D4Humanities. Deposit of Dissertation Data in Social Sciences and Humanities – A Project in Digital Humanities

Joachim Schöpfel, Hélène Prost

D4Humanities

Deposit of Dissertation Data in Social Sciences and Humanities – A Project in Digital Humanities

Authors

Joachim Schöpfel (corresponding author), GERiiCO Laboratory, University of Lille 3, France

Hélène Prost, associated member of GERiiCO Laboratory, CNRS, France

Abstract

Following our work on research data and electronic theses and dissertations since 2013, we are conducting a new research project between 2017 and 2018 called *D4Humanities* with three objectives – to develop the research data management and stewardship on our campus, to gain better insight into the nature of research data in social sciences and humanities and to produce empirical evidence on the development of dissertations. In particular, the project contains three components:

1. Qualitative survey on behaviours and knowledge in the field of research data with 50 scientists from the University of Lille Social Sciences and Humanities Department, with a special focus on the FAIR guiding principles of scientific data management and stewardship.
2. The creation of a workflow for the submission of research data related to PhD dissertations (deposit, preservation and dissemination of data via the NAKALA service Huma-Num)
3. Two conceptual studies on the definition and typology of research data in SSH and on the development of dissertations in the environment of e-Science and Open Science (content, format, structure, requirements).

In the following we present some preliminary results, in particular from the survey and from the conceptual studies, in order to enhance the understanding of research data in SSH and of the development of dissertations.

Acknowledgment: The project receives funding from the European Institute of Social Sciences and Humanities (MESHS Lille) and from the Regional Council (Conseil Régional Hauts-de-France). We would like to thank the D4Humanities project team for their contribution to the research underlying this paper, in particular Cécile Malleret, Eric Kergosien and Leslie Hyacinthe.

Keywords

PhD dissertations, research data, digital humanities, open access, open science, social sciences and humanities

Introduction

For more than ten years, one part of our professional and scientific work has been focused on PhD dissertations as one major document type of academic grey literature. We started with research on their production and findability (Paillassard et al. 2005) and then moved on to questions related to their accessibility, especially in the environment of electronic theses and dissertations (ETD), open access (OA) and institutional repositories (Schöpfel 2013, Schöpfel & Prost 2013, Schöpfel et al. 2015c). Three years ago, we began to study the research data produced by PhD students and submitted as complementary material together with the dissertation (Schöpfel et al. 2014), trying to establish the link between grey literature and e-Science in the field of ETDs. Our first questions were (are) operational: what could and should be done with this material, how can it be stored and preserved for a longer time, what are the conditions for sharing, publishing and reuse? However, these practical questions always included conceptual elements, about the definition and typology of data, about their identification and description, about their relationship with dissertations, and about the development of dissertations themselves and their potential for reuse with content mining tools. Because of the complexity of the field, we limited our research to the disciplines of social sciences and humanities (SSH).

In 2017 we launched a two-year project called *D4Humanities*[1] in order to transform our research work into operational service development on the campus and to enhance our knowledge of data and dissertation. The basic question is how to enable the exploration of research data in social sciences and humanities (textual or oral corpus, raw data, images...) with digital technologies (text and data mining, mapping, visualization ...) to convey a new meaning? The project *D4Humanities* is part of the Digital Humanities and a continuation of the recent research of the GERIICO laboratory and its partners at the University of Lille Humanities and Social Sciences (academic library, SSH graduate school, digitization centre ANRT...) with the objective of accelerating the research data management project in particular for PhD students and young researchers, and of fostering the preparation of an international research project.

We started our project in March 2017, and it will continue until fall 2018. So what we will do here is deliver some preliminary results on data behaviour and data management, including the development of a workflow for ETD related datasets, and first conceptual work on data and dissertations. This will be followed by an invitation to join our research consortium.

Data literacy (survey)

In 2015, we conducted a campus-wide survey at the University of Lille on research data management in social sciences and humanities. The survey received 270 responses, equivalent to 15% of all scientists, scholars, PhD students and administrative and technical staff; all disciplines were represented. The responses showed a wide variety of data, practice and usage; some differences seem related to job status and disciplines. Generally, 20-25% of the sample can be considered as pioneers in data management and sharing, and 25-30% are motivated; only 5-10% appear reluctant to make their data available (Schöpfel & Prost 2016).

