Open access to research data in electronic theses and dissertations: an overview

Joachim Schöpfel, Stéphane Chaudiron, Bernard Jacquemin, Hélène Prost, Marta Severo, Florence Thiault

OPEN ACCESS TO RESEARCH DATA IN ELECTRONIC THESES AND DISSERTATIONS: AN OVERVIEW

Joachim Schöpfel (corresponding author)

Stéphane Chaudiron

Bernard Jacquemin

Hélène Prost

Marta Severo

Florence Thiault

(GERiiCO, University of Lille 3, France)

Structured Abstract

Purpose – Print theses and dissertations have regularly been submitted together with complementary material, such as maps, tables, speech samples, photos or videos, in various formats and on different supports. In the digital environment of open repositories and open data, these research results could become a rich source of research results and datasets, for reuse and other exploitation.

Design/methodology/approach – After introducing electronic theses and dissertations (ETD) into the context of eScience, the paper investigates some aspects that impact the availability and openness of datasets and other supplemental files related to ETD (system architecture, metadata and data retrieval, legal aspects).

Findings – These items are part of the so-called "small data" of eScience, with a wide range of contents and formats. Their heterogeneity and their link to ETD need specific approaches to data curation and management, with specific metadata and identifiers and with specific services, workflows and systems. One size may not fit for all but it seems appropriate to separate text and data files. Regarding copyright and licensing, datasets must be evaluated carefully but should not be processed and disseminated under the same conditions as the related PhD theses. Some examples are presented.

Research limitations/implications – The paper concludes with recommendations for further investigation and development to foster open access to research results produced along with PhD theses.

Originality/value – Electronic theses and dissertations are an important part of the content of open repositories. Yet, their potential as a gateway to underlying research results has not really been explored so far.

Keywords: academic publishing, scientific communication, electronic theses and dissertations, institutional repositories, open access, escience, open data

Article Classification: Conceptual paper

New technology, formats and contents are challenging research communities and the information industry. Academic publishing has definitively left the Gutenberg era. Today, it has entered the world of the 4th paradigm where e-infrastructures enable "data-intensive scientific discovery" (Hey et al. 2009). Scientific information, in the words of Carlos Morais-Pires from the European Commission, has become a continuum between publication and data: "The trend is now towards a continuum of the scientific information space enabled by digital technologies"[1].

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

[1] http://www.grl2020.net/uploads/position_papers/Carlos_Morais_Pires.pdf

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). While academic publishers make usage of new technologies to enrich the content and functionalities of their online products ("article of the future", enhanced multimedia content etc.), universities have not so far really seized the opportunity of the supplementary files submitted together with electronic theses and dissertations (ETD).

The following paper explores research data related to electronic theses and dissertations as a specific part of the emerging e-infrastructure of research. After a brief overview of the context of eScience and open access, the paper will provide some elements for a better understanding of these resources and address the problems of identification and retrieval, system architecture and rights. The paper concludes with some recommendations for further research and investigation. The objective is to introduce PhD theses into the world of open, digital science and to improve access to scientific knowledge and results of publicly funded research.

1. eSCIENCE AND OPEN ACCESS TO DOCUMENTS AND RESEARCH DATA

Since the 19th century, journals have become the central vector of scientific communication (Fredriksson 2001), with most of the scientific work being disseminated via evaluated articles. The digital revolution deconstructed and fragmented 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).

Today, more and more scientific results are disseminated as datasets in digital formats, sometimes connected to, sometimes in competition with publications[2]. "Data are becoming an important end product of scholarship, complementing the traditional role of publications" (Borgman et al. 2007). The increasing mass of primary data, the need to manage the "data deluge", is among the main drivers of *eScience* or *e-/cyberinfrastructure* that consist mainly of data and not of literature or documents (Hey & Trefethen 2003, 2005). The American concept *cyberinfrastructure* is referred to as the "development and use of integrated and distributed information infrastructure that enables and accelerates the discoveries of science and engineering" and represents together with the European synonym *eScience*, "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).

Computing systems, data, information resources, networking, digitally-enabled sensors, instruments, virtual organizations, observatories, interoperable software services and tools – these are the technological components of cyberinfrastructure as defined by the US National Science Foundation (NSF 2007). Jim Gray, a researcher at Microsoft, describes this evolution as a paradigm shift. After empirical, theoretical and computerized research based on simulations of complex phenomena, science is moving to the "4th paradigm", i.e. data mining and integration of theories, simulations and experiments (Hey et al., 2009).

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). Some predict that eventually all publications, as traditional vectors for scientific communication will disappear in favour of a direct communication between machines. Availability of datasets and online access created another world, or rather, another way of seeing the world of scientific information. After the digital revolution, does this mean the end of scientific publications? "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).

