Dissertations and Data*

Joachim Schöpfel, HélèneProst, Cécile Malleret, (France);
Primož Južnič, Ana Češarek, Teja Koler-Povh (Slovenia)

Abstract

The keynote provides an overview on the field of research data produced by PhD students, in the context of open science, open access to research results, e-Science and the handling of electronic theses and dissertations. The keynote includes recent empirical results and recommendations for good practice and further research. In particular, the paper is based on an assessment of 864 print and electronic dissertations in sciences, social sciences and humanities from the Universities of Lille (France) and Ljubljana (Slovenia), submitted between 1987 and 2015, and on a survey on data management with 270 scientists in social sciences and humanities of the University of Lille 3.

The keynote starts with an introduction into data-driven science, data life cycle and data publishing. It then moves on to research data management by PhD students, their practice, their needs and their willingness to disseminate and share their data. After this qualitative analysis of information behaviour, we present the results of a quantitative assessment of research data produced and submitted with dissertations Special attention is paid to the size of the research data in appendices, to their presentation and link to the text, to their sources and typology, and to their potential for further research. The discussion puts the focus on legal aspects (database protection, intellectual property, privacy, third-party rights) and other barriers to data sharing, reuse and dissemination through open access.

Another part adds insight into the potential handling of these data, in the framework of the French and Slovenian dissertation infrastructures. What could be done to valorise these data in a centralized system for electronic theses and dissertations (ETDs)? The topics are formats, metadata (including attribution of unique identifiers), submission/deposit, long-term preservation and dissemination. This part will also draw on experiences from other campuses and make use of results from surveys on data management at the Universities of Berlin and Lille.

The conclusion provides some recommendations for the assistance and advice to PhD students in managing and depositing their research data, and also for further research.

Our study will be helpful for academic libraries to develop assistance and advice for PhD students in managing their research data, in collaboration with the research structures and the graduate schools. Moreover, it should be helpful to prepare and select research data for long-term preservation, curate research data in open repositories and design data repositories.

The French part of paper is part of an ongoing research project at the University of Lille 3 (France) in the field of digital humanities and research data, conducted with scientists and academic librarians. Its preliminary results have been presented at a conference on research data in February 2015 at Lille, at the 8th Conference on Grey Literature and Repositories at Prague in October 2015 and published in the *Journal of Librarianship and Scholarly Communication*. The Slovenian research results have not been published before.

Keywords Open science, open data, open access, institutional repository, data repository, research data, research data management, electronic theses and dissertations

---

* First published in the GL17 Conference Proceedings, February 2016.

1.Data-driven science

Scientific results are increasingly disseminated as digital datasets. “Data are becoming an important end product of scholarship, complementing the traditional role of publications” (Borgman et al. 2007). Data are not only the fuel of the digital economy[1] but also of science and engineering.

Five years ago, the European Commission met the challenge and defined the main strategy for the development of an ambitious scientific data policy in the European Research Area. This strategy includes a framework for a collaborative data infrastructure, additional funding, measuring and rewarding data value and training of experts[2]. The need to manage the “data deluge”, is among the main drivers of computationally intensive science or e-Science, “the powerful paradigm in which distributed computer and knowledge systems, and information and communication technologies are integrated to provide services to enable large-scale and collaborative sciences and engineering” (Wang & Liu 2009). Mathematical modelling, numerical analysis and visualization techniques are part of this new way of doing science (figure 1).

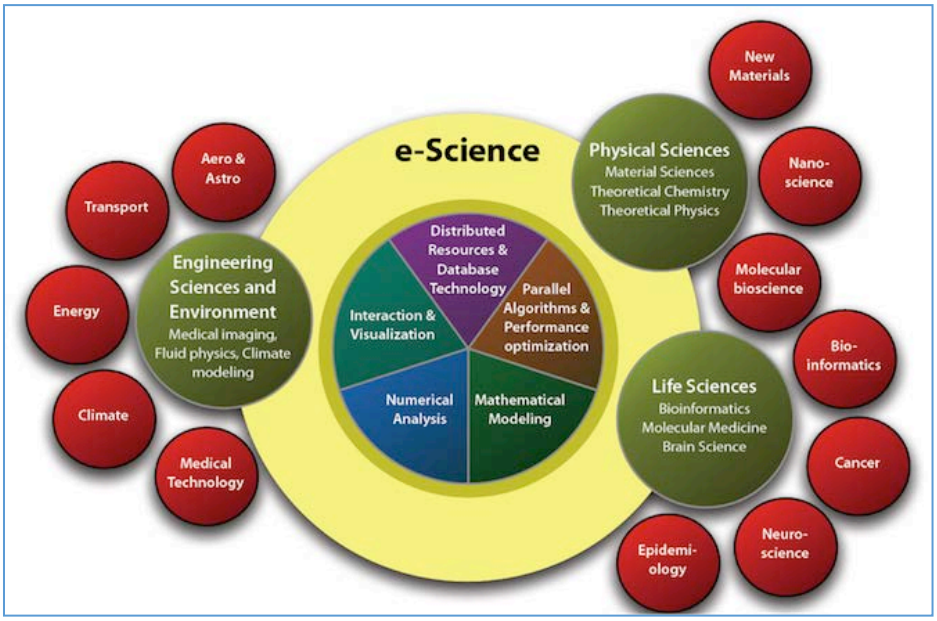


Figure 1 : Illustration of e-Science (by KTH Royal Institute of Technology in Stockholm[3])

E-Science affects all disciplines and research domains, even if some of them, such as life sciences or engineering, are more data-driven than, for instance, arts and humanities. But differences between different disciplines are diminishing and data is important for all of them, as research is based on it.

More than twenty years ago, academic publishing left the Gutenberg era and went digital. The digital revolution was the condition to enter the world of the “4th paradigm” of science where e-

[1] Cf. the French Secretary of State for Digital Affairs Axelle Lemaire opening Big Data Paris 2015 http://www.digitalforallnow.com/en/big-data-paris-2015-fuel-of-the-digital-economy/

[2] EU High Level Expert Group on Scientific Data, 2010. *Riding the wave. How Europe can gain from the rising tide of scientific data*. European Union, Brussels. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[3] https://www.kth.se/en/forskning/forskningsplattformar/ict/forskning/e-vetenskap-1.323973

infrastructures enable “data-intensive scientific discovery” through data mining and integration of theories, simulations and experiments (Hey et al. 2009). Scientific information has become a continuum between publication and data. Linking data to documents is crucial for the interconnection of scientific knowledge. One can imagine this inclusion of datasets and other materials as the “perfecting of the traditional scientific paper genre (...) where the paper becomes a window for the scientist to not only actively understand a scientific result, but also reproduce it or extend it” (Lynch 2009). The possibility to reproduce research outcomes, i.e. the ability of an entire experiment or study to be duplicated, has always been the basis of science and scientific research. Today, the availability of the research data contribute to this necessary reproducibility. At the same time, it may prevent plagiarism, fraud and falsification of data.

But what exactly is research data? What does it mean? There is no clear or unique definition of the term. Following the US OMB Circular 110[4], research data can be considered as “the recorded factual material commonly accepted in the scientific community as necessary to validate research findings.” The international directory for research data repositories re3data[5] distinguishes between fourteen different types of data (archived data, audio-visual data, configuration data, databases, images, network-based data, plain text, raw data, scientific and statistical data formats, software applications, source code, standard office documents, structured graphics, and structured text) but admits that there are other categories in the nearly 1,400 indexed repositories.

Large research projects, laboratories and technical apparatus produce what is commonly called “big data”, i.e. research data characterized by their important volumes, their availability in real time (velocity) and their large variety (Laney 2001). Yet e-Science is not only “big data”. The concept applies also to data output from individual scientists, smaller projects and research teams especially but not exclusively in social sciences and humanities. These data, defined as reusable research results, collected, observed, or created for purposes of analysis to produce original research results (University of Edinburg, cited by Burnham 2013), are produced in a large variety of formats, sources and types.

2.Data life cycle

Research data are part of the dynamic process of scientific research and discovery. In a very schematic and simplistic manner, two different functions of data can be distinguished in the research process:

- Data as material (input): at an early stage of the research process, data are collected and analysed from different sources and in different ways and formats, as material for exploration and hypothesis testing.
- Data as results (output): other data are produced during the whole process and at the end, together with publications, as research results.

The original raw or “input” data often are transformed and processed into derived products (metrics, graphics etc.). At the same time, research data, especially as research results (output) follow their own dynamic that can be described as a data life-cycle (figure 2).

[44] https://www.whitehouse.gov/omb/circulars_a110/
[5] http://www.re3data.org/

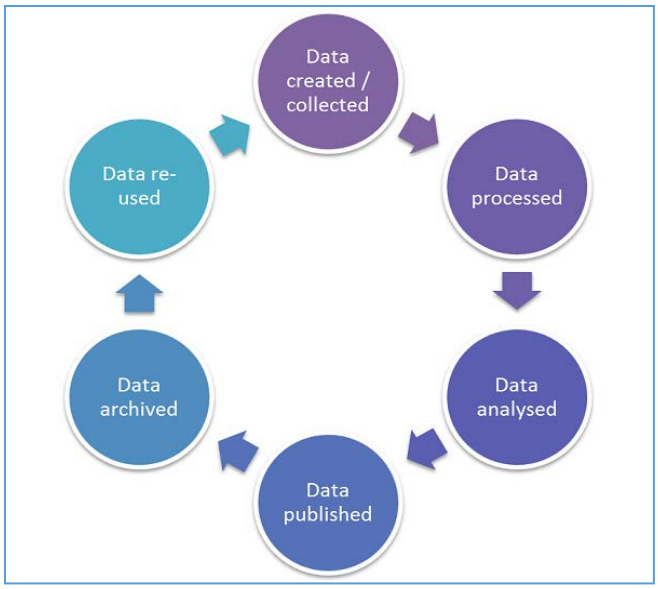


Figure 2: Data life-cycle (by Lancaster University Library[6])

We will not describe the whole life-cycle in detail, because of its complexity[7] and its great variability. "Although there may be significant differences in the individual stages, the life cycle is assumed to encompass the experimental design and capture, cleaning/integration, analysis, publication, and preservation processes, which occur in an iterative fashion" (Kowalczyk & Shankar 2011, p.251). We will just highlight two specific aspects:

- Data integration: Data integration means the process by which "disparate types of data (…) are identified and stored in a manner that facilitates novel associations among the data" (Bult 2002). This is more than preservation and sharing and goes beyond the capacities of most data repositories.
- Evaluation: Research data, as a part of the scientific "output" and together with publications, become increasingly involved in research assessment procedures, as for instance in the research portal of the King's College of London which integrates a current research information system and an institutional repository with published and unpublished results, including datasets[8].

Partly, research data management reflects the selection criteria of funding agencies and other scientific structures, with mandatory data management plans, long-term preservation guarantees and data sharing in open access.

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 (Neuroth et al. 2013). It is generally seen as a culture challenge to the whole profession, due to the lack of the skills and attitudes acquired and needed to cope with it (Cox et al, 2014). They contribute to the assessment of

[6] http://www.lancaster.ac.uk/library/rdm/plan/data-lifecycle/
[7] See for instance the much more detailed model of the JISC Digital Curation Centre, presented by Higgins (2008).
[8] https://kclpure.kcl.ac.uk/portal/en/

data management practices and needs (Simukovic et al. 2014) and to the development and evaluation of data repositories (Pampel et al. 2013, Lynch 2014).

