitical information. It also allows the general public to find important facts and data from a whole range of expert studies and

[110] **http://www.psp.cz/sqw/ppi.sqw?d=1**[cit. 2015-10-07]

other analyses that they would otherwise find difficult to access thanks to the database of the National Repository of Grey Literature and the Parliamentary Institute website. The aim of the Parliamentary Institute is to enter all studies or selected major papers intended for broader use in the database of the National Repository of Grey Literature, which should allow the general public to enjoy better accessibility to analyses and studies already having been compiled.

References

PARLAMENTNÍ INSTITUT. *Parlamentní institut, Kancelář poslanecké sněmovny: Ročenka* [online]. 12. ed. Praha: Kancelář Poslanecké sněmovny Parlamentu České republiky, 2014 [cit. 2015-11-25].

Poslanecká sněmovna Parlamentu České republiky: Parlamentní institut [online]. Praha [cit. 2015-11-25]. Available from: **http://www.psp.cz/sqw/hp.sqw?k=40**.