Perhaps the "data deluge" will not substitute academic publishing, the written word remaining essential for humans, but it already transforms publications in different ways. 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. They become windows on research results. At least three different ways can be distinguished in which documents contribute to data production and eScience.

[2] See for instance the US survey on institutional repositories by Lynch and Lippincott (2005) where a significant percentage of institutions either already accept datasets on their server or intend to do it in the future.

Document as data: Documents, such as conventional articles, theses, reports and conference abstracts or proceedings are exploited as primary data source for text mining, automatic extraction of meaningful information, intelligence etc[3]. This means changed publishing practices. "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 (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[4]. 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).

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. For instance, this is the case when theses and dissertations are published on an institutional repository while related spreadsheets or survey results are deposited on a data repository. Text and data files are interconnected but they have two different metadata sets and they are archived on different servers[5].

One part of these data is freely available, especially on servers that meet the criteria of open access. For example, over 8,500 social sciences datasets are aggregated in the European CESSDA-portal,[6] and in the arts and humanities, digital research data can be retrieved via the DARIAH project.[7] 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[8].

Shotton (2012) discriminates between four levels in which data can be made available together with published papers. The first two levels correspond to the "data vehicle" mentioned above, where data are published together with a paper or as a part of it:

"1 Supplementary information files available

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. However, the formats for these entities are not optimised for reuse — for example, figures and tables from PLoS journal articles are only available in TIFF or PNG image format.

2 Article data downloadable in actionable form

The data contained within the figures and tables of the article, and within its supplementary information files, are available in appropriate actionable formats, for example numerical data in downloadable numerical spreadsheets or CSV files."

The other two levels described by Shotton (2012) concern the situation where research data are published independently from a paper, i.e. where the paper has the function of a gateway to underlying datasets:

"3 Underlying datasets published

The full research datasets created during the research project, from which the sub-set of data included within the published article has been selected, are published in a permanent archive or repository, with a unique resolvable identifier (e.g. a URI or a DataCite DOI), with an open access data license or a CC Zero waiver and public domain dedication, and with sufficient descriptive metadata to enable their re-interpretation and reuse.

4 Data available to peer-reviewers

In cooperation with the journal, the datasets supporting a journal article are made available to peer reviewers, to assist in evaluation of the article. This is usually achieved privately, prior to the publication of these datasets at the same time as the article."

---

[3] See for instance the research project at the UK *National Centre for Text Mining* at Manchester http://www.nactem.ac.uk/research.php

[4] See Elsevier's launch of the "article of the future" initiative in 2009 http://www.articleofthefuture.com/

[5] For instance, while all proceedings of the international conferences on grey literature are published in open access on the OpenGrey platform, some of the underlying datasets are submitted to the Dutch DANS data repository.

[6] www.cessda.org

[7] http://www.dariah.eu

[8] See Tim Berners-Lee's five degrees of openness described as "Five Stars of Linked Open Data" http://www.w3.org/DesignIssues/LinkedData.html

The last two options fit better with strategies in favour of open, digital science and free and unrestricted access to public research results, because they guarantee high quality standards and allow liberal reuse and exploitation of datasets.

2. DIGITAL PHD THESES AND RELATED RESEARCH DATA

Significant part of academic "grey literature" (Schöpfel & Farace 2010), produced and published by universities, PhD theses 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 2012). However, more and more PhD theses are available in open access through institutional repositories, 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). In 2014, the international directory OpenDOAR listed more than 1,200 institutional repositories with electronic theses and dissertations, representing roughly half of all registered open archives. The academic search engine BASE provides more than 2.7 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).

Often, PhD theses contain the results of at least three years of scientific work, 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 theses and 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…). Academic libraries cared for the recording and preservation of this material but often, especially when on a different support, it was not maintained for reuse and, especially for theses and dissertations archived and disseminated on microforms, separated from the text and excluded from local and distant access through interlending and document supply.

In the new ETD infrastructures, such material is submitted and processed 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. An informal survey among 92 US academic libraries requiring ETD shows that roughly 40% allow deposit of supplemental material[9].

The problem is that most of this material is not produced massively or in standard formats, sometimes structured, often semi- or unstructured, and therefore not part of the so-called "big data" of eScience, like those data from important equipments or projects like the Large Hadron Collider or the Human Genome Project, or shared via open and networked data repositories such as the SkyServer[10] or the Galaxy Zoo[11]. Material accompanying ETD most often is small data or "little science", largely hidden and unexploited, with a Janus-faced status, both publicly funded and personal production, and its large variety (but also its sometimes uncertain copyright protection) affects the accessibility, openness and reusability of this material.