3.Data and/or publications

As said above, e-Science consists mainly of data and not of literature or documents (Hey & Trefethen 2005). The digital revolution deconstructed the unity of the text, eroding the notion of a monolithic 'document' in the hypertext paradigm and disintegrating the article in several distinct elements. At the same time and partly due to fragmentation, authors and publishers growingly enrich the article with new contents and features, such as multimedia, collaborative tools and data (Cassella & Calvi 2010). Some even predict that publications, as traditional vectors for scientific communication will disappear in favour of a direct communication between machines, at least in some specific research fields: “In the age of genomic-sized datasets, the biomedical literature is increasingly archaic as a form of transmission of scientific knowledge for computers” (Blake & Bult 2006). This may seem a little bit too futuristic, especially in the context of social sciences, arts and humanities where publications so far preserve their central role for the transmission of knowledge.

Until now, data were part of publications as support for argumentation and hypothesis testing or for illustrative purpose. In digital scholarship, new publication formats integrate data that can be updated, enriched, extracted, shared, aggregated and manipulated (McMahon 2010). Publications become live documents. The “next generation journal” will be enhanced and interactive, with video reference material, multimedia interactive graphs, links to datasets etc., making the article “more desirable for readers and more effective with regard to acquisition of the information” (Siegel et al. 2010, p.28). Publications become windows on research results.

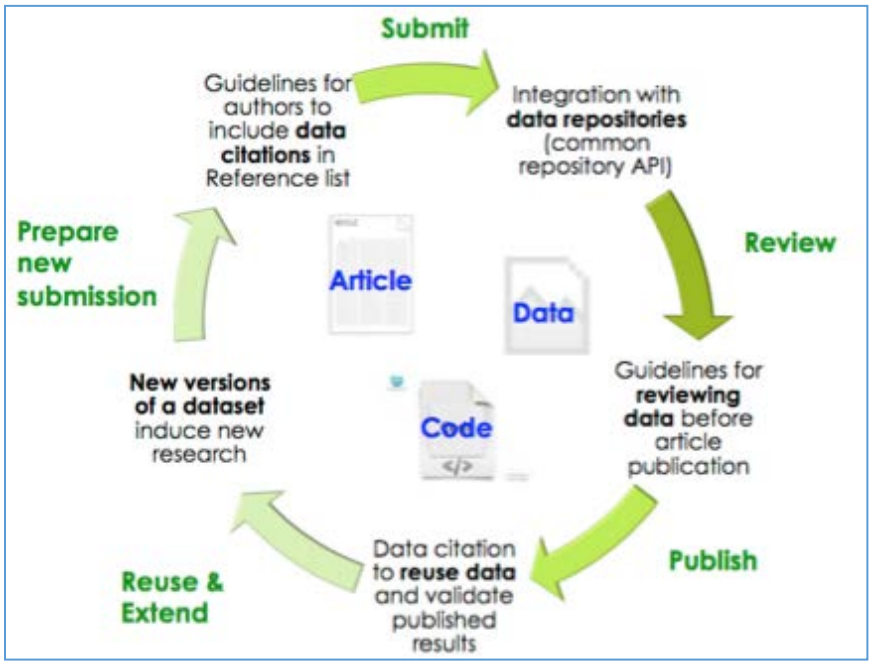


Figure 3: Diagram of an Automated Integrated Article and Data Publishing Workflow (Dataverse Project, Harvard University[9])

[9] http://datascience.iq.harvard.edu/blog/bridge-publishing-words-publishing-data

However, as figure 3 illustrates, other links are possible between data and publications. Perhaps the "data deluge" will not substitute academic publishing. Although we can be sure that it will change the way how academic publishing is today. At least three different ways can be distinguished in which documents contribute to data production and e-Science.

- Document as data: Documents, such as conventional articles, ETD, reports and conference abstracts or proceedings are exploited as primary data source for text mining, automatic extraction of meaningful information, intelligence etc. "Scientific journals will increasingly use standardized language and document structures in research publications" (Morris et al. 2005). The same remark applies to grey literature, including theses and dissertations[10] (see Murray-Rust 2007).
- Data vehicle: Enhanced publications or companion versions of published articles can serve as data carrier or database for content-dependent cross-querying of literature. For example, enriched articles can contain 'lively' and interactive content such as "interactive figures, semantic lenses revealing numerical data beneath graphs, pop-ups providing excerpts from cited papers relevant to the textual citation contexts (or) re-orderable reference lists" (Shotton 2012). Supplementary information files are available from the journal Web site, and/or the figures and tables containing research data within the article are available for download. Yet, their format does not necessarily allow or facilitate liberal reuse and exploitation.
- Gateway to data: Increasingly publications contain links to research data, either in the text or as part of the metadata. The reader (user) of the document can access the underlying research results but the data are not integrated into the document and both – data and document – can be used and reused separately. The full research datasets are published in a permanent archive or repository, with a unique identifier, with an open access data license or public domain dedication, and with sufficient descriptive metadata to enable their re-interpretation and reuse. This link can also be established when connecting a bibliographic database with a data repository, as INIS does with the Fukushima data archive (Savic 2015).

These different ways to link data and publications are represented in the STM data publication pyramid (figure 4). At the base line, raw data are just exploited for the publication of research results but neither cited nor connected. They remain hidden, even if they may be made available to other scientists and/or peer reviewers, on demand.

[10] In the following, we will use the term dissertation to designate the written contribution to obtain a PhD degree.

At the top level of the pyramid, data are published together with the article (or report, dissertation etc.). The third level, well-structured and documented databases and data collections foster a new vector of data publishing, i.e. so-called data journals, peer-reviewed journals for the publication of articles (data papers) on original datasets and collections[11].

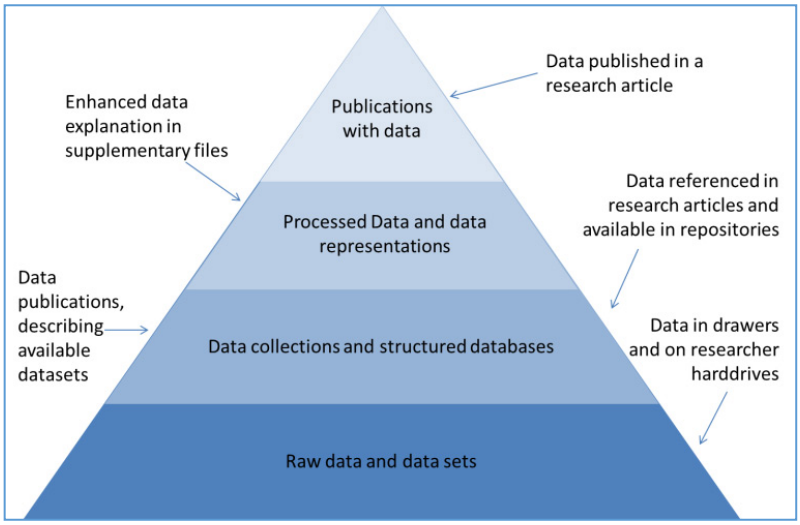


Figure 4: Data publication pyramid (from Reilly et al. 2011[12])

One part of these data is freely available, especially on servers that meet the criteria of open access. However, the access to many other research data is restricted or impossible. Savage & Vickers (2009) complain that the accessibility and the potential of datasets for reuse are often neither optimal nor effective, because of failing standards, metadata, identifiers or services. As for documents, availability and openness of data is not a simple on/off concept but a continuum between more or less open and restricted solutions ranging from the simple availability on the web of data on the lower end of the scale to data dissemination in non-proprietary formats and open standards as optimal openness.

4.The challenge of ETDs

While academic publishers make usage of new technologies to enrich the content and functionalities of their online products, universities can seize the opportunity of the supplementary files submitted together with electronic theses and dissertations (ETDs). Data sharing, long-term data storage and enrichment of dissertations by summary videos or data files were major topics of the 2015 USETDA conference at Austin, Texas, and the 2016 International Conference on Electronic Theses and Dissertation will be mainly about data and dissertations.[13]

Significant part of academic grey literature (Schöpfel & Farace 2010), produced and published by universities, dissertations are documents submitted in support of candidature for a PhD or doctorate degree presenting an author's research and findings (Juznic 2010). Theses and dissertations are “the most useful kinds of invisible scholarship and the most invisible kinds of useful scholarship” (Suber

[11] See for instance the list on the FOSTER website https://www.fosteropenscience.eu/foster-taxonomy/open-data-journals
[12] Figure from https://www.elsevier.com/connect/can-data-be-peer-reviewed
[13] http://etd2016.sciencesconf.org/

2012). However, more and more dissertations are available in open access through institutional repositories (Sengupta 2014), i.e. open archives "serving the interests of faculty – researchers and teachers - by collecting their intellectual outputs for long-term access, preservation and management" (Carr et al., 2008). The typical life-cycle of ETDs today includes institutional repositories as main vector of dissemination (figure 5). In November 2015, the international directory OpenDOAR listed more than 1,500 institutional repositories with electronic theses and dissertations (60%). The academic search engine BASE provides more than 3.9 million ETD via the OAI-PMH protocol. "At many institutions ETD are simply the lowest hanging fruit and new submission batches can generally be counted on each semester" (McDowell 2007). The European DART-Europe portal[14] gives access to more than 600,000 open access research ETD from 586 Universities in 28 European countries while the new international search engine "Global ETD Search" by NDLTD[15] references 4.3 million theses and dissertations.

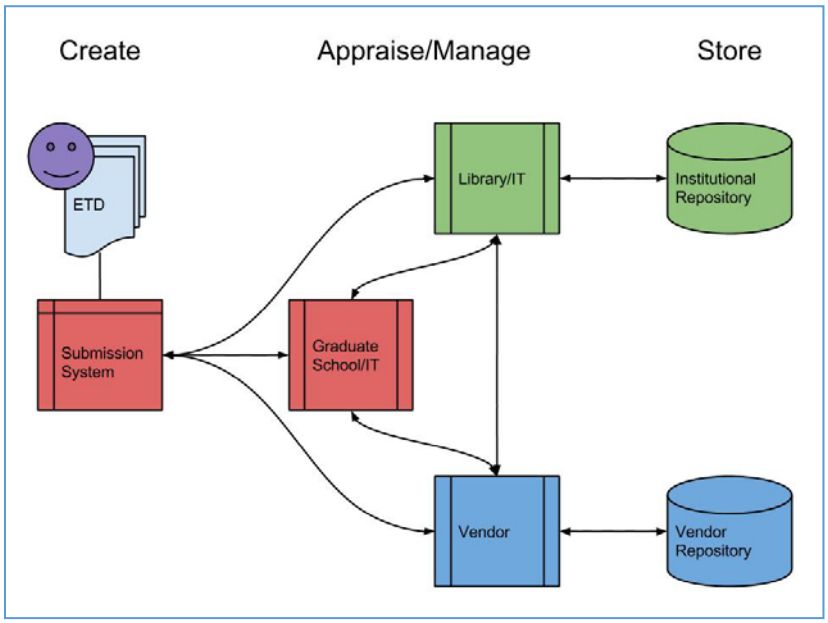


Figure 5: Life-cycle management of ETDs (Educopia project[16])

PhD dissertations contain the results of at least three years of scientific work. It is the result of cooperative work, between the PhD student and his tutor in the first place, but more generally, accomplished within a laboratory, a research team or an institute, school or company. These results may be presented as tables, graphs etc. in the paper or as additional material (annex). In the past, print dissertations have regularly been submitted together with supplementary material, in various formats and on different supports (print annex, punched card, floppy disk, audiotape, slide, CD-ROM…). Today, such material is submitted and processed as "complex content objects" (Schultz et al. 2014) together with the text files or as supplementary files in various formats, depending on disciplines, research fields and methods. If disseminated via open repositories, these research results could become a rich source of research results and datasets, for reuse and other exploitation. Thus, research results produced by PhD students could contribute to e-Science. However, there are three barriers.

