Research and Development in the Field of Research Data and Dissertations

Joachim Schöpfel, Hélène Prost, Cécile Malleret

10th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2017 [cit. 2017-10-15]. Available from: **http://nrgl.techlib.cz/conference/conference-proceedings/**. ISSN 2336-5021.

RESEARCH AND DEVELOPMENT IN THE FIELD OF RESEARCH DATA AND DISSERTATIONS

Schöpfel, Joachim

joachim.schopfel@univ-lille3.fr

GERiiCO laboratory, University of Lille SHS, France

Prost, Hélène

helene.prost@inist.fr

GERiiCO laboratory, INIST (CNRS), France

Malleret, Cécile

cecile.malleret@univ-lille3.fr

Academic library, University of Lille SHS, France

Abstract

The paper presents the research project D4Humanities conducted by the GERIICO laboratory at the University of Lille in the field of research data management (RDM). In particular, it describes the development of a local workflow for the submission of research data related to PhD dissertations and the connection to the national RDM infrastructure

Huma-Num (deposit, preservation and dissemination of research data via the NAKALA service), along with the RDM training program for PhD students provided by the Graduate School in Social Sciences and Humanities at the University of Lille.

Keywords

Research data, electronic theses and dissertations, data management, data sharing, data literacy, PhD training program, social sciences and humanities

Introduction

Research data management (RDM) has been described as a "wicked problem" without easy answers, perhaps even insoluble, at least temporarily (Awre et al., 2015). However, each research performing institution must define its own RDM strategy in order to provide optimal work conditions for its faculty and scientists. So even if (or just because) there may be no "best model" for RDM, institutions can learn from each other's experiences and successful initiatives.

Two years ago, at the 2015 Conference on Grey Literature and Repositories in Prague, we presented empirical results on research data in electronic theses and dissertations (ETDs) and on RDM behaviours and needs of scientists and PhD students on the social sciences and humanities (SSH) campus of the University of Lille (Schöpfel et al. 2015). The basic assumption of our research was (and still is) the observation that research results produced by PhD students can contribute to data-intensive scientific discovery (Schöpfel et al. 2014). They are "hidden treasures" (Prost et al. 2015); however, a few issues must be addressed in order to make them visible, available and, moreover, reusable for scientific research.

In 2015, we published an institutional approach to RDM in the field of PhD dissertations as a White Paper (Chaudiron et al. 2015), promoting five leading principles for the development of campus-based research data support services (see Schöpfel et al. 2015, figure 5). After validation by our research department, we started to implement this approach on the campus, together with the SSH graduate school, the GERiiCO research laboratory[1], the academic library and the ANRT service (National Centre for the Reproduction of PhD Dissertations). The implementation will take three years (2015-2018); one part of the implementation was integrated in a research project called *D4Humanities*[2] and received funding from the Regional Council (Conseil Régional Hauts-de-France) and the European Institute of Social Sciences and Humanities in Lille (MESHS).

Where are we now? What have we learned? Our paper will present the actual advancement of the implementation process, in the particular context of our institution and country, and it will make some statements and recommendations for similar initiatives.

[1] Information and communication sciences, see **http://geriico.recherche.univ-lille3.fr/**

[2] Deposit of Dissertation Data in Social Sciences and Humanities – A Project in Digital Humanities, see **http://d4h.meshs.fr/**

The context

To foster uptake and increase efficiency and outcomes, the design and implementation of such a program must evaluate the specific conditions of the immediate and wider environment, including the expressed needs of the community being served. Here are the essential context features for the Lille program.

National context

- Since 2006, the French universities have a centralized system for the deposit, indexing and preservation of ETDs called STAR[3]. The ETDs are reported in the French academic union catalogue SUDOC and on the platform for PhD dissertations[4]. The author can opt for open access via an institutional or other academic open repository.
- In 2018, the digital deposit of PhD dissertations on STAR will become mandatory for all universities, faculties and departments.
- STAR allows the deposit of supplementary files including datasets but does not provide specific tools for their curation and publishing.
- With funding from the CNRS (National Centre for Scientific Research) and some universities, the consortium Huma-Num is developing an infrastructure for the SSH research communities. This infrastructure includes a platform for the deposit, curation, recording, publishing and preservation of datasets (NAKALA).
- The CNRS is also developing online services to improve the data literacy of scientists and professionals and to facilitate the preparation of data management plans[5].