Searching through library catalogues, bibliographic databases and repositories (DANS, DART Europe, NDLTD, SUDOC, EThOS etc.) but also in data repositories and with aggregators such as the DataCite metadata search engine produces anecdotic evidence on this material, in all disciplines but mostly non-indexed, in a great variety of forms and contents, e.g. maps, tables, speech samples and other sound recordings, photos, videos, reproductions of originals or manuscripts, spreadsheets, raw data about molecular modelling, non-encoded survey responses, questionnaires, worksheets, notebooks or collaborative activity, interview transcriptions, graphics, didactic sequences, statistics, glossaries and terminology, canvas and groundwork, graphical data or workflows models, evaluation checklists etc.

Sometimes this material is given as a data annex after the text, where the PhD thesis functions a kind of "data vehicle". An example: a 2012 PhD thesis in law from the University of Milan contains two annexes, one with questionnaire raw data, and another with tables of survey results, on more than fifty pages. The thesis is available without any restriction in the Milan institutional repository AIR[12], the datasets are recorded with a unique

[9] Survey launched in 2011 by Dorothea Salo and Sarah L. Shreeves and available at https://docs.google.com/spreadsheet/ccc?key=0AtSgIIhGWCkpdHJvOUNSZUZyRC04UXRUa0w3UmgtYWc&hl=en_US#gid=0

[10] http://www.skyserver.org/

[11] http://www.galaxyzoo.org/

[12] http://air.unimi.it/handle/2434/215991

identifier (DOI) and harvested by the DataCite content service[13] but as they are part of the text file in PDF, they can not be reused – they are open but not exploitable. Another example is a 2014 dissertation in architecture at the TU Delft, with one DOI attributed to different datasets (again, fifty pages of statistics, measures etc.) which are part of the PDF text file[14].

Sometimes the text and data files are submitted separately, and the PhD thesis plays the role of a "gateway to data". This is the case of a 2009 PhD thesis at the University of Sheffield in media and communication sciences[15] - two files with research data have been deposited on Figshare[16] with attribution of a DOI which allows for retrieval with the DataCite content service[17]. Here, the research results – thousands of labels and values in ODS spreadsheet format – are not only openly available but can be reused, reanalyzed, merged with other data. In a more systematic way, the US Inter-university Consortium for Political and Social Research (ICPSR) at Ann Arbor, Mass., started to link datasets to PhD theses from the ProQuest database (Walker 2011).

Often, ETD 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 2010 Master thesis in physics from the University of Magdeburg and DESY, with several figures and tables in the text and an annex with a complete source code – all in PDF, open, with a DOI for datasets (for the code, not for the figures and tables) but not reusable[18]. Text and data mining are necessary to identify and exploit all this information, and the dissertation is at the same time "data vehicle" and primary data.