[14] http://www.dart-europe.eu/
[15] http://search.ndltd.org/
[16] http://educopia.org/research/electronic-theses-and-dissertations

- First, dissertations must be freely available in open access, deposited in institutional or other repositories and disseminated with sufficient user rights to allow re-use. However, up to now a significant portion of the digital dissertations are not online, not open, not freely available but embargoed or under restricted access (Schöpfel et al. 2015a).
- 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).
- Dissertations are a most important part of the scientific community, despite various critiques of both the romantic notion of authorship and the epistemological assumptions that form traditional notions of independent scientific and scholarly research. This makes it hard to define "authorship" regarding data produced with dissertations.

When this material is submitted as a kind of data appendix, the dissertation becomes a "data vehicle", where data are published together with the dissertation or as a part of it. Sometimes the data are available on a distant server and without the text of the dissertation, transforming the dissertation in a "gateway to data". Yet, too often the data are simply not available; or data, methodology, tools, primary sources are mingled, not indexed, badly described, and unrelated with the text, unconnected with other files.

Often, dissertations will be somewhere "in-between" with some data integrated in the text (tables, graphs, illustrations) and others published as annex. One example among thousands: a dissertation in archaeology from Slovenia contains plenty of photographs, maps and tables and has an annex with an excavation map, the results of chemical equations and a comprehensive description of analysed bodies. Beside this annex there is another volume with almost 300 pages of photographs from the excavations, drawings of the objects and their descriptions. This could be a perfect example for reuse of this comprehensive data if the data would exist in digital form, especially with the suitable searchable platform. Even so, text and data mining would be necessary to identify and exploit all this information, and the dissertation is at the same time "data vehicle" and primary data.

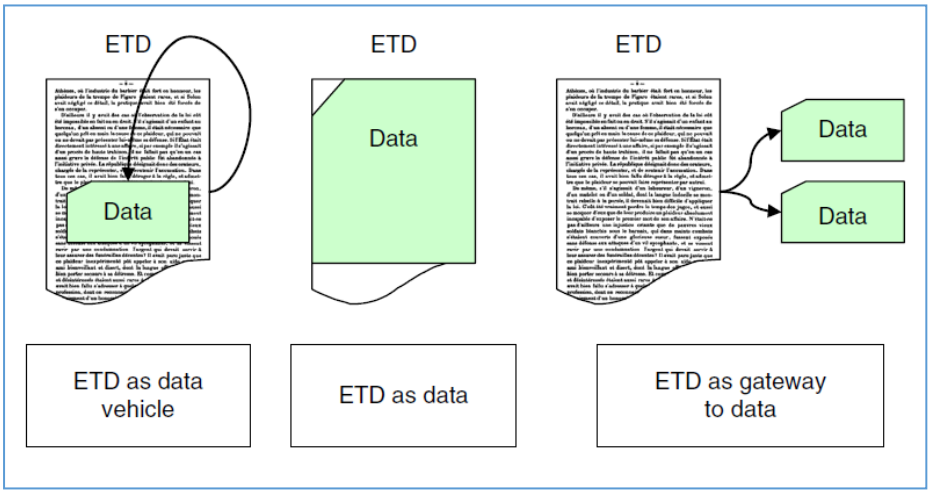


Figure 6: ETD as data, data vehicles and gateway to data

Obviously, ETDs can contribute to e-Science as primary data source, they contain (potentially) interactive content, and they are linked to datasets and research results (figure 6). But linking does not necessarily mean providing access. Supplementary material from ETD does not often even meet the criteria of simple availability on the web with an open licence, for different reasons, including dissemination under copyright or with more restrictive licences. This material continues to challenge the academic library. Research results, methodology, tools, primary sources are mingled, often not indexed, badly described, unrelated with the text, non-connected with other files, and virtually unavailable. In an academic environment that claims open access not only to scientific publications but also to research results, this situation is not satisfying.

5.Empirical evidence on data and dissertations

Up to now, there is very few empirical evidence and insight in the field of research data and other content related to dissertations, and only a small number of research projects and innovative workflows have been documented so far. Two exceptions are the Dataverse pilote project at the Emory University (Doty et al. 2015) and the ongoing research project ETDplus on preservation and curation of ETD research data and complex digital objects, funded by the Educopia Institute.[17] In the following we present some empirical results from surveys conducted by the Universities of Lille 3 and Ljubljana.

5.1.Research data management: practice and needs

In a survey on research data management and sharing in social sciences and humanities 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 - 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. These results are compliant with a German survey with 117 PhD students of different disciplines (STM and SS&H) at the University of Humboldt, Berlin (Kindling 2013).

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

[17] https://educopia.org/research/grants/etdplus

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.

5.2.Research data in dissertations: a French-Slovenian survey

Following a first survey on 283 dissertations from the University of Lille 3 between November 2014 and January 2015 (Prost et al. 2015), we analysed two other, complementary samples:

- ETDs in the fields of social sciences and humanities from the Universities of Lille 1, 2 and 3.
- Dissertations in the fields of social sciences and humanities from the University of Ljubljana.

The survey is based on a German research project at the Humboldt University at Berlin (Simukovic et al. 2014). The methodological approach has been described by Prost et al. (2015). Altogether, the sample contains 780 digital and print dissertations from more than 15 different disciplines (figure 7).

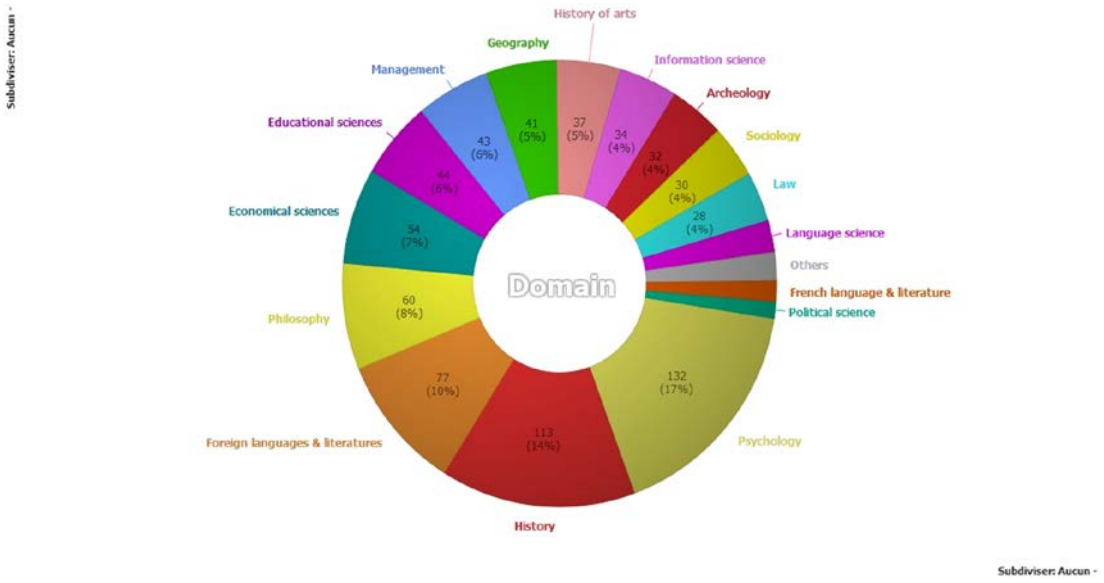


Figure 7: Scientific disciplines of the survey sample (N=780 dissertations)

The sample consisted of 353 digital dissertations (45%) and 427 print dissertations (55%), from 1987 to 2015. In our sample, Psychology, History, Foreign Languages and Literature (English and American, Spanish, Slavonic, Hebrew…), Philosophy and Economics were the most represented disciplines, followed by Education, Management, Geography and History of Arts.

All dissertations have been analysed either in digital or print format or on microform. Each dissertation has been checked by at least one of the authors, either in the library holdings (print or microform) or on the institutional repository server. We tried in particular to identify research data added to the end of the dissertation. In our sample of 780 dissertations, 522 contain one or more appendices with some kind of research data (67%). The length of these appendices varies widely, from one to 829 pages, with a median of 37 pages, and totalling more than 45,000 pages (Figure 8).

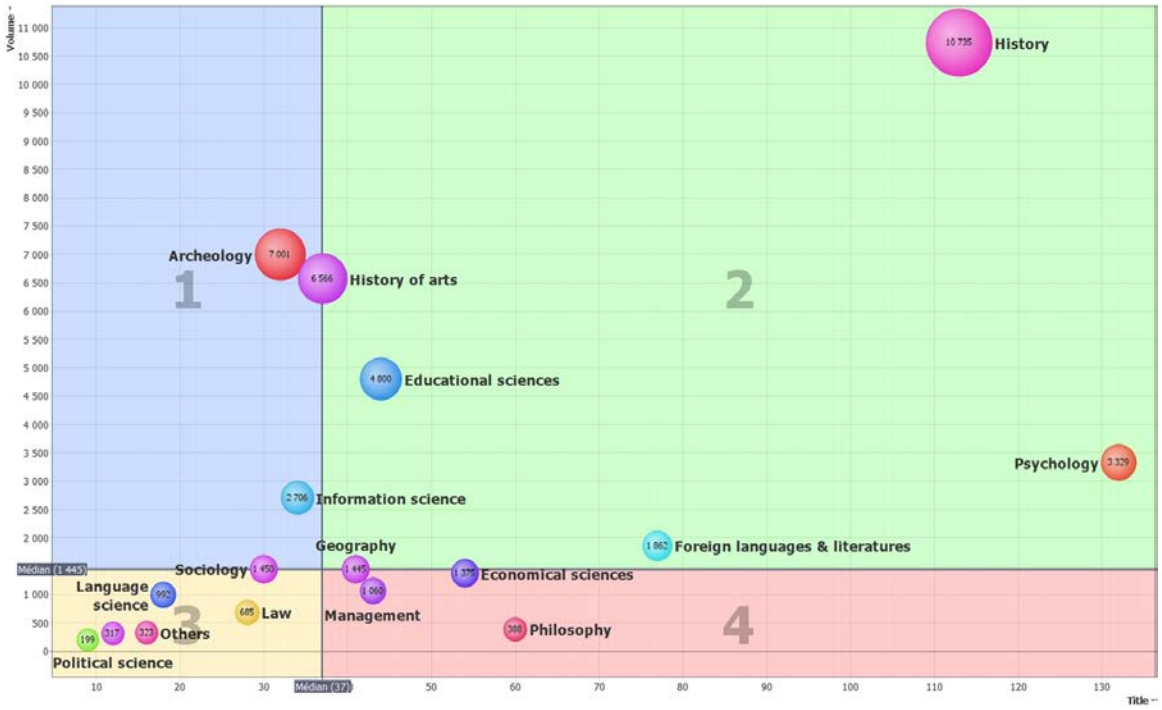


Figure 8: Size of data appendices (in pages, N=522 dissertations)

Even if each appendix holds some kind of research data, this does not mean that one can find research results (data) stricto sensu on all pages. Some pages contain empty questionnaires or survey forms, experimental procedures, bibliographies etc. which cannot be considered as data.

5.3.Disciplines

In the first Lille survey, the distribution of disciplines per dissertation with appendices was more or less the same than for the overall sample, with a significant linear determination coefficient R2 between both variables of .91 (Prost et al. 2015). The differences were not significant – in some domains such as Psychology, Philosophy and Linguistics, we found fewer dissertations with data appendices than the average (67%); in others there were relatively more (Information Sciences, History of Art). In Education and in Archaeology and Egyptology, all dissertations of the sample contained some form of data appendices.