Local context

- The University of Lille has a mandatory ETD policy; all PhD dissertations must be submitted in digital format, for deposit in the national STAR system.
- A new institutional repository is under development (Dspace). It is uncertain to which extent it will accept the deposit of datasets, especially from PhD students.
- The SSH campus provides a solution for the sharing of files in the cloud; however, the storage capacity is limited, and the server does not guarantee long term preservation

Needs

- Following our own and other survey results (Schöpfel & Prost 2016), scientists express above all a need for storage and long term preservation solutions, along with advice and assistance for RDM, for both the deposit as well as for the description and legal aspects. Data sharing, especially open access, is not a priority.
- PhD students have less experience with RDM but are more motivated than other scientists in data sharing. They are also motivated by the RDM related criteria of project calls and funding agencies' programs, especially by the EU Framework Programme for Research and Innovation[6] which requires the submission of a data managemen plan (DMP) along with the project proposal[7]; they know they must be compliant with these criteria in order to get funding for their research.

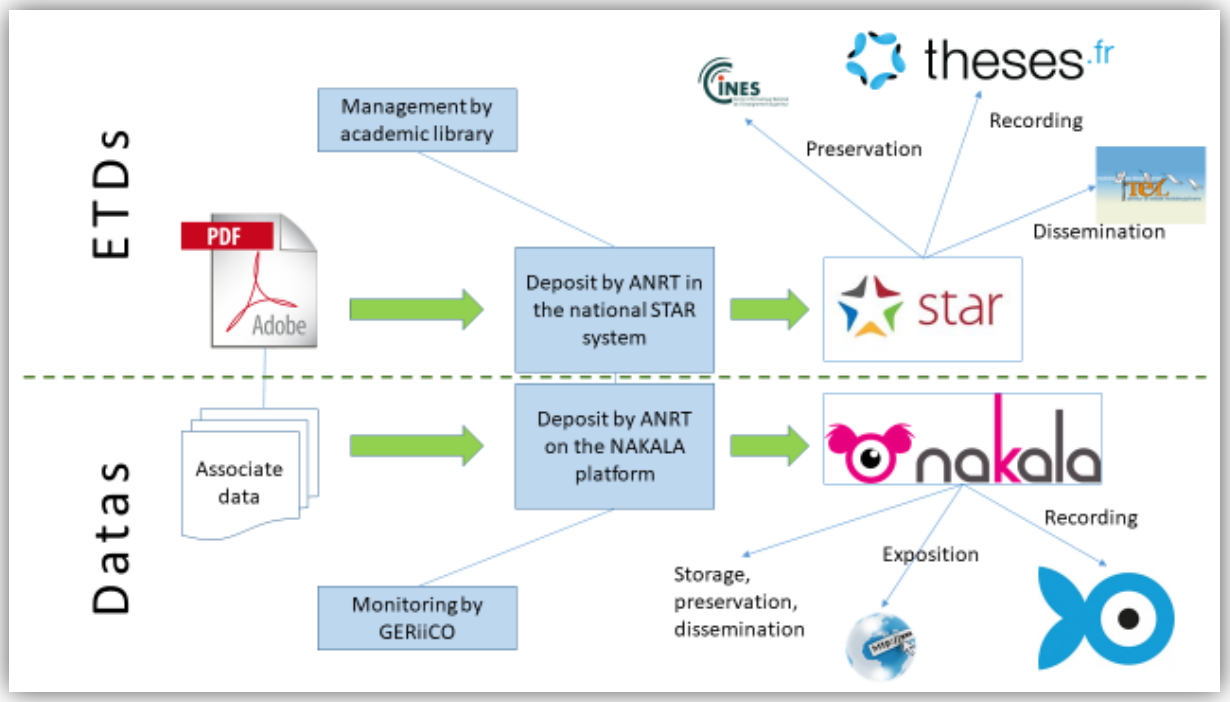
[3] http://star.theses.fr/
[4] http://www.theses.fr/
[5] Platform DoRANum http://doranum.fr/ and service DMP OPIDoR https://opidor-preprod.inist.fr/
[6] Horizon 2020 or H2020, see https://ec.europa.eu/programmes/horizon2020/
[7] See the H2020 guidelines http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm

The development of a local workflow

Compliant with our surveys on the emerging environment of data repositories[8] and, in particular, on similar projects[9] we decided to create a local workflow for the deposit of research data by PhD students. Our assumptions:

- No new data repository (institutional/local or disciplinary) but a connection to existing infrastructure in SSH.
- A solution linked to the national ETD system by metadata and identifiers but independent of this system.
- A separation between data and dissertations from the beginning and onwards (separate deposit).
- A "by default" solution, complementary to specific data repositories.
- An integrative, complete solution covering all needs (recording, preservation, dissemination...).

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. Figure 1 presents the main aspects of our solution.



As before, the ETDs will be submitted to the national STAR system by the academic library which in 2018 will integrate the actual ANRT staff; via the STAR system, the ETDs will be preserved by the national CINES agency, with their metadata being disseminated by the national academic union catalogue SUDOC and the national ETD portal Theses.fr. Following

[8] Nearly 2,000 sites indexed by the international directory re3data http://www.re3data.org/

[9] 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)

the PhD students' choice, the text of the dissertation can be published on the national open access TEL server, the Lille institutional repository and/or on another platform.

The same staff will deposit the associated datasets on the NAKALA platform, create the metadata and link them to the dissertation on STAR. After formal validation and acceptation of the files, NAKALA will guarantee the preservation, the dissemination (following the students' choice), the exposition of the metadata on the web and the indexing by the Huma-Num discovery tool ISIDORE[10]. Like the ETD deposit, the academic library will supervise the submission of datasets together with the research laboratory GERiiCO[11], which will be in charge of the scientific follow-up of the workflow.

Figure 1: Local ETD/data workflow

Thus, our main problem was not the creation of a new system but the connection between existing systems, with questions related to compliance and interoperability. The discussion with the NAKALA team identified eleven specific issues where action has to be taken:

Content/coverage

- Granularity: what exactly should be defined as a dataset for deposit? We have discussed this question in two communications (Schöpfel et al. 2016, 2017). 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 format: which formats can be accepted? While the national ETD system only accepts PDF files, the NAKALA data repository supports all file formats that can be accepted by the national academic digital archive in Montpellier[12] so that we can use their checklist FACILE as a filter for the validation of acceptable file format[13].
- Database: how should larger databases (surveys, inventories, text samples etc.) be dealt with? What are the limits for deposits on the NAKALA platform? This issue is part of the tests with the Huma-Num team.

Metadata

- Indexing: who should do the indexing? Our idea is that the indexing should be done and supervised by information professionals, based on the basic metadata provided by the PhD students for the national ETD system STAR.
- Data structure: how should data be described and structured? Our option would be to apply the Metadata Encoding & Transmission Standard of the Library of Congress[14] but we still have to assess the compliance of METS with the NAKALA platform.
- Referential: we decided to index five elements of the Dublin Core following a qualified metadata schema (file name, data type, creator, date, title). This means that we have to

[10] ISIDORE combines a search engine and a metadata harvester for all kind of SSH data from the Huma-Num infrastructure **https://www.rechercheisidore.fr/**

[11] Information and communication sciences, **http://geriico.recherche.univ-lille3.fr/**

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

[13] **https://facile.cines.fr/**

[14] METS **http://www.loc.gov/standards/mets/**

prepare precise descriptions and term lists and determine what is acceptable for these DC elements. These metadata together with the ETD and data identifiers will be used for the connection between dissertations on STAR and data on NAKALA.

- Identifier: which unique identifier should be used for the datasets? Even if France is part of the DataCite consortium for the assignment of DOIs[15], we opt for the moment for the handle system which is applied by the Huma-Num infrastructure, but we remain open for future adoption of the DOI.
- Source code: a last issue is how to describe sources code-related to datasets. How can this information be included in metadata? So far we have no solution. Perhaps, this is out of scope, at least for the moment and/or for this project.