On the basis of the results of this first survey, we prepared a small qualitative survey with academic "volunteers" on the Lille SSH campus, among researchers and PhD students from various disciplines. We wanted to gain more insight in personal research data management behaviour and data literacy, in particular those contributing to the compliance with the FAIR principles for data management (Wilkinson et al. 2016). The investigation is not over; for the moment, we have conducted 27 interviews with researchers from history, archaeology, literature and language studies, psychology and information sciences. First results and comments:

Interest and motivation: finding volunteers on the campus was not easy this time; obviously, for many colleagues RDM is not a "hot topic" to spend one hour or more in a semi-directive interview on data practice and literacy. At least, it does not appear as priority or relevant.

Funding agencies: one half of the volunteering respondents (14) has conducted or participated in one or more research projects funded by the European Commission (H2020 program) and/or the

[1] Deposit of Dissertation Data in Social Sciences and Humanities. A Project in Digital Humanities https://d4h.meshs.fr/

French National Research Agency (ANR programs). But only 10 have knowledge of requirements (such as of the H2020 program), guidelines or recommendations for RDM.

Privacy: 13 respondents use or produce personal data as defined by the French CNIL commission, or confidential data. 6 submitted a research protocol to the university's ethics committee.

Standards, description: 8 participants reported assigning codes to their data, 9 people have already drafted a data management plan, and 5 participants follow standards for describing their data.

Dissemination and sharing: data collection, analysis and storage are often carried out by the researcher him/herself or together with the research team. 16 participants agree to share their data with others, which means above all with other colleagues from the project team. 10 participants have already submitted their data to an online server, 2 others intend to do so; only one refuses for security reasons.

Need for advice: generally, the respondents need advice on querying databases, formatting and naming data; they seek advice on licensing and legal protection of sensitive data; they want to know more about the services offered by the deposit platforms. So far, they have been seeking advice on RDM not at the library but with people from the IT department (system security, storage) and from the ethics committee.

Need for data services: the services requested by the researchers relate mainly to the different aspects related to data storage: to know what data to store, under which formats, on which server, with which guarantees of duration and security. They want to encourage exchanges between researchers and information professionals.

So far, we have observed very large differences between disciplines and research domains, but also between research methods and tools in the same field. Some scientists have a long experience with RDM and apply standard and transparent data procedure, even if they don not always call it RDM. This data literacy can mainly be explained by legal issues (privacy laws, especially in psychology, education, sociology, and projects in public health) or ethics rules, less (up to now) by requirements from funding agencies. However, application of standards in RDM remains exceptional, such as data publishing and sharing. We did not encounter significant reluctance or even opposition to RDM and data sharing, but rather ignorance or lack of interest.

Data workflow

Similar to other ETD projects[2] we are developing a local workflow for the deposit of research data by PhD students. The main characteristics of this workflow are:

- Data and dissertations are submitted on different servers,
- The local deposit is interconnected with existing infrastructures, in particular with the French SSH data platform NAKALA,
- Data and dissertations are stored and preserved on various platforms but linked via their metadata and identifiers.

Figure 1 shows the workflow and the separation of data and dissertations from the beginning on (deposit). The guiding principle was to provide an interface (with technical assistance) on our campus for the deposit of research data on the NAKALA platform of the national infrastructure for SSH communities. For a detailed description, see Schöpfel et al. (2017b).

[2] For instance, the ETDplus project funded by Educopia https://educopia.org/research/grants/etdplus and the workflow at the University of Bielefeld, see Vompras & Schirrwagen (2015)