Yet, in the larger sample the differences between disciplines are elsewhere (Figure 8). Some disciplines "produce" rather large appendices, with an average number of pages above the mean of the whole sample, such as History, Education and Foreign Languages, while others most often contain shorter appendices (Sociology, Law, Political Sciences…).

5.4.Support, presentation and format

In France, all files of digital PhD dissertations should be deposited with the text, and the French national computer centre for Higher Education (CINES) maintains a list of accepted file formats for long term preservation. However, there is no control of this deposit, if really all files with data and other material have been deposited or not. Also, nearly all files are in PDF (image or text), and other formats are very rare. In the French sample, only one dissertation has been submitted with audio-visual files in audio and video file formats on CD-ROM.

The French and Slovenian official guidelines for PhD dissertations do not specify how to structure or present an appendix. Some dissertations have poor or no table of contents for their appendices, like

a dissertation in History with a table of content for the text volume but not for the two volumes that contain rich material, including 1,581 figures and images.

As mentioned above, 45% of our samples are electronic dissertations. Compared to the print dissertations, they contain slightly more appendices (Figure 9).

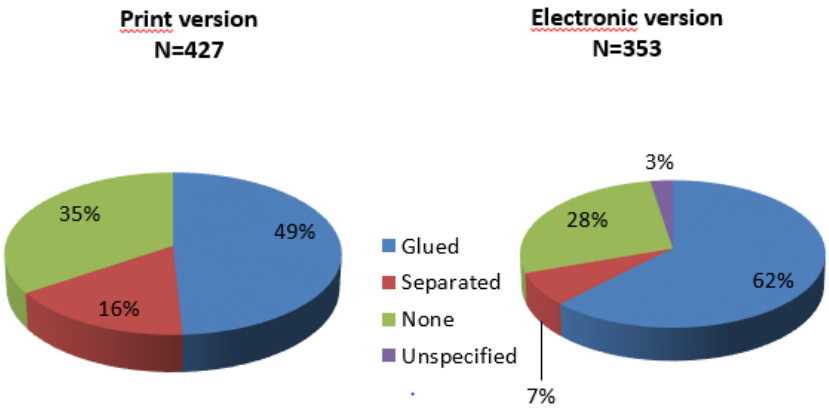


Figure 9: Link between text and appendices (in %, N=780)

Also, electronic dissertations often do not separate text and appendices but glue them together into the same file (62%), worse than the presentation of appendices in print dissertation (49%). Here the dissertations are not gateways to data but play the role of data vehicles, yet with data that are more or less useless or rather, not really reusable because of the format and missing structure.

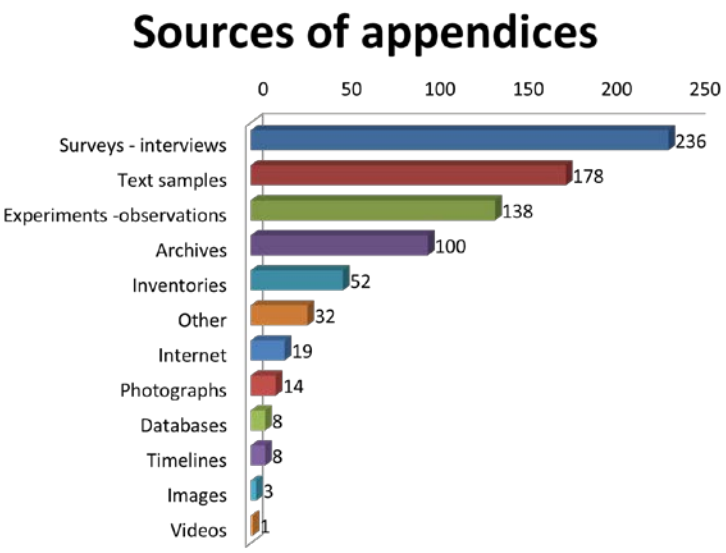
Nevertheless, some dissertations demonstrate a real effort of data management and curation by the PhD student. For instance, a French dissertation in Egyptology on late Egyptian steles contains an exhaustive inventory of those steles on CD-ROM with indexing of geographical origin, general characteristics, specific particularities and dating, together with the transcription of the inscription and a justificatory supporting the provenance of the stele. The PhD student also delivers a user manual for the navigation in the database. Two examples from Slovenia: a dissertation in archaeology (Early Iron Age) provides short guidelines for the use of the annex which encompasses more than 50% of the dissertation; in a dissertation in history of arts, the annex is on a CD-ROM together with an installation file for the software needed to open the content files.

Another French dissertation in Linguistics presents a diachronic analysis of the vocabulary from 49 political speeches and 10 manifestos, pamphlets and articles, with a lexical analyser software (Wmatrix corpus analysis and comparison tool). The appendix contains the complete list of all words with their frequency of usage ranking.

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 at the University of Lille 3: a dissertation on the Renaissance elite of the old Flemish town of Douai 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.

5.5.Research data sources

The PhD students used a great variety of sources for their scientific work, with four major data



sources, i.e. surveys, text samples (corpora), experiments (observations) and archives (Figure 10).

Figure 10: Data sources per dissertation (N=780)

Other less exploited sources are inventories, Internet sources, photographs and images, databases, timelines and videos. The distribution of data sources is to some extent specific for each discipline.

	Archives	Databases	Experiments - observati...	Images	Internet	Inventories	other	Photographs	Surveys - interviews	Text samples	Timelines	Vidéos	Tous
Archeology	1		14			25		5		3			30
Economical sciences		1	27		3	1			14	18			43
Educational sciences			6		3			1	33	13	2		38
Foreign languages & literatures	4		5		1		11		14	35	1		46
French language & literature									2	3	1		6
Geography	3	2	27		1		5		15	8		1	33
History	76		4	1		7		2	9	21	2		88
History of arts	10		4			17	7	2	3	4	2		28
Information science	1	1	1		6		6		17	10			28
Language science			1						3	5			7
Law					2				2	5			7
Management	1	3	7		1	1			21	14			30
Others		1	4						1				6
Philosophy	1		2					1	1	9			11
Political science									6	3			6
Psychology			30				3	2	71	12			91
Sociology	3		6	2	2	1		1	24	15			28
Tous	100	8	138	3	19	52	32	14	236	178	8	1	526

Figure 11: Data sources and disciplines (N=780)

Here are some examples of heavily used sources:

- History: archives, text samples
- Psychology: surveys, experiments
- Philosophy: text samples
- Foreign Languages and Literature: text samples
- Information and Communication Sciences: surveys, text samples, Internet
- History of Art: inventories
- Linguistics: text samples, surveys
- Archaeology and Egyptology: inventories, photographs

However, the situation is more complex, and Figure 11 reveals as well specific data-profiles for each discipline as disciplinary profiles for each data source: inventories for instance are typical for archaeology and history of arts, experiments and observations are specific for psychology, economics and geography, and so on. These are typical research data sources for the social sciences and humanities. Compared to the Berlin survey, other data sources like simulations, statistics, reference data or log files (usage data) are unusual or missing. For instance, many PhD students from the Humboldt University reported that they made use of measurement series, statistical analyses and spectra (Kindling 2013) – except for the statistics, we did not find such data sources in our SS&H sample.

5.6.Typology of research data

Which are the research data produced by the PhD students and present in the appendices? Our evaluation reveals several different and heterogeneous data types (Figure 12).

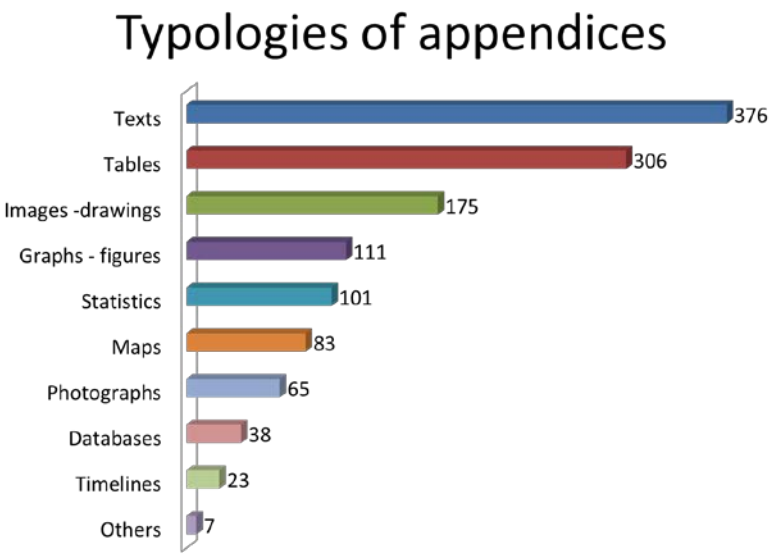


Figure 12: Data types, per dissertation (N=780)

However, two types of data are significantly more produced in these dissertations, i.e. text samples and tables (spreadsheets). Other data are produced in form of images, drawings, graphs and figures, while statistics, maps, photographs, databases and timelines (chronologies) are also part of the data

appendices but less often. We found only one dissertation with audio-visual media (interviews) and no dissertation at all with geolocation data.

Again, as for the data sources, there are some discipline-specific data type profiles (Figure 13). In some disciplines, one or two data types are predominant. This is the case in Philosophy and Language Science where text samples represent more than half of the data. Other disciplines are characterized by a wider number of different research data. Some examples (in brackets, the percentage of this data type for all data appendices in this discipline) :

- History: ten different data types, including text (74%), tables (50%) and images (44%).
- Information and Communication Sciences: ten different data types, including text (71%) and tables (43%).
- Archaeology and Egyptology: nine different data types, including images (73%), maps (60%), text (50%) and photographs (37%).
- Psychology: eight different data types, including tables (71%), statistics (60%) and text (53%).
- Economics: six different data types, including text (84%), tables (72%) and graphs (37%).

	Databases	Graphs - figures	Images - drawings	Maps	Others	Photographs	Statistics	Tables	Texts	Timelines	Tous
Archeology	4	2	22	18		11	1	16	15	1	30
Economical sciences		16	1	5			2	31	36		43
Educational sciences		8	14	1			5	25	29	1	38
Foreign languages & literatures	1	1	20		1	1	6	21	36	1	46
French language & literature		1					1		5	1	6
Geography		13	7	13		5	3	27	23		33
History	16	22	39	27		26	14	44	65	12	88
History of arts	6		17	8	1	8		4	20	1	28
Information science	2	7	7	3	4	2	5	12	20	1	28
Language science	1	1	1				1	1	7		7
Law		1	3	2				4	5		7
Management	2	12	10	1		1	7	26	22	2	30
Others	1	2						2	4	1	6
Philosophy		2	2		1	1		1	11		11
Political science	1	1	4				1	6	2		6
Psychology	2	15	20	1		4	55	65	48		91
Sociology	2	7	8	4		6		21	28	2	28
Tous	38	111	175	83	7	65	101	306	376	23	526

Figure 13: Data types and disciplines (N=780)

The research data are very different. Some examples: a great number of images and photographs on the religious life in the French town of Etaples from the beginnings to 2000, statistics on prisons and prisoners in Northern France during the French Third Republic, the mapped tours and comments of children in a dissertation on two exhibitions, or a large corpus of old documents and archaeologic findings for the reconstruction of the organisation of banquets in Anglo-Saxon England. In another Slovenian example from history the annex contains on 400 pages a comprehensive list of short bibliographies from priests living in Carniola during the second half of the 18th century.