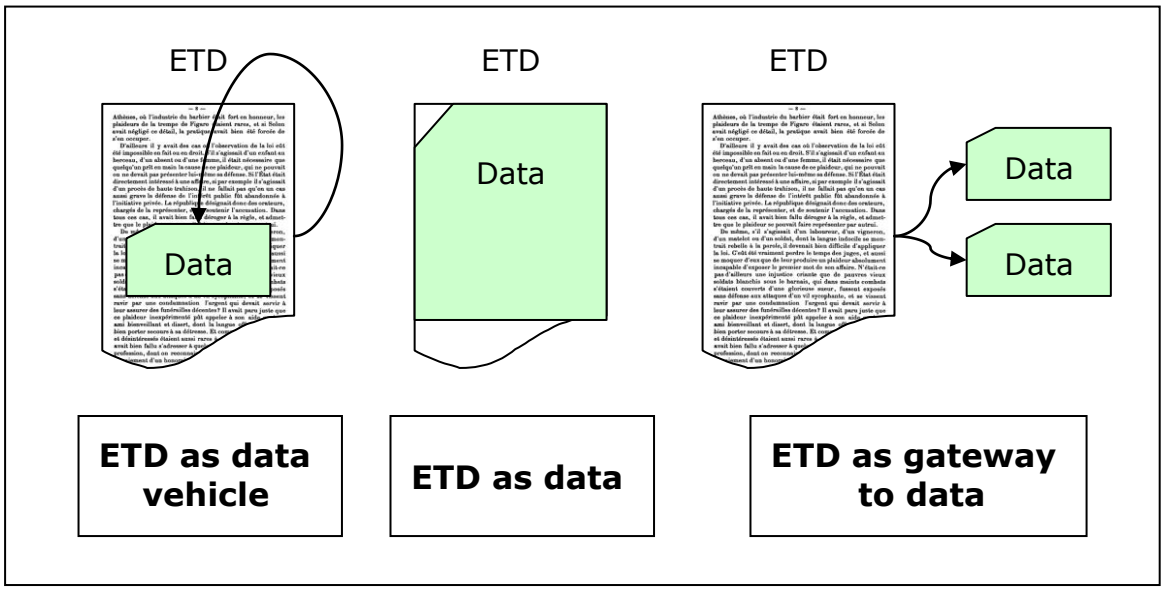

Figure 1: Electronic theses as data, data vehicles and gateway to data

Obviously, electronic theses and dissertations can contribute to eScience in each way described above: they can be exploited as primary data source, they contain (potentially) interactive content, and they are linked to datasets and research results (see Figure 1). But linking does not necessarily mean providing access. Supplementary material from ETD does not often even meet the criteria of Berners-Lee's and Shotton's first level of openness, i.e. 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 some kind of way, they belong to the dark data of eScience (Heidorn 2008). Yet, in an academic environment that claims open access not only to scientific publications but also to research results, this situation is not satisfying.

---

[13] http://data.datacite.org/10.13130/ZULETA-FERRARI-MARIANA_PHD2012-12-17
[14] http://data.datacite.org/10.7480/ABE.2014.2
[15] http://www.worldcat.org/title/reading-news-images-in-japan-visual-semiosis-in-the-context-of-television-representation-volume-1/
[16] http://figshare.com/articles/Thesis_Data/744800
[17] http://data.datacite.org/10.6084/M9.FIGSHARE.744800
[18] http://data.datacite.org/10.3204/DESY-THESIS-2011-001

3. CHALLENGES

Up to now, these items appear more or less out of scope for eScience and open data strategies. Surveys on small data (see for instance Simukovic et al. 2014) do not mention the link with theses and dissertations. Scientists report that they produce research results through observations, experiments, simulations, surveys, interviews and so on, but they report this production as completely separated from their publishing activity. There is no reliable and representative empirical evidence of the volume, distribution and characteristics of datasets and other files submitted together with ETD, except for a recent study at the University of Illinois at Urbana-Champaign that evaluated the part of ETD with some form of supplementary file since 2010 as 3%, in particular in life and physical sciences, most often with data, protocols or codes in a wide variety of formats, such as PDF or Excel (Shreeves 2013). However, the study insists on the preliminary nature of these findings and concludes that more structured description of supplementary files are needed, together with more investigation into when supplementary files are included and when they are not.

Providing access to research data related to digital PhD theses is a challenge for academic libraries and infrastructures. The following overview puts the focus on three crucial topics: which could or should be the information system that fits best? How to facilitate the retrieval of these datasets? And which are the legal conditions for their dissemination, access and reuse?

3.1. WORKFLOW AND SYSTEM ARCHITECTURE

Should ETD 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? 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. This is the first point.

The second point is that 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 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?

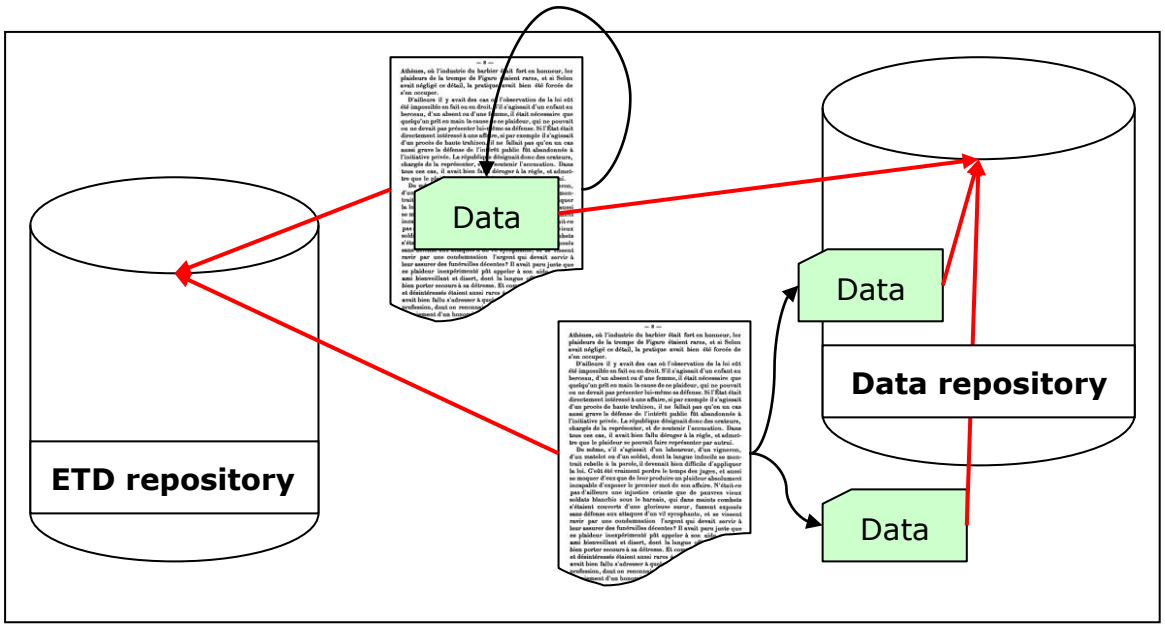

Figure 2: Storage of electronic theses as related datasets

Third point: 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 2). 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 ETD 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).