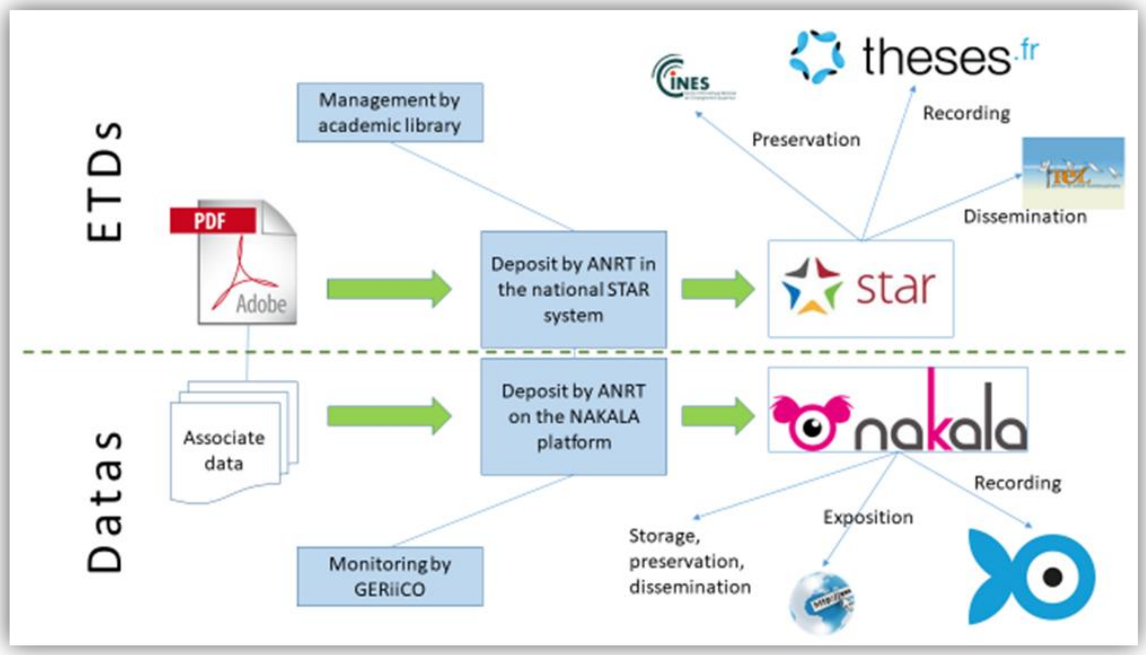
ETDs
Datas
PDF
Adobe
Management by academic library
Deposit by ANRT in the national STAR system
star
CINES
theses.fr
Preservation
Recording
Dissemination
Associate data
Deposit by ANRT on the NAKALA platform
nakala
Monitoring by GERiiCO
Storage, preservation, dissemination
Exposition
Recording

Figure 1: Local ETD/data workflow

Our intention is to offer young scientists a “default solution”, complementary to existing disciplinary data repositories, accompanied by technical assistance and a PhD training program for research data management delivered by our Graduate School.

The preparation and development of the workflow raised several issues, some of them familiar to the grey community:

- Granularity: what exactly should be defined as a dataset for deposit? We have discussed this question in two communications (Schöpfel et al. 2016, 2017a). There are no clear rules or guidelines. The pragmatic solution is to accept datasets on a granularity level which makes sense for understanding (validation) and reuse, and to allow deposit of dataset collections with a hierarchical structure.
- Data structure and description: how are data to be described and structured? Our option is to apply the Metadata Encoding & Transmission Standard of the Library of Congress.
- Identifier: which unique identifier should be used for the datasets? Even if France is part of the DataCite consortium for the assignment of DOIs, we opt for the handle system which is applied by the Huma-Num infrastructure but remain open for future adoption of the DOI.
- Legal aspects: we anticipate legal issues like copyright, third party rights, privacy etc. Our approach is twofold: we provide legal advice as part of the library's data service, and we ask the students to provide a declaration (template) that they have the permission to upload the datasets on NAKALA.
- Quality: the question was raised about the quality of datasets. Should all datasets provided by PhD students be accepted? Should we set up a kind of validation procedure? If so, which criteria should be applied? Who should evaluate? For the moment, we will not filter

submitted data files otherwise than by formal criteria (size, format...), similar to other projects and data repositories. But the question remains open.

The tests of the new workflow started end of September. The workflow will be operational in 2018.

Data definition

But what exactly are data and datasets? The issue was raised during the preparation of the data workflow. Therefore, we carry out a conceptual analysis of the meaning and content of the term of research data as a vital complement to the workflow development and survey. The first results were presented during a workshop at the University of Toulouse in May 2017 (Schöpfel et al. 2017a). Figure 2 resumes the main characteristics of our approach which is based on a synthesis of recent French and international reviews and definitions.