Some data types are present in all or nearly all disciplines, like text samples, tables, images or graphs – an observation which confirms the Berlin results where texts, tables (spreadsheets) and databases

were dominant. Others, in particular inventories or audio-visual material, are at least in our sample specific for one or two disciplines. We compared print dissertations and e-dissertations and performed a chi-squared test but found no significant differences neither for research data sources nor for data types (on .05 level). Obviously, these differences are more related to disciplinary methodologies than to support.

5.7.The special case of ETDs in engineering sciences[18]

In order to learn more about research data in dissertations, in particular in the field of applied sciences, another study was conducted on 86 ETDs deposited in the institutional repository DRUGG[19] of the Faculty of Civil and Geodetic Engineering of the University of Ljubljana, Slovenia (UL FGG). The repository was established in 2011, it is listed in the OpenDOAR and ROAR directories and compliant with the OpenAIRE infrastructure criteria (Koler-Povh et al. 2014). In September 2015, it contained 2,625 items, mostly Diploma theses (2,180) and Master theses (125) (Koler-Povh & Lisec 2015). 86 dissertations of 100 doctoral dissertations in civil engineering published between 2008–2014 were included in the survey on research data.

At UL FGG all dissertations since 2008 have been archived in print and digital version. All print and electronic versions are identical both in terms of content and layout. In the field of civil engineering, all dissertations contain some kind of research data. Together, we found 18,981 datasets in 86 analysed dissertations. Usually these data are integrated in the content, i.e. they are part of the text (the dissertation as the raw data source). Further, in 28 dissertations (33%), some data are published at the end of the dissertation, as a separate and distinct part of the dissertation. When submitted in digital format, the data are published in a single data file, in PDF or JPEG or, especially for large datasets, compressed in a ZIP file, as requested by the IR DRUGG's editorial board. We found 247 different data files for these 28 dissertations. All files have their own metadata.

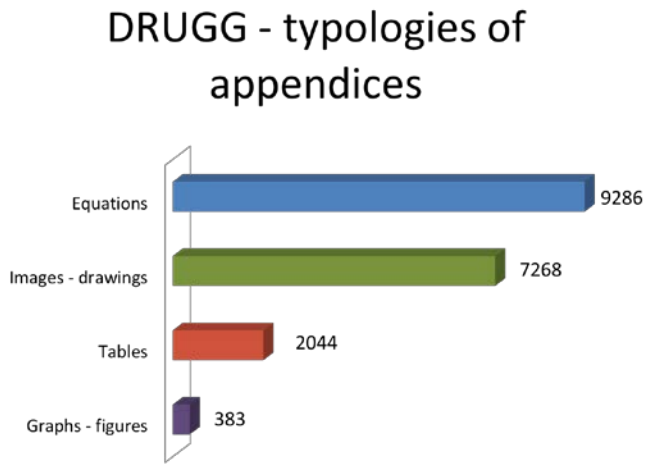


Figure 14: Data types in the DRUGG repository (civil engineering, N=86)

Most datasets published in the dissertations are equations (n=9286), which appear in 76% of all dissertations (66), with a high number in 14 dissertations. There are two dissertations with 424 and 400 equations each. In the former, the number of figures is also high (n=120), in the latter the number of other types of appendices is low, i.e. below 50 for all types.

[18] Study conducted by Teja Koler Povh
[19] http://drugg.fgg.uni-lj.si

There are two more dissertations with more than 300 equations and five with more than 200. Equations are always the most frequent type of datasets. We conclude that in civil engineering equations are the most used type of appendices.

Compared to equations, we found less images, even though they are present in all dissertations. 29 dissertations (34%) contain a high number of this type of data. Following UL FGG's guidelines (Koler-Povh and Turk, 2011), “figure” is a broader term, which includes graphs, pictures, figures, schemes, notes, windows, plans, and charts. Nevertheless, some students use the term “figure” only for schemes, pictures, and photos. In one dissertation the photos are separated from figures. Nine students distinguish between figures and graphs. In the doctoral dissertations from UL FGG, figures are the most frequently used type of appendices; one dissertation contains 252 figures, 11 others published 100 or more figures each. Sometimes figures from other author(s) and sources are used. The counting of such figures is not the same for all dissertations. Two dissertations count imported, i.e. adopted figures, separately from (author's) figures. More dissertations present the adopted figures by stating the source in the legend of the figure, but they do not list them separately in an index of imported figures.

There is also the problem of classifying maps, sometimes they are photos (e.g. satellite photos), sometimes they are maps, many times they are an appendix to the dissertation due to the high paper format.

In 35 dissertations two types of appendices are presented in a similar frequency, for 15 of them the similarity is high (the difference in the frequency is lower than 25%). In the whole sample of 86 dissertations, these 15 present 17%. We can conclude that the number of single type of appendices is regulated by authors on purpose.

6.Potential re-use of research data in dissertations

All these datasets are important to understand a dissertation’s framework, methodology and results. They are helpful for the evaluation of the author’s interpretations and conclusions, and they provide raw material useful to verify and validate the overall research. Moreover, many, if not all data could also 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.

Our study reveals at least three barriers to open data:

- Incomplete, inadequate or missing description of the whole datasets and/or individual data. In some dissertations, especially in History, History of Art and Archaeology, inventories, photographs, maps etc. are well described and indexed. But these are exceptions and often descriptions are simply missing. This problem includes, too, the lack of citability when datasets are not correctly identified.

- 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. In print copies, this means that data are not clearly separated from the dissertation text. In electronic dissertations, this means that data and text are glued together in a PDF file instead of being separated and published in adequate file formats (spreadsheets, image files, text files, database formats, XML…).

Other problems are related to the choice of media, e.g. compact disc, DVD, online server, USB flash drive... For instance, the dissertation on Egyptian steles inform about an online database with restricted access but does not provide login and password. For some retro-digitized dissertations, the online version does not include the data appendix submitted together with the print version.

All these problems make it difficult to find a dissertation's underlying data. The dissertation itself functions more like a kind of barrier instead as a gateway. Without metadata, without identifiers and referencing, it is virtually impossible to find these data otherwise than reading the dissertation. Lack of searchability is a direct consequence of missing data management and curation.

And they make it difficult if not impossible, too, to exploit these data with tools of text and data mining. Text and data mining tools need great volumes of open data, and hidden data in text or protected data files are not really useful for this purpose.

7.Legal aspects

Applying copyright to dissertations is already rather complicated (Schöpfel & Lipinski 2012). And regarding the sharing of research data, "the law makes all of this far more complicated than it need be. For those seeking to pick and choose which reuses of another's data may be permitted by law, regrettably, the answers (…) are more context dependent than many would like" (Carroll 2015). The legal environment of data and other supplementary materials is not the same as for ETD (see Murray-Rust 2008). Both must be considered independently, even if related and interconnected. Non-adapted licensing or (over) protection by copyright can be legal barriers to their deposit, dissemination and reuse. Linking datasets to the copyright protection of ETD creates a potential conflict with open data policy.

The European Commission and several national governments promote the dissemination of datasets under the minimalist open licence, limited to the attribution of authorship (CC-BY). On the other side, authors and service providers of ETD often adopt a more restrictive sharing policies that prohibit modifications and for-profit use, apply the full protection of the intellectual property law or limit the dissemination to campus-wide access (Schöpfel & Prost 2013). This is too restrictive to realize the potential for reuse of data and to be in conformity with the wish of the European Commission to make it "a general rule that all documents made accessible by public sector bodies can be re-used for any purpose, commercial or non-commercial, unless protected by third party copyright."

Some content may be protected by privacy or confidentiality concerns, for instance personal (human) data and sensitive or strategic information, including professional secrecy. Other research results may be subject to specific sui generis database property rights, and sometimes open access policy may be in conflict with legitimate interests (publishing for scientific career) and fear of plagiarism. In any case, author and institution must reconsider the legal condition of the deposit and dissemination of datasets and other material, but they should do so applying a policy of open data allowing for a maximum of reuse and exploitation. Unlike ETD, datasets should not only be free (in terms of the open access movement) but also "libre", i.e. reusable.

In our survey, at least two legal problems are related to the deposit and dissemination of research results in dissertations:

- Privacy 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, at least with regards to the French law they need special processing and careful handling.
- Third party copyright. 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.

Sometimes, the authorship of the research data remains uncertain. These problems should be addressed as a part of doctoral education on data management, well ahead of decisions on preservation and dissemination. Because "restrictions in the use of research data directly affect research data curation (they) must (…) be taken into account right from the beginning (in matters such as policies, technology, etc.)" (Neuroth et al. 2013). Legal requirements, metadata, back and front office of research data handling have to be considered as a whole, interconnected, and interdependent. Important are two aspects: document and data must be distinguished and separated, intellectually, logically and physically; and the whole approach must be designed in a framework of open data, open access and open science.

Last but not least, open access to digital dissertations and data can be helpful to prevent plagiarism, as a specific form of academic integrity breaches. There was always some concern that plagiarism might occur easier and more often, if dissertations are in open access and freely accessible. On the other hand, open access makes it also easier to detect plagiarism, even if some forms of plagiarism are not easy to detect by anti-plagiarism software. Here, open access to data can be helpful when both, content (text) and supplementary data files are considered as a compound dissertation. The originality of PhD should be also proven by open access datasets especially in the cases when these data were collected by the author and are not part of some collective research project, which is usually the case in social science and humanities. This can be a good way to prevent or deter possible falsification of data.

8.ETD processing and workflows

The data publication workflows should be incorporated to dissertation submission process (Vompras & Schirrwagen 2015). But should dissertations and related datasets be processed together or separately? Should they be disseminated on the same or on different repositories? Should they be preserved on the same or on different servers? How should they be linked?

The Educopia Institute's *Guidance Documents for Lifecycle Management of ETDs* (Schultz et al. 2014) suggests the separation of the dissertation text files and the related "complex content objects" whenever possible. "Embedding multimedia components within the full text might seem advantageous in that they would then be inseparable. However, when the time comes that it is necessary to migrate either the full text itself or one of the multimedia components, having separate files would greatly simplify matters" (p.5-9). Metadata and persistent identifiers like handle, PURL, ARK or DOI are supposed to provide the "glue" that binds together the text, multimedia and data files. Preservation-worthy research data (survey data, measurements, laboratory notebooks, measured spectra etc.) might be stored as part of an "ETD package" or, after transformation into archival files, in a separate data repository, and "links to the data repository from the ETD metadata would then enable researchers to access these research data in the future" (p5-11).

The Dataverse ETD pilot program at Emory, Atlanta, is similar to this approach (Doty et al. 2015). In the Emory workflow, dissertation and research data are deposited separately, and while the dissertation is stored and disseminated on a publication (or document) server, the PhD student is invited to submit the research data to an appropriate disciplinary data repository or, if not available, to the Dataverse pilot. The Emory program supports tabular file formats (.xls, .csv, .xlsx, .dta, .R, SAS Files), software code (.hpp, .py, .rst, .cpp) and geospatial data (.mxd, .kmz, .kml, .gpx, etc…). In a very schematic way, the ETD program runs as follows (Figure 15).