At the first EUDAT conference, Laure & Livenson (2012) presented a draft service for small datasets called "Simple Store", based on the CERN Invenio technology and designed for the "long tail of small data and the researchers/citizen scientists creating/manipulating them (with) YouTube or DropBox like functionality (and) inclusion of data presentation and discussion layer (players, feedback, etc.)". This is another model for the emerging infrastructure of repositories for small data. Yet, they must be adapted to the specificities of ETD related research results.

3.2. IDENTIFICATION AND RETRIEVAL

Appropriate metadata are crucial for the management of research results and the development of repositories and other services. "Metadata are important for helping preserve understanding and building new research environments for data sharing and repurposing" (Yang et al. 2010); without metadata, there would be no innovative web portals, discovery services etc. Neuroth et al. (2013) insist on the importance of descriptive metadata "providing systematically differentiated details, which shed light on the criteria used in selecting the object of investigation, the methods of examination, measurement and surveying, their application as well as the results of the examination." Without metadata, the research data will remain hidden, non-exploitable by other scientists, and they cannot be connected with other data, exported to other systems or stored in long-term repositories. To be fully accessible, sharable and reusable over the time, research data need curation and management throughout their whole life cycle (Neuroth et al. 2013, Schubert et al., 2013).

Formats: Because of their specificity and because of their different service environment, datasets and ETD should be described by different metadata sets. In particular, metadata formats on datasets should be adapted to the specific needs of the ETD environment. One part of both metadata sets will be shared, and these metadata should be generated whenever possible from the ETD metadata, instead of creating new data. CERIF, the European standard format of current research information systems (CRIS) can be helpful as exchange format

between data and text repositories, facilitating incidentally the integration of ETD and related datasets into research evaluation (Schöpfel et al. 2014)[19].

Standards: Standards are necessary for interconnections and interoperability. In the particular case of ETD and related data, this means standardized metadata for both. Qualified Dublin Core metadata is the minimum for ETD (see for instance Park & Richard 2011); more detailed and appropriate are standard formats like TEI, MODS (Library of Congress), the UK ETD Metadata Core Set, TEF (France ETD infrastructure STAR) and above all ETD-MS (NDLTD). Each format allows for some more or less controlled and detailed description of accompanying material. Concerning supplementary material, the exploration of standard metadata for datasets and other types of files should build on the ongoing work on metadata for datasets federated by EUDAT (de Witt 2012) or DANS[20]. Adapting those models will probably face one major problem, i.e. the great variety of these small data. Along with research data, these metadata must be documented (perhaps even standardized) and stored with the description of the technical requirements (e.g. to document or even archive hardware and software frameworks).

Identifiers: While unique identifiers for ETD already exist and are widely used, especially in national and international ETD infrastructures, the lack of permanent names for archived data and of the attribution of persistent and unique object identifiers to scientific datasets such as an URN, handle or digital object identifiers (DOI) are still a problem. The international project DataCite, member of the Research Data Alliance (RDA), already attributed more than three millions DOI to research data[21]. At present, datasets linked to PhD theses represent less than 1% of the DataCite content service. The International Standard Randomised Controlled Trial Number Register (ISRCTN) recently decided to introduce DOI to all records in the Current Controlled Trials database and to link datasets to publications. So far, no major institution or network appears to have taken the same kind of decision for digital PhD theses.

Linking ETD to data: Datasets should be connected to the text via the ETD metadata, not only by attribution of the unique ETD identifier or via an URL address but by means of their own persistent identifier. ETD should mention the existence of related datasets as a specific element of their metadata set, as detailed and controlled as possible. Linking online ETD to their underlying datasets should also be considered in the emerging context of semantic publishing (Rinaldi 2010) and rich Internet publications (Breure et al. 2011).

Formats, standards, identifiers and appropriate linking are necessary conditions for data retrieval but they do not guarantee accessibility and reusability. Non-adapted or missing discovery services and functionalities can be bottlenecks on the way to open access to ETD and supplementary data files. Data integration and handling need specific service and workflow management. "Creating a production grid environment poses several significant technological problems related to security, accounting, information provisioning, resource brokering, and data management" (Dooley et al 2006). If hosting institutions or service providers want to increase usage and impact of open repositories, they should develop a user-centred process of service development prioritization (Halbert 2007). According to Morris et al. (2005), "biomedical computing will permit scientists to extract biologically meaningful information from datasets of ever-increasing size, heterogeneity and complexity", a prognosis that can be extrapolated to eScience in general and ETD in particular but is conditioned by appropriate metadata, retrieval tools and other data-related services.

