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.

Keywords

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

---

Dissertations and data

Open access to PhD dissertations[1] 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,

[1] 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".

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 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[2]. 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[3] 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%)[4]. 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).

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

[3] 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.

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

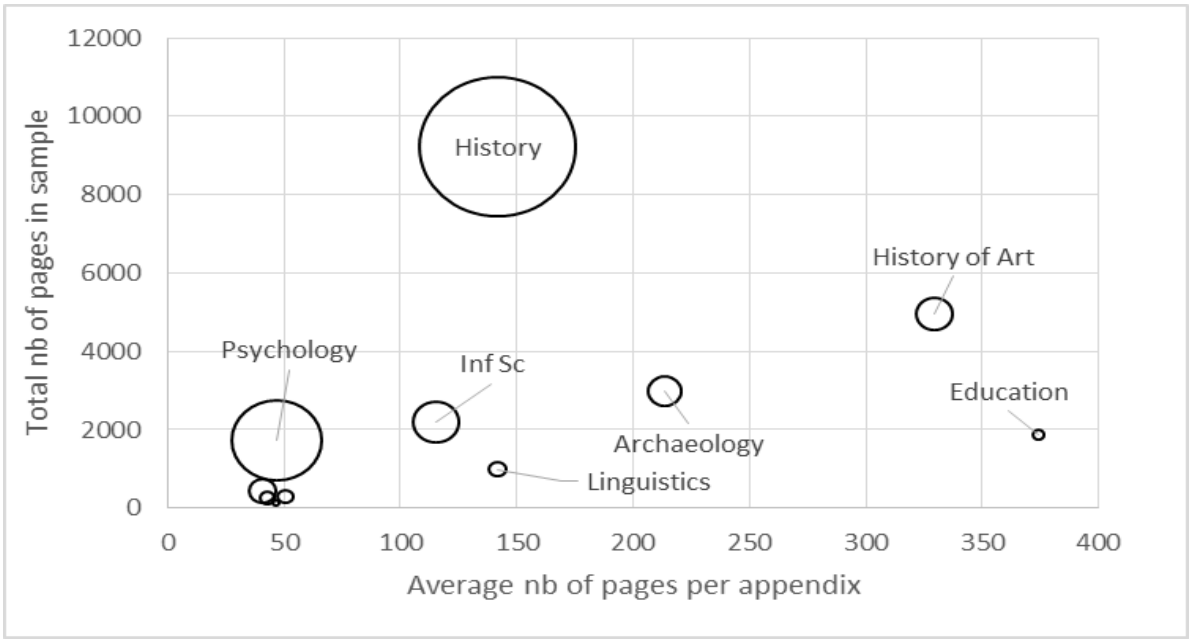

Figure 1 Size of appendices (circle size = number of dissertations, N=188)

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).

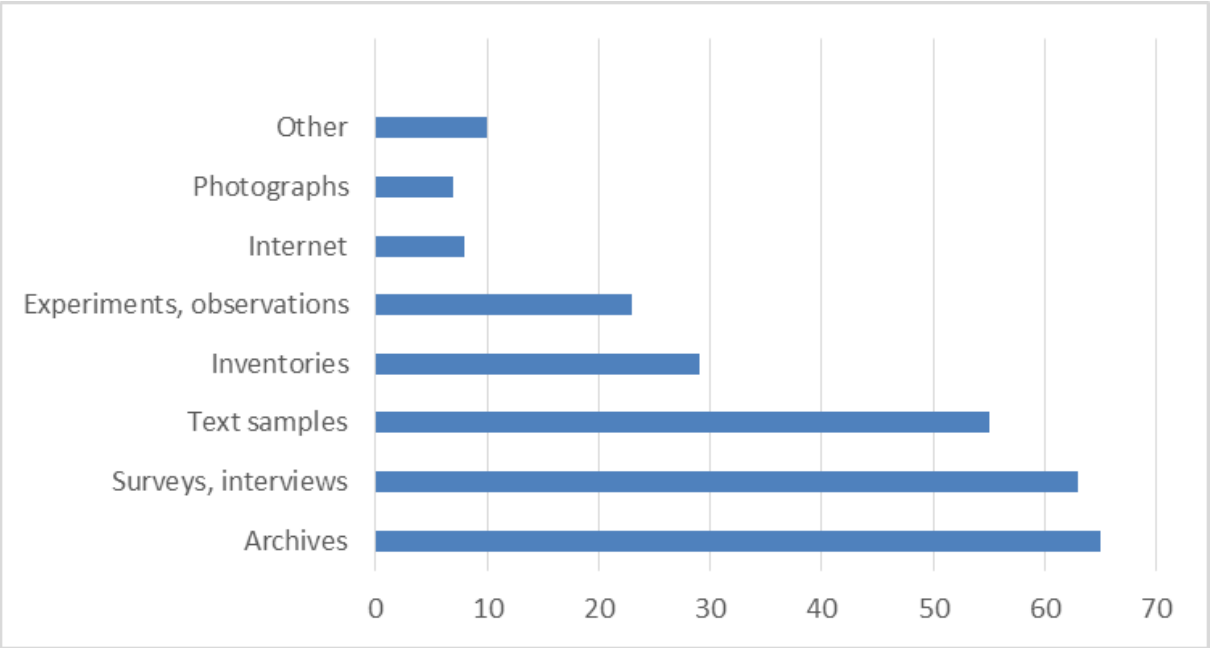

Figure 2 Data sources per dissertations (N=188)

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).