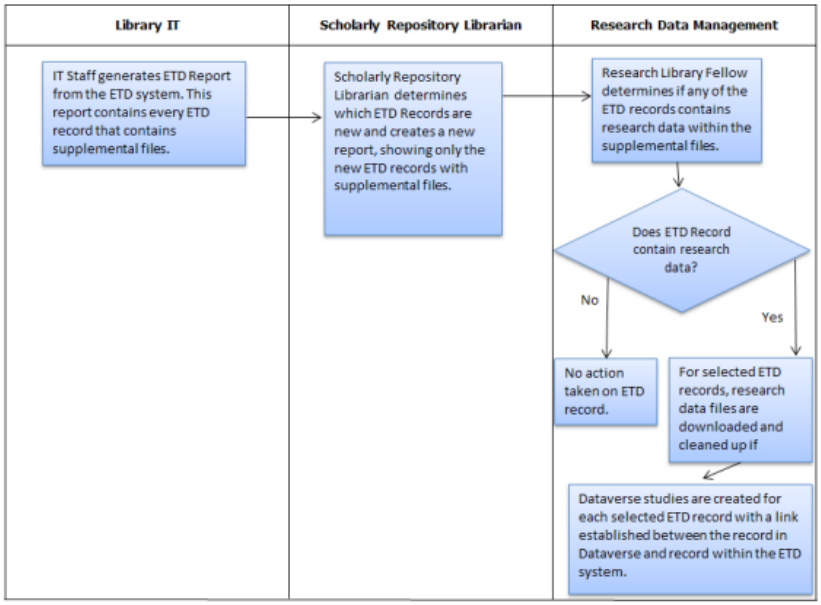


Figure 15: Workflow for Dataverse ETD Pilot Program at Emory (Doty et al. 2015)

The workflow includes also a phase of “cleaning up” research data, i.e. data curation just before the deposit and the connection to the dissertation via identifiers and direct linking between the Datavers pilot and the ETD system.

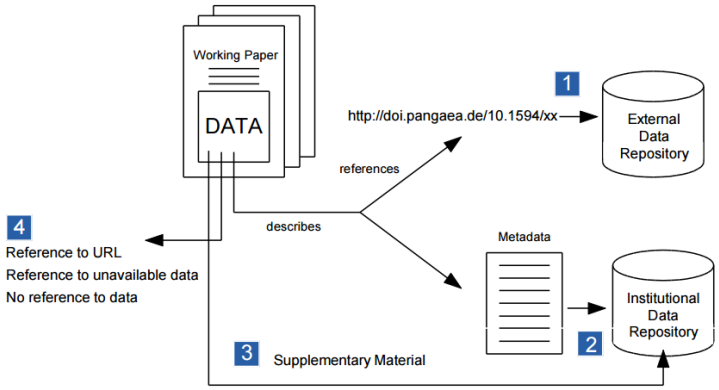


Figure 16: Linking grey literature with research data as discrete resources (Vompras & Schirrwagen 2015)

This process can be done in smooth and simple way, as a growing number of repositories show, such as the Bielefeld PUB[20] or the TU Delft institutional repository.[21] Vompras & Schirrwagen (2015) describe the linking options as follows (here for working papers):

Both workflows have in common that dissertation and data are separated, that this separation is operated before or during the deposit of the dissertation, that the deposit is preceded by data curation, and that dissertation and data are not stored and dissemination on the same repository.

For universities in Slovenia, especially University of Ljubljana, there was a long way how the legal backgrounds have been prepared or revised to support a mandatory process of ETD (Ojsteršek et al, 2014). The process is still not finished and although the problem of research data was identified, due to the other, more basic, legal and even competence and authority problems it was not really tackled yet. This may change as they will have to adapt and embrace the policy of open access that includes also research data. In the Slovenian National strategy 2015-2020 on Open access, research data has become one of the priorities: "The research data financed by public funds should be as far as possible open, accessible with minimal restrictions. Open information must be given to locate or access them evaluate and understand to be useful for others and, if possible, interoperable, coherent with certain quality standards. Open access to research data is relating to the right to online access and re-use of digital research data under the conditions specified in the grant agreements. Accessing, mining, exploitation, reproduction and dissemination are free of charge. Justified exceptions must be explained, for example, in the interests of national security, protection of personal data and intellectual property rights of private co-financiers. Customer Information Control Systems (CICS) must be compliance with legal and ethical requirements to ensure open access. If the access to research data for justified exceptions is limited, at least a freely accessible metadata must be available, from which it is clear where and under what conditions, research data are available."[22]

A particular challenge the existence of two distinct systems, one maintained by the University with its institutional repository and the other by the National Library. The actual laws on university libraries and the National Library do not mention digital dissertations, just that university libraries must obtain and process the compulsory copies of material that is created and published within the framework of the university, including graduate and masters theses and doctoral dissertations, and two copies of (print) doctoral dissertation are to be sent to the National Library. Electronic versions can be uploaded in Digital library of Slovenia, maintained by National and University Library of Slovenia, only with the written permission by authors.

A workflow comparison of the French STAR and the UK EThOS infrastructures with ProQuest's global schema (Walker 2011) and the TARDis project at the University of Southampton (Simpson & Hey 2006, Hey & Hey 2006) suggests that there may be no unique ideal solution but different options, depending on legal and technical conditions. For instance, research data can be handled and disseminated via centralized data management systems or decentralized collaborative systems (social networks) with reduced costs and customizable interfaces (Wang & Liu 2009). Data repositories can be institution-based (such as most ETD repositories) but also run by third-party service providers, such as Dryad, Zenodo or Figshare. One size does not fit all.

As a matter of fact, such heterogeneous datasets cannot be compared to the kind of big data produced by CERN and other large facilities but are more similar to personal data, even if the main

[20] https://pub.uni-bielefeld.de/
[21] http://repository.tudelft.nl/
[22] https://www.arrs.gov.si/sl/obvestila/15/odprti-pristop-20152020.asp (in Slovenian)

challenges are roughly the same, covering issues of up-dates, enrichment and reuse, submission policy, handling of copyrighted material, standards, technical infrastructure or long-term preservation. The point is that the ideal system architecture should combine features of personal data stores (small data) with characteristics of institutional information systems (big data). For instance, how to decide on the inclusion and deposit of supplementary files? In big data repositories with automatic input, these decisions are taken ad hoc and upstream. Because of the link to copyright protected documents, and because of the personal nature of these research results, the same strategy cannot be applied here, and decisions have to be taken on a different level, probably case by case. But on which criteria?

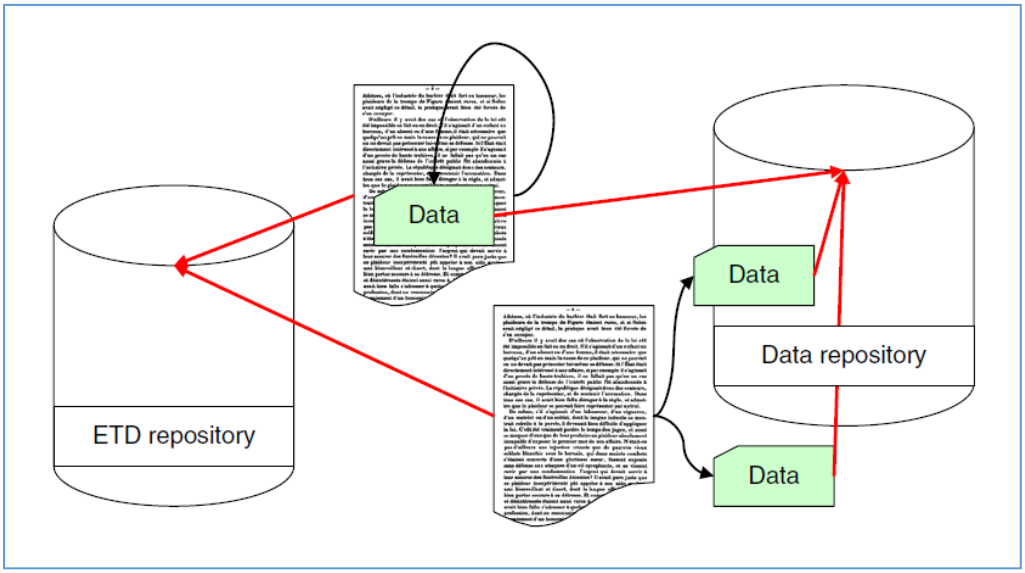


Figure 17: Storage of ETD as related datasets

Because of the specific nature of data and supplementary files (see above), it appears appropriate not to store text and data files in the same repository but to distinguish between document server and data repository and to deposit text and data files on different platforms, or at least to separate them on an early stage of the workflow and to handle them in different information system environments (see Figure 17). For instance, Sun et al. (2011) developed a database and associated computational infrastructure for datasets with different metadata submission forms for different topics. Supplementary material should not only be available as appendix or illustration to the related dissertations but also extractable and reusable without link to the thesis, as an independent dataset and interconnected to other data. In the Berlin survey cited above, scientists seem to prefer a local data repository (department, laboratory) to other solutions which means that they are realistic enough to require a combined institutional and disciplinary environment for their data (Simukovic et al. 2014, Prost & Schöpfel 2015).

9.Helping PhD students to manage their data

Considering these empirical results and the scientific interest, the University of Lille 3 decided to foster the data education of PhD students in social sciences and humanities, as a central part of its

global approach to research data and open access[23]. Following the work of Reznik-Zellen et al. (2012) at the University of Massachusetts Amherst, the Lille project team develops three tiers of research data support services for PhD students on our campus, including education, consultation and infrastructure (Figure 18).

Figure 18: Research data support services

Education: The University of Lille 3 organized three conferences on research data between February and April 2015, especially designed for PhD students in social sciences and humanities[24]. A first transdisciplinary doctoral seminar on research data management will be launched in January 2016, with seven units on data management plans, data life cycle, data description, storage, sharing etc. Another seminar will put the focus on data exploitation. At the same time, the project team will edit guidelines for good data practice and make them available for the PhD students, together with frequently asked questions and updates on data management, open data etc. The University of Lille 3 is not the first one to teach data management and exploitation, and it can build on experiences from other campuses, like the "University of Minnesota Data Management Course" with seven modules.[25] Another example is the "Data Management Bootcamp for Graduate Students" workshop series, a joint program of Virginia Tech and four other Virginia universities[26]. This Bootcamp provides data curation training with seven modules, including formats and transformation, data protection, and preservation, sharing and licensing. Also, the University of Virginia hosts a "Graduate Student Data Management Portal" that offers help to understand the research and data lifecycle (cf. Figure 20), practical guidance and links to useful tools.[27]

Advice and assistance: Probably as a part of its future Learning Centre, the University of Lille 3 will develop personalized help and assistance for PhD students, able to provide answers and advice to their specific questions and problems. This might include, too, advice and guidance regarding methods for de-identification of protected, sensitive personal data (health information, surveys etc.), such as the Safe harbour methods and following the Privacy rule[28]. Also, the role of supervisors should be valorised. Yet, a recent study reveals that most of them know a small amount or nothing

[23] See the Lille 3 White Paper on research data in PhD dissertations http://hal.univ-lille3.fr/hal-01192930v1 (Chaudiron et al. 2015)
[24] http://drtdshs2015.sciencesconf.org/
[25] https://sites.google.com/a/umn.edu/data-management-workshop-series/home-1?pli=1
[26] https://www.research.vt.edu/announcements/data-management-bootcamp-offered-graduate-students
[27] https://pages.shanti.virginia.edu/SciDaC_Grad_Training/
[28] http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#standard

about research data management or digital curation, and most of them have no or only limited skills or expertise in this area (Abbott 2015). Moreover, advice and assistance should build on external resources whenever possible, even if the same study points that the "use of specific external resources is low at under 10% and awareness for all specialised external resources was under 20%. This represents a missed opportunity in terms of outsourcing as much training as possible to dedicated experts" (loc.cit, p.15).

Infrastructures: The Lille 3 approach is based on intermediation, not on research and development. Probably, some basic tools for temporary storage and metadata will be developed and implemented on the campus. For instance, this might be a "data vault" for temporary storage of data files and/or a data asset register, synchronized with the institutional repository or the research management system[29]. Yet, the main idea is to partnership with existing data networks and repositories, including agreements if necessary and delegation of the deposit. Sometimes, especially for small "orphan" datasets, the solution may also simply be Zenodo or FigShare.