When asked to indicate services they would like to have for their research data, scientists rank first secured and backed-up storage, followed by advice and guidance on legal issues (e.g. access restrictions, sensible data, licensing) and on technical issues (e.g. metadata, standards, long-term preservation) (Simukovic et al. 2014). Options for copyright clearance, such as dissemination under a Creative Commons or an Open License can be part of these added value services upstream of the repository. In order to exploit and explore datasets, scientists will need "the assistance of specialized search engines and powerful data mining tools" (Hey & Hey 2006).

Open and institutional repositories develop a large spectrum of added value services such as training events, guidelines and assistance, mailing lists, different entry points and separate indexes, several browsing and search options, customisation, reference management such as export of bibliographic listings in different formats, social media tools, multi-dimensional data processing, virtual organizations, location based service, usage statistics, videos or print on demand (see for instance Bester 2010). Some of them may be relevant for data repositories.

---

[19] See the proposal of data-related model extensions by the CERIF task group at http://cerifsupport.org/2013/04/02/data-in-cerif/

[20] *Data Archiving and Networked Services* run by the Dutch Royal Academy of Arts and Sciences http://www.dans.knaw.nl/

[21] https://www.datacite.org/

In a general way, the specifications for services and functionalities must allow the linking of documents and datasets, while facilitating the independent use and reuse of each item and their exploitation together with other research data ("cross-domain data-mashup"), including harvesting by other service providers. This means that the data deposited in the form of supplementary files should be reusable independently of the related thesis (yet clearly linked to it). Another related issue may be the service, support and incentives for authors to publish their data together with their thesis (see Costello 2009).

All these services and functionalities are designed to increase the availability, searchability and retrieval of research results. Some of these developments are possible even with a "very modest programming staff available to deploy on ad hoc projects" (Halbert 2007) while others ask for more investment and resources. On a more general level, the repository services for the dissemination of ETD should be flexible, with a capacity for rapid adaptation; the software should be user-friendly and reliable, perhaps also open to other service providers and integrated in different service environments, such as extended (distant) learning.

3.3. LICENSING AND REUSE RIGHTS

Recently, a librarian asked the following question on the international ETD listserv: "Suppose that at your publicly-funded university, a student completes a full-length documentary film as part of a PhD in Gender Relations and Social Justice. (…) Does your university allow the student to refuse to put the film in your library, and to retain full and exclusive rights to its distribution?" Applying copyright to PhD theses is already rather complicated (Schöpfel & Lipinski 2013). The legal environment of data and other supplementary materials is not the same as for ETD. 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."[22]

The legal regime for datasets should not be the same as for the dissemination of theses and dissertations. Murray-Rust (2008) suggested for open data the following categories:

- scientific data deemed to belong to the commons (e.g. the human genome)
- infrastructural data essential for scientific endeavour (e.g. in Geographic information systems)
- data published in scientific articles which are factual and therefore not copyrightable
- data as opposed to software and therefore not covered by Open Source licenses and so potentially capable of being misappropriated.
- maps and other artifacts required for communal infrastructure.

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.

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.

[22] http://europa.eu/rapid/press-release_IP-11-1524_en.htm

4. 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 eScience, 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 theses. Today, new investment is needed for the curation of research data produced and deposited with PhD theses. The development of ETD infrastructures, open repositories and eScience makes it possible to find an appropriate solution for the management and reuse of small data produced along with PhD theses. Our paper provides an overview on characteristics, bottlenecks and perspectives of these resources.

Our first observation is that in spite of some empirical evidence, precise knowledge about of supplementary files submitted together with PhD theses is lacking. We know that this material exists, either as a part of the thesis (annex with tables of results etc.) or deposited as data files with different formats and independently from the text of the dissertation. But this evidence does not allow a deeper understanding of characteristics and requirements due to formats, scientific disciplines and specific research fields, and more investigation on volume and variety is needed for the development of appropriate systems and services.

PhD theses 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, PhD theses should be valorised as "gateways to data" which implies incentives for the deposit of related datasets and other supplementary files and specific 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 small data repositories should be, whenever possible, integrated in CRIS environments.

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 (Suber 2012). 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 (Simukovic et al. 2014) 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 local policy in favour of open access and open data.