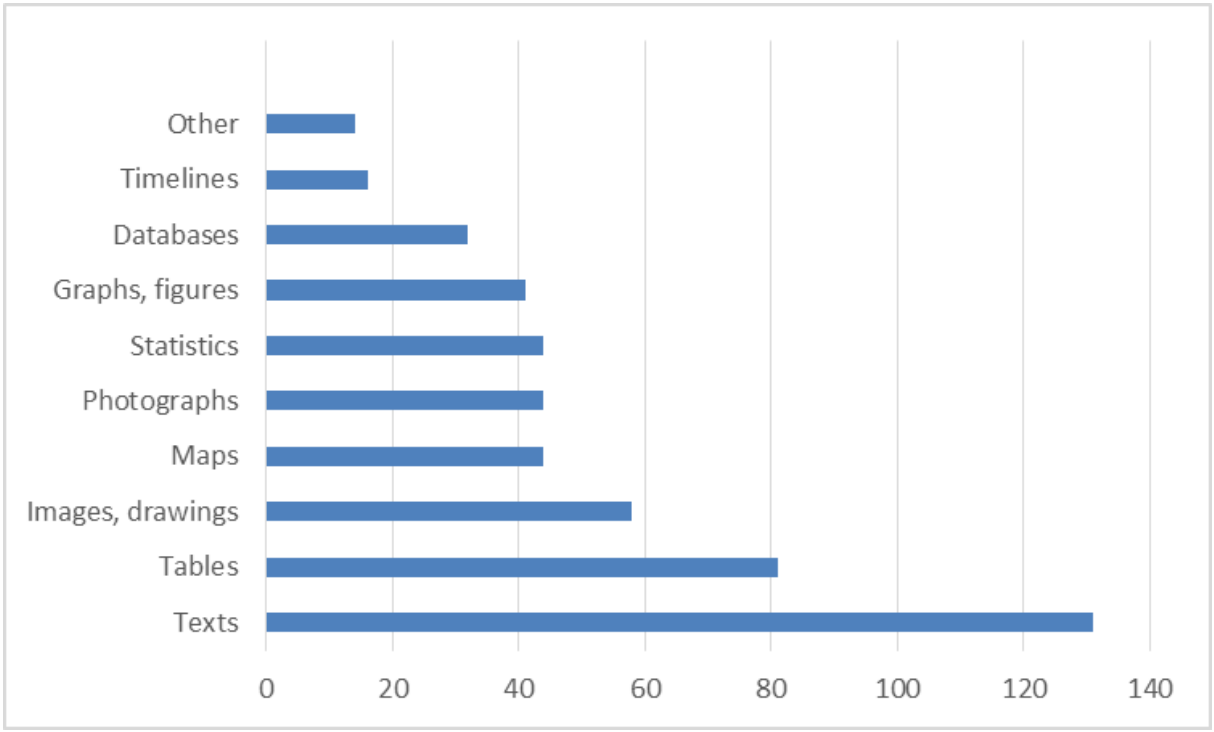

Figure 3 Data types, per dissertations (N=188)

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[5] maintains a list of accepted file formats for long term preservation[6]. 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[7] 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.

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

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

[7] 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[8] - 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.

[8] 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).

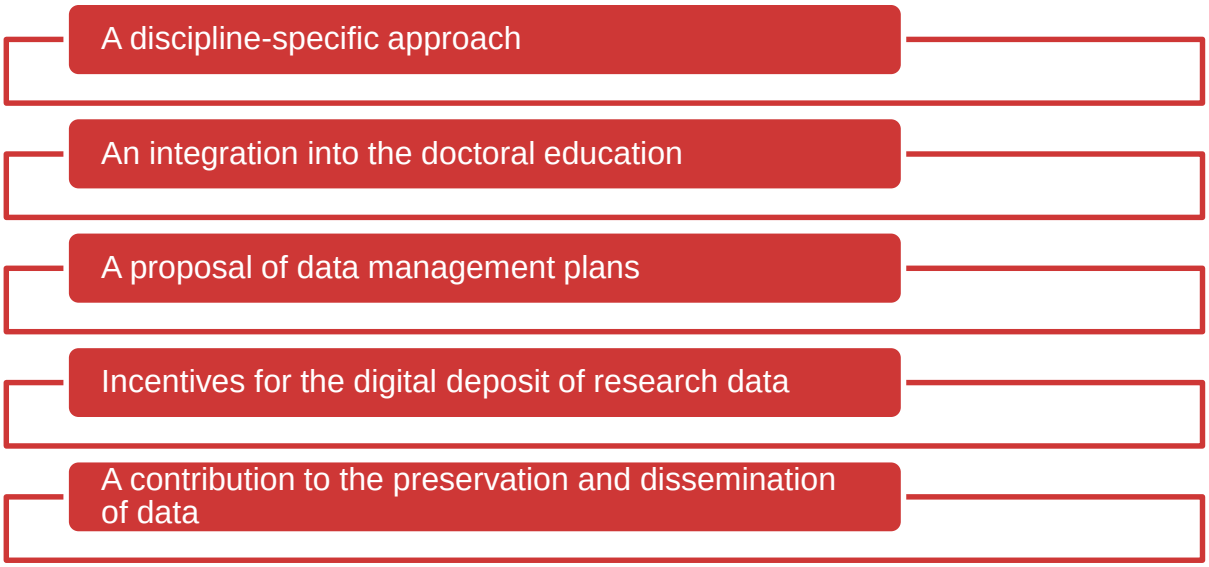

Figure 5 Five leading principles for the implementation of research data support services

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[9] and part of the national open repository HAL[10].

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.**

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*

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

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.