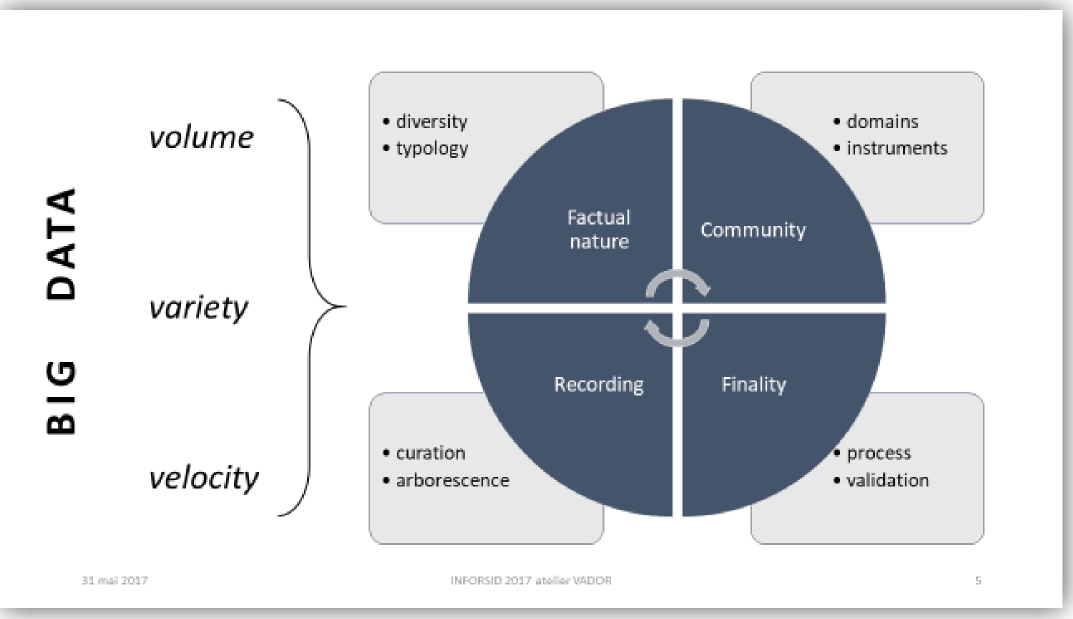

Figure 2: Elements of a data definition

We can identify five key elements of research data:

1. Link to the concept of big data: even if one part of research data is considered as small or "smart" data, the link with the "3 V's" of the big data is always present[3], in particular the diversity of data, their large number and size and the continuous stream of data input and output.
2. Factual nature: definitions of research data often insist on their factual nature, at least implicit, as primary material in need for processing, analysis and interpretation. This often implies a more or less detailed typology of data.

[3] See the consensual definition on big data by De Mauro et al. 2016: "Big Data is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".

3. The link to community: research data are embedded in disciplinary and institutional context, are specific to large instruments, research infrastructures and scientific domains.
4. Finality: research data are also embedded in the research process (cycle), are dynamic, with different functions and requirements. The most basic distinction is between input and output, primary and secondary data, data as resources and data as results of scientific work. Among the various functions, the most important (in a mainly STI and library perspective) are the validation of results and hypotheses (replication) and their preservation along with publications.
5. Recording: the need for recording (curation, preservation) is the last key element of research data definitions. Part of the research data management, data curation raises issues like granularity, identification and data arborescence (hierarchical structure of data and datasets).

Actually, we complete this synthesis with an assessment of the re3data[4] typology, their distribution and definition especially in the field of social sciences and humanities. Special attention is paid to the content of large data types (raw data, images) and the "other" categories of the more than 500 repositories in SSH.[5]

Data impact: evolution of PhD dissertations

The fourth and last work package of the D4Humanities project is intimately associated with the research and debates in the grey community. Our question is: how does the new environment of research data management and text and data mining impact the characteristics and requirements of ETDs? The discussion is open whether or not PhD dissertations should still be considered as grey literature in the digital age and how (Schöpfel & Rasuli 2017); but it seems obvious that the potential of text and data mining and the availability of datasets related to dissertations will have (or already have) substantial effects on the writing, content, format, length and submission and processing of dissertations, perhaps even on their legal status and licensing.