Other issues

- Legal aspects: we anticipate legal issues like copyright, third party rights, privacy etc. Our approach is twofold: we provide basic 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.
- Deposit: who has access to the NAKALA platform? Who is an authorized user? Our first choice is to limit access to the project team (i.e. information professionals of the academic library, with a generic address and identification via the national academic IT network RENATER) and to prohibit self-archiving. However, this may change in the future.
- Data size: actually, we don't know exactly what will be the potential data volume. In average, 60-80 PhD dissertations are submitted per year on our campus, representing roughly 2 GB. But except for very few dissertations, these deposits in the national ETD system do not contain data files. So all we can do is try to make some estimations, perhaps also with our international partner projects.

Four other issues have been raised but they are not directly linked to the development of the workflow:

- Long term preservation: Up to now, the NAKALA platform does not guarantee long-term preservation of submitted datasets. But they have an agreement with the national CINES agency which ensures long term preservation of backups of the different Huma-Num platforms' content, which means that the NAKALA datasets could be recovered if necessary.
- 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.
- Promotion: we have already discussed how to promote the new data service - who should do this, what the best communication vectors are, and what should be the message. As mentioned before, the main message will not be "PhD students must share their data" but rather "we can provide a solution for the preservation of your research data". And the message will be communicated via the Graduate School, the Research Department and research laboratories and the academic library. Also, our intention is not to make the

[15] **http://www.inist.fr/?DOI-Assignment&lang=en**

deposit of datasets mandatory but to promote and incite data deposit as a form of good scientific practice.
- Technical documentation: after the launch of the new workflow, we will have to write the technical documentation, a procedure on two levels, one for the professional staff, the other for the students in the form of guidelines or recommendations to facilitate the process of submission and deposit.

The tests of the new workflow started at the end of September 2017. The workflow will be operational in 2018. We will perhaps customize the Huma-Num interface for the submission of datasets and the creation of metadata but this is not essential for the project.

We mentioned above that the University of Lille has started to develop a new institutional repository (Dspace). A priori, this would not modify the workflow for ETD related datasets, as it would not modify the submission of ETDs to the national STAR system. As Dspace is able to harvest metadata and to integrate different identifiers and outbound links, connecting the NAKALA datasets to the institutional repository should not be a problem.

The PhD training program

Our assumption is that RDM will become a part of basic scientific skills and good practices of research work, like sampling, statistics, surveys or systematic reviews. PhD students will have to obtain data literacy as part of their scientific education and training program. We started to work with PhD students on RDM nearly three years ago. Our experience is that most of them get some elements of data literacy through their research practice (e.g. privacy issues, ethics, backups) but they lack an overall ability of data curation, including description, preservation, and sharing[16].

Following this assumption we launched a training program to develop and enhance RDM skills for PhD students in SSH, together with the research support service of the academic library but as part of the Graduate School training program, not as a library course. The integration in the official program of the Graduate School and the mixed pedagogic team (four scientists and one academic librarian) partly explains the success of the program, together with the newness and interest of the topic itself. Since the beginning, about 40 PhD students have participated.

Our first training program (2015) consisted of three seminars (3 x 6 hours), organized together with scientists from the University of Lille and other institutions, on research data management, legal aspects and potential reuse and exploitation of data, including content mining.

In 2016 and 2017, we organised the program in a more traditional way, as a seminar with seven sessions (6 x 3 hours and 1 x 2 hours), one session per month from January to June, on different topics (figure 2).