So, how to do the job in the best and most efficient way, how to make the first step in this field in order to improve access to scientific knowledge and the transition to open, digital science? Five suggestions may be helpful. First, evaluate the specific needs and questions of the different research communities on the local level; they produce data and they ask for solutions especially for the storage and preservation. Second, raise awareness through debate and communication about the public and scientific interest of these research data, again in particular about those aspects that mostly concern the scientists (conservation, storage…). Three, learn from successful models and initiatives, exchange with experts and stakeholders, invite them for conferences and presentations but keep in mind the need for local solutions. Four, prepare a research data management plan appropriate to the scientific community and adapted to disciplinary particularities, make lobbying in the institution but again, keep in mind that it should offer a solution for the researchers' needs, not (only) a mandatory policy implying more workload for scientists and staff. And last not least, five, facilitate the free and open access to PhD theses as a fundamental condition for a related data management plan, accompany, inform and advice young scientists during the preparation of their dissertation. This is not only a task for academic libraries and graduate schools but for all senior scientists concerned and interested in open, digital science.

5. BIBLIOGRAPHY

Bester E. (2010), "Les services pour les archives ouvertes de la référence à l'expertise", *Documentaliste - Sciences de l'Information*, Vol. 47 No. 4, pp.4-15.

Blake J.A. and Bult C.J. (2006), "Beyond the data deluge: Data integration and bio-ontologies", *Journal of Biomedical Informatics*, Vol. 39 No. 3, pp.14-320.

Borgman C.L., Wallis J.C. and Enyedy N. (2007), "Little science confronts the data deluge: habitat ecology, embedded sensor networks, and digital libraries", *International Journal on Digital Libraries*, Vol. 7 No.1-2, pp.17-30.

Breure L., Voorbij H. and Hoogerwerf M. (2011), "Rich Internet Publications: "Show What You Tell"", *Journal of Digital Information*, Vol. 12 No. 1.

Bult C.J. (2002), "Data integration standards in model organisms: from genotype to phenotype in the laboratory mouse", *TARGETS*, Vol. 1 No. 5, pp.163-168.

Carr L., White W., Miles S. and Mortimer B. (2008), "Institutional Repository Checklist for Serving Institutional Management", In *Proceedings of the 3rd International Conference on Open Repositories 2008, Southampton, United Kingdom 1-4 April 2008.*

Cassella M. and Calvi L. (2010), "New journal models and publishing perspectives in the evolving digital environment", *IFLA Journal*, Vol. 36 No. 1, pp.7-15.

Costello M.J. (2009), "Motivating Online Publication of Data", *BioScience*, Vol. 59 No.5, pp.418-427.

Dooley R., Milfeld K., Guiang C., Pamidighantam S. and Allen G. (2006), "From Proposal to Production: Lessons Learned Developing the Computational Chemistry Grid Cyberinfrastructure", *Journal of Grid Computing*, Vol. 4 No. 2, pp.195-208.

Fredriksson E.H. (2001), *A century of science publishing*. IOS Press, Amsterdam.

Halbert M. (2007), "Integrating ETD Services into Campus Institutional Repository Infrastructures Using Fedora", In *Proceedings of ETD 2007 10th International Symposium on Electronic Theses and Dissertations, Uppsala, Sweden, June 13-16, 2007.*

Heidorn P.B. (2008), "Shedding Light on the Dark Data in the Long Tail of Science", *Library Trends*, Vol. 57 No. 2, pp.280-299.

Hey T. and Trefethen A.E. (2003), "The Data Deluge: An e-Science Perspective", In *Grid ComputingMaking the Global Infrastructure a Reality*, pp. 809-824. Wiley, Chichester.

Hey T. and Trefethen A.E. (2005), "Cyberinfrastructure for e-Science", *Science*, Vol. 308 No. 5723, pp.817-821.

Hey T. and Hey J. (2006), "e-Science and its implications for the library community", *Library Hi Tech*, Vol. 24 No. 4, pp.515-528.

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

Juznic P. (2010), "Grey Literature produced and published by Universities: A Case for ETDs", In Farace D.J. and Schöpfel J. (2010), *Grey Literature in Library and Information Studies*, pp.39-51, De Gruyter Saur, München.

Laure E. and Livenson I. (2012), "Simple Store. An Overview of a Potential New EUDAT Service", In *Proceedings of EUDAT 1st Conference, Barcelona, October 22-24, 2012*.

Lynch, C. and Lippincott, J.K. (2005), "Institutional repository deployment in the United States as of early 2005", *D-Lib Magazine*, Vol. 11 No. 9.

Lynch C. (2009), "Jim Gray's Fourth Paradigm and the Construction of the Scientific Record", In Hey T., Tansley S. and Tolle K. (2009), *The fourth paradigm. Data-intensive scientific discovery*. Microsoft Corporation, Redmond, WA, pp.177-183.

Mauch J.E. and Park N. (2003), *Guide to the successful thesis and dissertation: a handbook for students and faculty*. Marcel Dekker, New York.