These three tiers of research data support services will be launched progressively between 2015 and 2018. Their development will follow five guiding principles (Figure 19).

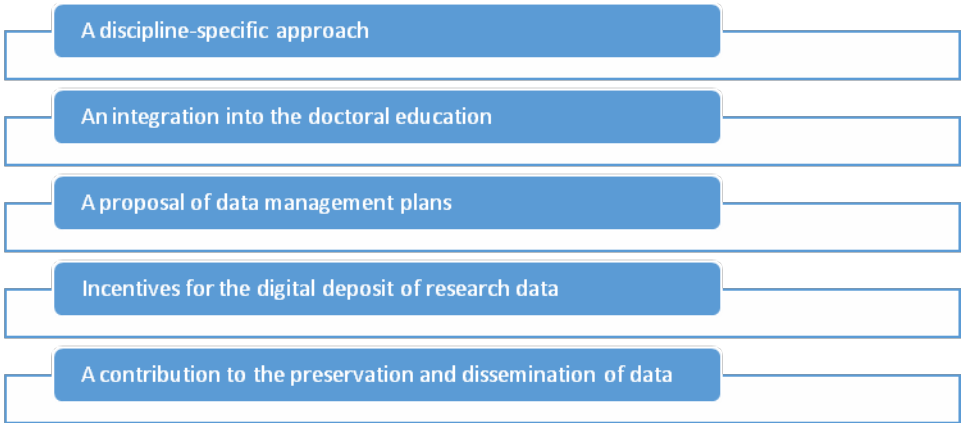


Figure 19: Five leading principles for the implementation of research data support services

1. A discipline-specific approach: 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.
2. An integration into the doctoral education: Data management and sharing must become part of the mandatory doctoral education syllabus, such as project management, scientific writing or data analysis. The Lille 3 data literacy program will contribute to the creation of a culture of data management.
3. A proposal of data management plans: The University of Lille 3 will develop its own templates for data management plans, compliant with social sciences and humanities and the criteria of the European program H2020. The DMPonline tool developed by the JISC Digital Curation Centre to help write data management plans may be a helpful model.[30]

[29] See Stuart Lewis' blog posts on the Ediburgh Research Data Blog http://datablog.is.ed.ac.uk/
[30] https://dmponline.dcc.ac.uk/

4. Incentives for the digital deposit of research data: 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.
5. A contribution to the preservation and dissemination of data: Finally, as mentioned above, the University of Lille 3 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[31] and part of the national open repository HAL[32].

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, scientists and students obviously have not identified the academic library as a potentially useful structure for their data (Prost & Schöpfel 2015). 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 for "data librarianship" (LIBER 2012).

One part of the new library function could be the promotion of research data citation by applying persistent identifiers to research data, such as DOIs. For instance, the Purdue University Research Repository provides DOI for research data. French libraries already assign a persistent identifier (code) to each dissertation. Yet, the ability to connect dissertations with the underlying data needs a consistent way to assign an appropriate level of granularity to sub-sections, appendices and related content. Moreover, due to this new function the academic library may also take responsibility in the field of persistent identifiers for authors (such as ORCID), for instance through assistance and advice for PhD students and young scientists to create and manage their identifiers.

10.Changing the way of doing PhDs

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.

[31] https://www.ccsd.cnrs.fr/
[32] http://hal.univ-lille3.fr/

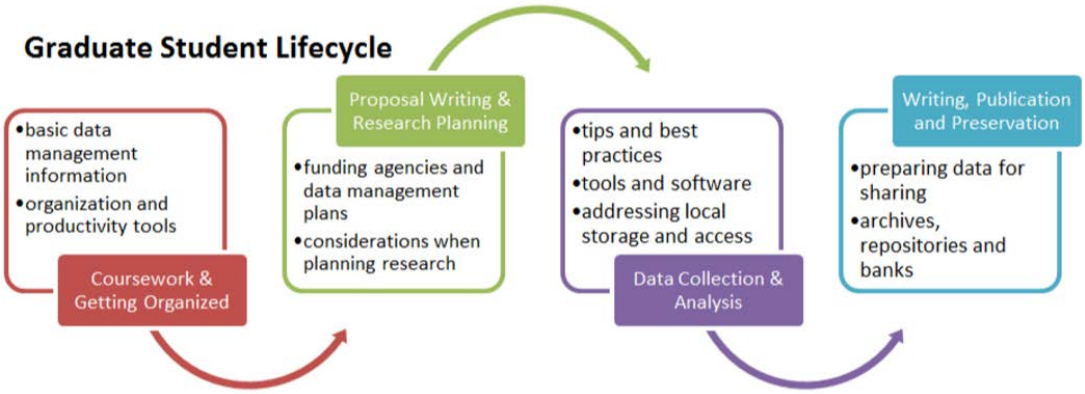


Figure 20: Research and data lifecycle (source: University of Virginia Data Management Portal)

Research data management and data curation cannot wait until the final stage of the PhD project. The student researchers must be aware of the data lifecycle from the beginning on, they must anticipate the legal and technical challenges of their collected and produced data, and they must know how to describe, store and share their data.

In an ongoing survey launched by NDLTD, 52 out of 104 US and Canadian universities that require electronic submission of dissertations allow deposit of supplemental material for doctoral work. In order to do this properly, in a way described above, text and data must be separated, with different metadata and identifiers. Also, the research data files or databases must be well structured and documented, with a detailed and organized tagging (markup) of the datasets. The data must be described in a standard language and format, with sufficient detail for retrieval and data mining.

The PhD students, with assistance from supervisors, colleagues and professionals, must make a thorough choice of formats and supports appropriate for the temporary storage, sharing and future deposit of their data in a data repository. Whenever possible, open formats should be preferred, to facilitate long-term preservation and re-use. Often, data repositories suggest data deposit in the original file format.

We already mentioned the need of clearing of privacy and copyright issues. Just like ethical aspects, these issues cannot wait and must be anticipated from the very beginning on, to be compliant with legal rules and to be able, at the end of the PhD work, to store and share the research results whenever possible.

Data management and curation change the way of doing PhDs, in two ways. The overall planning must include the different stages of the research data lifecycle, from the collection and creation to the preservation and enabling of re-use. On the other side, as the dissertation becomes a gateway to data, the structure and the format of the dissertation must allow the link to related, underlying data. The way to do this will be different between disciplines, domains and research communities. But it seems probable that the text writing and editing of a PhD dissertation will be facilitated because the data and other material moves out of the text. Some dissertations, at least some parts of them, might even become similar to data papers.

11.Perspectives

Open, digital science is work in progress. Along with documents and publications, research data become an essential part of scientific information. Electronic theses and dissertations have the

potential to contribute to the emerging landscape of e-Science, as “data vehicles” as well as “gateways to data”. Higher Education and research organizations invested into infrastructures, repositories and library systems in order to facilitate the transition from print to digital dissertations. Today, new investment is needed for the curation of research data produced and deposited with ETD. The development of ETD infrastructures, open repositories and e-Science makes it possible to find an appropriate solution for the management and reuse of small data produced along with dissertations.

Dissertations often are “data vehicles” where research results are published together with the text of the dissertation. This makes sense in the print world but appears inappropriate in the digital environment of the 4th paradigm. Curation, retrieval and reuse would be largely facilitated if this material would be separated from the PhD text files and handled in a different way. This means, dissertations should be valorised as “gateways to data” which implies incentives for the deposit of related datasets and other supplementary files, minting DOIs (or other persistent identifiers) for research data and innovative procedures and workflows in graduate schools and academic libraries.

Furthermore, service functionalities from institutional repositories and data stores should be adapted to these specific items, with a flexible, user-centred approach. To increase accessibility and reuse and to avoid isolated data silos with multiple metadata entries, all developments should be as standardized as possible and with maximal interconnectivity, based on the OAI protocol. This means also that small data repositories should be, whenever possible, integrated in CRIS environments.

The JISC identified five key areas for actions in favour of research data management UK universities, i.e. policy development and implementation, skills and capabilities, infrastructure and interoperability, incentives for researchers and support stakeholders, and business case and sustainability (Brown et al. 2015). This framework describes the challenges research data projects in the field of dissertations have to face. Even if the basic idea of open access is simple, it is easy to underestimate the cultural barriers and the time required to work through them. The first step is always the hardest. Costello (2009) points out the fact that lack of support is one of the reasons why scientists don’t deposit their data in open repositories. Scientists remain committed to the values, norms and services of their institution and discipline which means that developing an infrastructure for electronic theses and dissertations and supplementary files will be successful if and only if supported by an explicit policy in favour of open access and open data. This policy can be implemented locally and serve as a good example or show case, or nationally as the part of the accreditation systems. Either way, the awareness of the importance of open research data in dissertations should be a good basis for universities to change their policy to PhD dissertations accordingly.

References

Abbott, D., 2015. Digital curation and doctoral research. International Journal of Digital Curation 10 (1), 1–17. http://dx.doi.org/10.2218/ijdc.v10i1.328

Blake, J. A., Bult, C. J., 2006. Beyond the data deluge: Data integration and bio-ontologies. Journal of Biomedical Informatics 39, 314–320. http://dx.doi.org/10.1016/j.jbi.2006.01.003

Borgman, C. L., Wallis, J. C., Enyedy, N., 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries 7 (1-2), 17–30. http://escholarship.org/uc/item/6fs4559s

Brown, S., Bruce, R., Kernohan, D., 2015. Directions for research data management in UK universities. JISC, Bristol. http://repository.jisc.ac.uk/5951/4/JR0034 RDM report 200315 v5.pdf

Bult, C. J., 2002. Data integration standards in model organisms: from genotype to phenotype in the laboratory mouse. TARGETS 1 (5), 163–168. http://www.sciencedirect.com/science/article/pii/S1477362702022158

Burnham, A., 2013. An introduction to managing research data for researchers and students. University of Leicester. http://www2.le.ac.uk/services/research-data/documents/an-introduction-to-managing-research-data

Carr, L., White, W., Miles, S., Mortimer, B., 2008. Institutional repository checklist for serving institutional management. In: Third International Conference on Open Repositories 2008, 1-4 April 2008, Southampton, United Kingdom. http://pubs.or08.ecs.soton.ac.uk/138/

Carroll, M. W., 2015. Sharing research data and intellectual property law: A primer. PLoS Biol 13 (8), e1002235+. http://dx.doi.org/10.1371/journal.pbio.1002235

Cassella, M., Calvi, L., 2010. New journal models and publishing perspectives in the evolving digital environment. IFLA Journal 36 (1), 7–15. http://www.ifla.org/files/assets/hq/publications/ifla-journal/ifla-journal-36-1 2010.pdf

Chaudiron, S., Maignant, C., Schöpfel, J., Westeel, I., 2015. Livre blanc sur les données de la recherche dans les thèses de doctorat. Université de Lille 3, Villeneuve d'Ascq. http://hal.univ-lille3.fr/hal-01192930

Costello, M. J., 2009. Motivating online publication of data. BioScience 59 (5), 418–427. https://researchspace.auckland.ac.nz/bitstream/handle/2292/7173/bio.2009.59.5.9.pdf

Cox, A., Verbaan, E., Sen, B., 2014. A spider, an octopus, or an animal just coming into existence? Designing a curriculum for librarians to support research data management. Journal of eScience Librarianship 3 (1). http://dx.doi.org/10.7191/jeslib.2014.1055

Doty, J., Kowalski, M. T., Nash, B. C., O'Riordan, S., 2015. Making student research data discoverable: A pilot program using dataverse. Journal of Librarianship and Scholarly Communication 3 (2). http://dx.doi.org/10.7710/2162-3309.1234

EU High Level Expert Group on Scientific Data, 2010. Riding the wave. how europe can gain from the rising tide of scientific data. European Union, Brussels. http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

Halipré, A., Malleret, C., Prost, H., 2015. Les données de la recherche dans les thèses en SHS de l'Université de Lille 3 (poster). In: Journées ABES, 27-28 mai 2015, Montpellier. http://www.abes.fr/Media/Fichiers/Footer/Journees-ABES/JABES 2015 Poster SCD Lille 3

Heidorn, P. B., 2008. Shedding light on the dark data in the long tail of science. Library Trends 57 (2), 280–299. https://www.ideals.illinois.edu/bitstream/handle/2142/10672/heidorn.pdf?sequence=2

Hey, T., Trefethen, A. E., 2005. Cyberinfrastructure for e-Science. Science 308 (5723), 817–821. http://dx.doi.org/10.1126/science.1110410

Hey, T., Hey, J., 2006. e-Science and its implications for the library community. Library Hi Tech 24 (4), 515-528. http://dx.doi.org/10.1108/07378830610715383

Hey, T., Tansley, S., Tolle, K. (Eds.), 2009. The fourth paradigm. Data-intensive scientific discovery. Microsoft Corporation, Redmond, WA.