[16] Concerning the global concept of RDM, see the overview by Neuroth et al. 2013

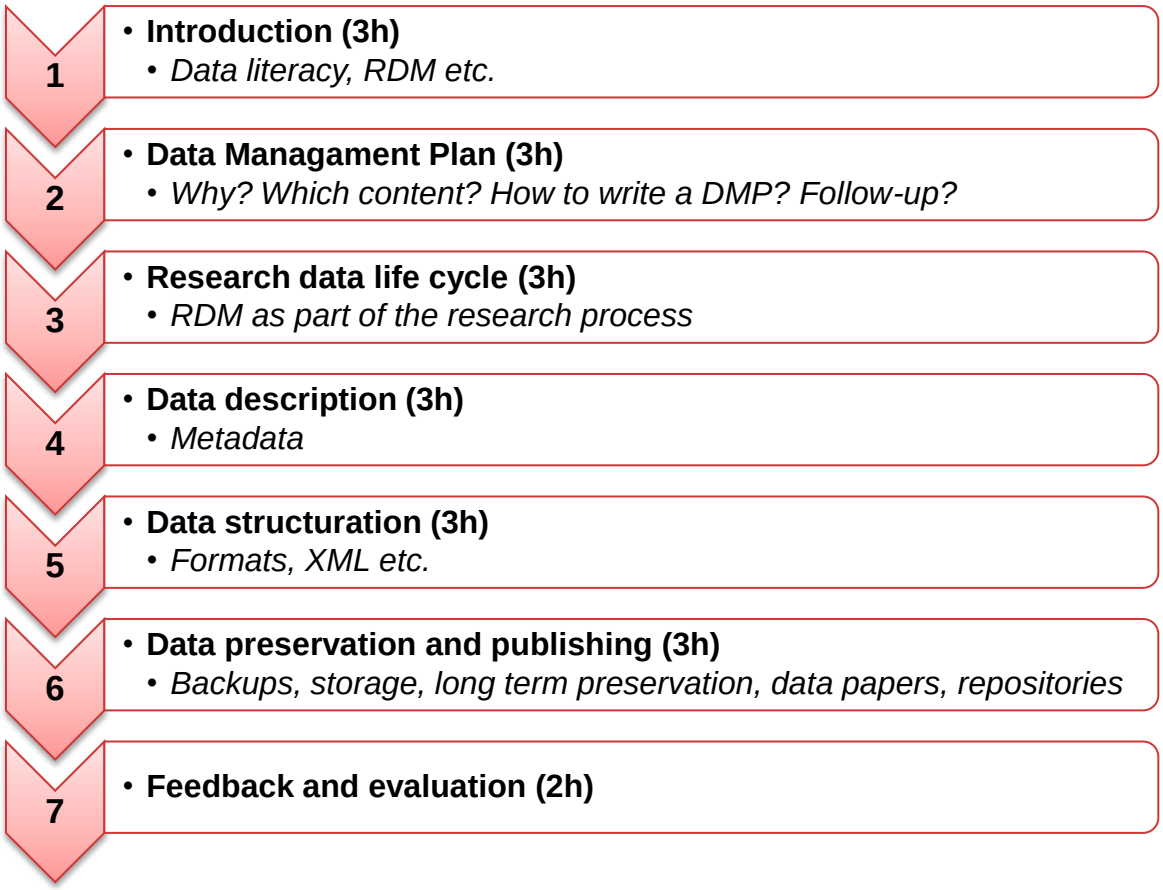


Figure 2: RDM training modules

We try to cover the most important elements of RDM, with a double objective: provide theoretical and operational knowledge on research data and data management (data literacy), and develop practical skills (data behaviour). Based on the feedback of the first two years, our 2017 seminar focuses on the PhD students' own scientific experience and data behaviour; in each session, the introduction and overview is followed by discussion, practice and individual follow-up. Each session takes place in a computer room.

The operational goal of the seminar is the writing of a data management plan for each PhD project on the OPIDoR platform[17], which is the French adaptation of the JISC DMPonline service[18]. Each student creates his/her personal account on OPIDoR, writes a DMP with the European Commission's H2020 template and shares the DMP with the seminar's training staff, which provides individual follow-up, comments, suggestions etc. directly on the platform. At the end of the seminar, the staff downloads the final version of each DMP for evaluation and direct feedback.