In the past years, we tried to assess which kind of data are related to PhD dissertations, especially in social sciences and humanities, how they are linked to the dissertation and how they should be curated (Prost et al. 2015, Schöpfel et al. 2015a, b); furthermore, we started to re-examine the meaning of dissertations in the light of text and data mining, considering dissertations as data (Schöpfel et al.2016). Content mining tends to make the borders between text and data increasingly blurred, even insignificant, and revives the discussion on the distinction between publications (documents) and data.

The D4Humanities project contributes to this research field from a special perspective, i.e. the guidelines, prescriptions and laws ruling the writing and submission of digital PhD dissertations. In 2018, the project team will conduct a landscape study together with academic and corporate partners, including a state of the art on recent research and papers on dissertations and data and a small-scale survey on the development of PhD prescriptions.

Perspectives

This last work package is just a beginning. In fact, its objective is threefold:

[4] The international registry of research data repositories, available at http://www.re3data.org/

[5] See Kindling et al. (2017) for some general elements of these repositories. Our own results are stored on a wiki and available on request at http://d4hdata.pbworks.com

1. An overview on ongoing research in order to define questions and hypotheses for further research.
2. The setting up of a scientific consortium around a core project team (GERiiCO laboratory at Lille and Institute of Scientific Networking at Oldenburg).
3. And third, the preparation of an international research project on new forms of PhD dissertations, with European (H2020) or French-German funding (ANR/DFG). For the time being, the project's code name is *xDiss*, for "Special Dissertations".

Therefore our conclusion is an appeal to the members of the grey community: if you are interested, contact us and join our consortium.

References

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.

Jacquemin, B., Prost, H., Schöpfel, J., Severo, M., Thiault, F., 2013. Ouvrir les données de la recherche pour la veille scientifique. Le cas des thèses électroniques. In: *VSST'2013*, Nancy, 23-25 octobre 2013.

Kindling, M., et al., 2017. The landscape of research data repositories in 2015: A re3data analysis. *D-Lib Magazine* 23 (3/4).

De Mauro, A., Greco, M., Grimaldi, M., Apr. 2016. A formal definition of big data based on its essential features. *Library Review* 65 (3), 122-135.

Paillassard, P., Schöpfel, J., Stock, C., 2005. How to get a French doctoral thesis, especially when you aren't French. *Publishing Research Quaterly* 21 (1), 73-93.

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+.

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.

Schöpfel, J., 2013. Adding value to electronic theses and dissertations in institutional repositories. *D-Lib Magazine* 19 (3/4).

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)).

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.

Schöpfel, J., Juznic, P., Prost, H., Malleret, C., Cesarek, A., Koler-Povh, T., 2015a. Dissertations and data (keynote address). In: *GL17 International Conference on Grey Literature*, 1-2 December 2015, Amsterdam.

Schöpfel, J., Prost, H., Malleret, C., 2015b. 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.

Schöpfel, J., Prost, H., Piotrowski, M., Hilf, E. R., Severiens, T., Grabbe, P., 2015c. A French-German survey of electronic theses and dissertations: Access and restrictions. *D-Lib Magazine* 21 (3/4).

Schöpfel, J., Kergosien, E., Chaudiron, S., Jacquemin, B., 2016. Dissertations as data. In: *ETD2016*, Lille 11-13 July 2016.

Schöpfel, J., Prost, H., 2016. Research data management in social sciences and humanities: A survey at the University of Lille 3 (France). *LIBREAS. Library Ideas* 29, 98-112.

Schöpfel, J., Prost, H., Rebouillat, V., 2016. Research data in current research information systems. In: *CRIS 2016*, St Andrews, 8-11 June 2016.

Schöpfel, J., Kergosien, E., Prost, H., 2017a. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. In: *Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017*, 31 mai 2017 Toulouse (France).

Schöpfel, J., Prost, H., Malleret, C., 2017b. Research and development in the field of research data and dissertations. The D4Humanities project at the University of Lille (France). In: *10th Conference on Grey Literature and Repositories*, National Library of Technology (NTK), 19 October 2017, Prague, Czech Republic.

Schöpfel, J., Rasuli, B., 2017. Are electronic theses and dissertations (still) grey literature in a digital age? a FAIR debate. *The Electronic Library* 35 (4).

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.

Wilkinson, M. D., et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, sdata201618+.