McDowell C.S. (2007), "Evaluating Institutional Repository Deployment in American Academe Since Early 2005", *D-Lib Magazine*, Vol. 13 No. 9/10.

McMahon B. (2010), "Interactive publications and the record of science", *Information Services and Use*, Vol. 30 No.1, pp.1-16.

Morris R.W., Bean C.A., Farber G.K., Gallahan D., Jakobsson E., Lui Y., Peng G.C., Roberts F.S. Twery, M., Whitmarsh J. and Skinner K. (2005), "Digital biology: an emerging and promising discipline", *Trends in Biotechnology*, Vol.23 No. 3, pp.113-117.

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.*

Murray-Rust P. (2008), "Open Data in Science", *Serials Review*, Vol. 34 No.1, pp.52-64.

Neuroth H., Strathmann S., Osswald A., Ludwig J. (2013), "*Digital curation of research data. Experiences of a baseline study in Germany*", Verlag Werner Hülsbusch, Glückstadt.

National Science Foundation Cyberinfrastructure Council (2007), *Cyberinfrastructure Vision for 21st Century Discovery*, Report NSF 07-28, National Science Foundation.

Park E.G. and Richard M. (2011), "Metadata assessment in e-theses and dissertations of Canadian institutional repositories". *The Electronic Library*, Vol. 29 No. 3, pp.394-407.

Rinaldi A. (2010), "For I dipped into the future", *EMBO reports*, Vol. 11 No. 5, pp.345-349.

Savage C. J. and Vickers A.J. (2009), "Empirical Study of Data Sharing by Authors Publishing in PLoS Journals", *PLoS ONE*, Vol. 4 No. 9, pp. e7078+.

Schöpfel J. and Farace D.J. (2010), "Grey Literature", In Bates M.J. and Maack M.N. (2010), *Encyclopedia of Library and Information Sciences*, Third Edition, pp.2029-2039, CRC Press, London.

Schöpfel J. and Lipinski T.A. (2012), "Legal Aspects of Grey Literature", *The Grey Journal*, Vol. 8 No. 3, pp.137-153.

Schöpfel J. and Prost H. (2013), "Back to Grey. Disclosure and Concealment of Electronic Theses and Dissertations", In *Proceedings* of *GL15, the Fifteenth International Conference on Grey Literature. The Grey Audit: A Field Assessment in Grey Literature. CVTI SR, 2-3 December 2013, Bratislava*, pp.110-118, Amsterdam. TextRelease.

Schöpfel J., Zendulkova D. and Fatemi O. (2014), "Electronic Theses and Dissertations in Current Research Information Systems", In *Proceedings of the 12th International Conference of Current Research Information Systems (CRIS 2014), Rome, 13-15 May 2014.*

Schubert C., Shorish Y., Frankel P. and Giles K. (2013), "The evolution of research data: strategies for curation and data management", *Library Hi Tech News*, Vol. 30 No. 6, pp.1-6.

Shotton D. (2012), "The Five Stars of Online Journal Articles - a Framework for Article Evaluation", *D-Lib Magazine*, Vol. 18 No. 1/2.

Shreeves S.L. (2013), "Supplementary Files in Electronic Theses and Dissertations: Implications for Policy and Practice", In *Proceedings of the 8th International Digital Curation Conference, IDCC 2013, Amsterdam, Netherlands, 14-16 January 2013.*

Simpson P. and Hey J. (2006), "Repositories for research: Southampton's evolving role in the knowledge cycle", *Program: electronic library and information systems*, Vol. 40 No. 3, pp.224-231.

Simukovic E., Kindling M. and Schirmbacher P. (2014), "Unveiling Research Data Stocks: A Case of Humboldt-Universität zu Berlin", *Proceedings of iConference, 4-7 March 2014, Berlin*, pp.742-748.

Suber P. (2012), *Open access*. MIT Press, Cambridge Mass.

Sun S., Chen J., Li W., Altintas I., Lin A., Peltier S., Stocks K. Allen E.E., Ellisman M., Grethe J. and Wooley J. (2011), "Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource", *Nucleic Acids Research*, Vol. 39 Suppl. 1, pp.D546-D551.

Walker E.P. (2011), "What We Can Learn from ETD: Using ProQuest Dissertations & Theses as a Dataset", In *Proceedings of USETDA 2011: The Magic of ETD…Where Creative Minds Meet. May 18-20, Orlando, Florida.*

Wang S. and Liu Y. (2009), "TeraGrid GIScience Gateway: Bridging cyberinfrastructure and GIScience", *International Journal of Geographical Information Science*, Vol. 23 No. 5, pp.631-656.

Witt (de) S. (2012), "Metadata and EUDAT', In *Proceedings of EUDAT 1st Conference, October 22-24, 2012, Barcelona.*

All web sites accessed in April and May 2014.