Higgins, S., 2008. Draft DCC curation lifecycle model. International Journal of Digital Curation 2 (2), 82–87. http://dx.doi.org/10.2218/ijdc.v2i2.30

Juznic, P., 2010. Grey literature produced and published by universities: A case for ETDs. In: Farace, D., Schöpfel, J. (Eds.), Grey Literature in Library and Information Studies. De Gruyter Saur, pp. 39–51.

Kindling, M., 2013. Doctoral theses' research data and metadata documentation. In: ETD 2013 Hong Kong 16th International Symposium on Electronic Theses and Dissertations 25 September 2013. http://lib.hku.hk/etd2013/presentation/Maxi-ETD-20130925.pdf

Koler-Povh, T., Lisec, A., 2015. Geodetski vestnik and its path to better international recognition. Geodetski vestnik 59 (02), 289–319. http://dx.doi.org/10.15292/geodetski-vestnik.2015.02.289-319

Koler-Povh, T., Mikoš, M., Turk, G., 2014. Institutional repository as an important part of scholarly communication. Library Hi Tech 32 (3), 423–434. http://www.emeraldinsight.com/doi/full/10.1108/LHT-10-2013-0146

Koler-Povh, T., Turk, G., 2011. Instructions for theses designing and citing on UL FGG =Navodila za oblikovanje zaključnih izdelkov študijev na FGG in navajanje virov. Ljubljana: Fakulteta za gradbeništvo in geodezijo, 63 p. ISBN 978-961-6167-97-0.

Kowalczyk, S., Shankar, K., 2011. Data sharing in the sciences. Annual Review of Information Science and Technology 45 (1), 247-294. http://courses.washington.edu/geog482/resource/9_Kowalczyk_DataSharingSciences.pdf

Laney, D., 2001. 3D data management: Controlling data volume, velocity and variety. Tech. rep., Gartner META Group, Stamford CT. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

LIBER working group on E-Science / Research Data Management, 2012. Ten recommendations for libraries to get started with research data management. LIBER, The Hague. http://libereurope.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

Lynch, C., 2009. Jim Gray's fourth paradigm and the construction of the scientific record. In: Hey, T., Tansley, S., Tolle, K. (Eds.), The fourth paradigm. Data-intensive scientific discovery. Microsoft Corporation, Redmond, WA, pp. 177–183.

Lynch, C., 2014. The need for research data inventories and the vision for SHARE. Information Standards Quarterly 26 (2), 29+. http://dx.doi.org/10.3789/isqv26no2.2014.05

Malleret, C., Prost, H., 2015. Les données de la recherche dans les thèses en SHS de l'Université de Lille 3. In: Séminaire DRTD-SHS "Les données de la recherche dans les humanités numériques", 2 février 2015, Lille.

McDowell, C. S., 2007. Evaluating institutional repository deployment in American academe since early 2005. D-Lib Magazine 13 (9/10). http://dx.doi.org/10.1045/september2007-mcdowell

McMahon, B., 2010. Interactive publications and the record of science. Information Services and Use 30 (1), 1–16. http://iospress.metapress.com/content/f4th457822023783/fulltext.pdf

Morris, R. W., Bean, C. A., Farber, G. K., Gallahan, D., Jakobsson, E., Liu, Y., Lyster, P. M., Peng, G. C. Y., Roberts, F. S., Twery, M., Whitmarsh, J., Skinner, K., Mar. 2005. Digital biology: an emerging and promising discipline. TRENDS in Biotechnology 23 (3), 113–117. http://cmbi.bjmu.edu.cn/news/report/2004/biotech/24.pdf

Murray-Rust, P., 2007. The power of the electronic scientific thesis. In: ETD 2007 10th International Symposium on Electronic Theses and Dissertations, June 13-16, 2007, Uppsala, Sweden. http://epc.ub.uu.se/ETD2007/sessions/keynote-2.html

Murray-Rust, P., 2008. Open data in science. Serials Review 34 (1), 52–64. http://www.dspace.cam.ac.uk/bitstream/1810/194892/1/opendata.html

Neuroth, H., Strathmann, S., Oßwald, A., Ludwig, J. (Eds.), 2013. Digital curation of research data. Experiences of a baseline study in Germany. vwh, Glückstadt. http://www.nestor.sub.uni-goettingen.de/bestandsaufnahme/Digital Curation.pdf

Ojsteršek, M., Brezovnik, J., Kotar, M., Ferme, M., Hrovat, G., Bregant, A., Borovič, M., 2014. Establishing of a Slovenian open access infrastructure: a technical point of view. *Program* 48 (4), 394–412. http://dx.doi.org/10.1002/asi.20663

Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., Goebelbecker, H.-J., Gundlach, J., Schirmbacher, P., Dierolf, U., 2013. Making research data repositories visible: The re3data.org registry. PLoS ONE 8 (11), e78080+. http://dx.doi.org/10.1371/journal.pone.0078080

Prost, H., Malleret, C., Schöpfel, J., 2015. Hidden treasures. opening data in PhD dissertations in social sciences and humanities. Journal of Librarianship and Scholarly Communication 3 (2), eP1230+. http://dx.doi.org/10.7710/2162-3309.1230

Prost, H., Schöpfel, J., 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. http://hal.univ-lille3.fr/hal-01198379v1

Reilly, S., Schallier, W., Schrimpf, S., Smit, E., Wilkinson, M., 2011. Report on integration of data and publications. ODE Opportunities for Data Exchange, The Hague. http://www.stm-assoc.org/2011 12 5 ODE Report On Integration of Data and Publications.pdf

Savage, C. J., Vickers, A. J., 2009. Empirical study of data sharing by authors publishing in PLoS journals. PLoS ONE 4 (9), e7078+. http://dx.doi.org/10.1371/journal.pone.0007078

Savic, D., 2015. INIS: Nuclear grey literature repository. In: 8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic. https://www.techlib.cz/en/83294-conference-on-grey-literature

Schöpfel, J., Chaudiron, S., Jacquemin, B., Prost, H., Severo, M., Thiault, F., 2014. Open access to research data in electronic theses and dissertations: An overview. Library Hi Tech 32 (4), 612–627. http://www.emeraldinsight.com/doi/abs/10.1108/LHT-06-2014-0058

Schöpfel, J., Farace, D. J., 2010. Grey literature. In: Bates, M. J., Maack, M. N. (Eds.), Encyclopedia of Library and Information Sciences, Third Edition. CRC Press, London, pp. 2029–2039. http://dx.doi.org/10.1081/e-elis3-120043732

Schöpfel, J., Lipinski, T. A., 2012. Legal aspects of grey literature. The Grey Journal 8 (3), 137–153. http://archivesic.ccsd.cnrs.fr/sic 00905090/fr/

Schöpfel, J., Prost, H., 2013. Degrees of secrecy in an open Environment. The case of electronic theses and dissertations. ESSACHESS - Journal for Communication Studies 6 (2 (12)). http://www.essachess.com/index.php/jcs/article/view/214

Schöpfel, J., Prost, H., Malleret, C., 2015a. Making data in PhD dissertations reusable for research. In: 8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic. https://www.techlib.cz/en/83294-conference-on-grey-literature

Schöpfel, J., Prost, H., Piotrowski, M., Hilf, E. R., Severiens, T., Grabbe, P., 2015b. A French-German survey of electronic theses and dissertations: Access and restrictions. D-Lib Magazine 21 (3/4). http://www.dlib.org/dlib/march15/schopfel/03schopfel.html

Schultz, M., Krabbenhoeft, N., Skinner, K. (Eds.), 2014. Guidance Documents for Lifecycle Management of ETDs. Atlanta, GA. http://metaarchive.org/public/publishing/Guidance Documents for Lifecycle Management of ETDs.pdf

Sengupta, S. S., 2014. E-thesis repositories in the world: A critical analysis. Ph.D. thesis, Savitribai Phule Pune University. http://pqdtopen.proquest.com/doc/1696933497.html?FMT=ABS

Shotton, D., 2012. The five stars of online journal articles - a framework for article evaluation. D-Lib Magazine 18 (1/2). http://dx.doi.org/10.1045/january2012-shotton

Siegel, E. R., Lindberg, D. A. B., Campbell, G. P., Harless, W. G., Goodwin, C. R., 2010. Defining the next generation journal: The NLM–Elsevier interactive publications experiment. Information Services and Use 30 (1), 17-30. http://dx.doi.org/10.3233/isu-2010-0608

Simpson, P., Hey, J., 2006. Repositories for research: Southampton's evolving role in the knowledge cycle. Program: electronic library and information systems 40 (3), 224-231. http://eprints.soton.ac.uk/41240/1/ProgramAug2006simpsonrev1final2.pdf

Simukovic, E., Kindling, M., Schirmbacher, P., 2014. Unveiling research data stocks: A case of Humboldt-Universität zu Berlin. In: iConference, 4-7 March 2014, Berlin. pp. 742–748. https://www.ideals.illinois.edu/handle/2142/47259

Suber, P., 2012. Open access. MIT Press, Cambridge Mass. http://mitpress.mit.edu/books/open-access

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., Allen, E. E., Ellisman, M., Grethe, J., Wooley, J., 2011. Community cyberinfrastructure for advanced microbial ecology research and analysis: the CAMERA resource. Nucleic Acids Research 39 (suppl 1), D546-D551. http://dx.doi.org/10.1093/nar/gkq1102

Vompras, J., Schirrwagen, J., 2015. Repository workflow for interlinking research data with grey literature. In: 8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic. https://www.techlib.cz/en/83294-conference-on-grey-literature

Walker, E. P., 2011. What we can learn from ETDs: Using ProQuest dissertations & theses as a dataset. In: USETDA 2011: The Magic of ETDs…Where Creative Minds Meet. May 18-20, Orlando, Florida. https://conferences.tdl.org/USETDA/USETDA2011/paper/view/368

Wang, S., Liu, Y., 2009. TeraGrid GIScience gateway: Bridging cyberinfrastructure and GIScience. International Journal of Geographical Information Science 23 (5), 631–656. http://www.cigi.illinois.edu/publications/2009/GIScienceGateway-IJGIS-Wang-etal.p

All references are here: http://www.citeulike.org/user/Schopfel/tag/gl17