[17] Hosted by INIST https://dmp.opidor.fr/
[18] Hosted by the JISC Digital Curation Centre https://dmponline.dcc.ac.uk/

Concretely, each student completed the H2020 initial DMP template, i.e. a general description of the research project followed by five issues with specifications for each dataset:

- Dataset reference and name
- Dataset description
- Standards and metadata
- Data sharing
- Archiving and preservation (including storage and backup)

For instance, a student preparing an anthropological and ethnographic study on the perception of the 2016 Olympic Games by the people from Rio de Janeiro described three different data types, based on questionnaires, interviews and photos, and explained how he will index the data sets, with whom he will share them, and how he will store and preserve them.

Our experience with this approach is promising: with personal guidance, follow-up and feedback, the PhD students not only learn to write their own DMP but, in doing so, they learn to anticipate the essential issues of research data curation, like standard description, systematic back-up and secure storage and, in the end, long-term preservation and publishing. They also learn to prepare DMP in good and due form, compliant with the H2020 criteria, which will be an essential asset for future research work and project submission. Also, their DMP will contribute, as didactic material, to further training.

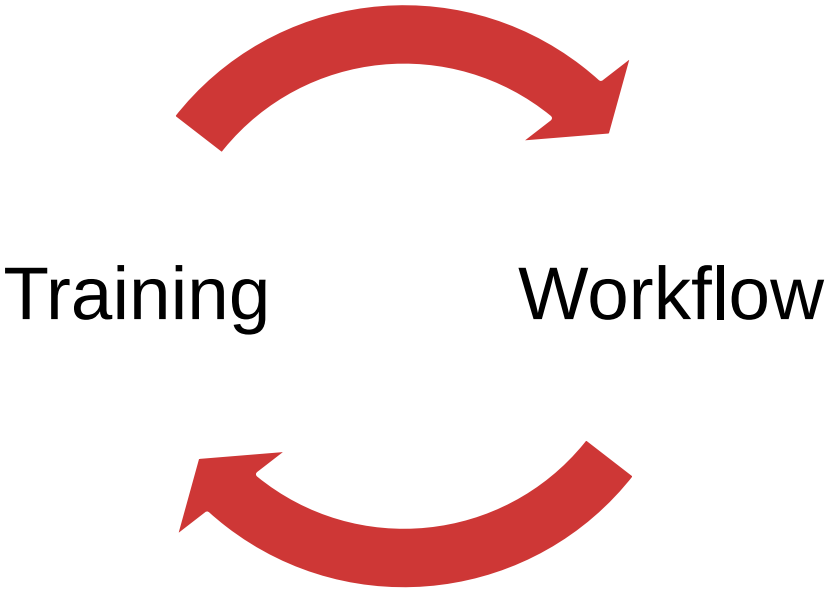
Nevertheless, we also observed that students who are just beginning their PhD have different questions and needs from those in second or third year of graduation who already have their methodology and often also datasets. In the future (2018), we will probably divide the seminar in two parts, the first one (modules 1-3, see figure 2) for "beginners" and the second (modules 4-7) for "advanced". Even so, the writing of a DMP will be part of both.

Lessons learned

We must keep in mind that our primary project target is not to foster or increase open access to/in research data. Our main goal is to help young scientists in SSH develop their data literacy, i.e. their RDM related skills, to raise awareness about the challenges of data curation and publishing, and to provide a solution by default for their datasets. Our project is an investment in the future, and we expect the young scientists to have efficient and modern data behaviours, just as we expect them to develop new methodologies and conceptual approaches in their research field. Data sharing is part of the game, but not the only or most important purpose.

1 - Our experience confirms that in many respects, RDM is not just a technical problem but a “people problem” (Ward et al. 2011). “Improving tools are not the only steps necessary to overcome barriers. The next steps will likely involve training for scientists (...)” (Tenopir et al. 2015). In fact, solutions are often available (cf. Kindling et al. 2017). What is needed are new skills[19], information, promotion and incitation to use these solutions; paradoxically, we can

[19] Of course, this includes also the acquisition of new data skills by the project team.



say that our focus is not on data but on people. The most important decision of our project was perhaps the choice of a specific target group - young scientists and specifically PhD students in SSH. It was this choice that made our project intelligible and distinctive, on our campus as well as in the French HE landscape.

Figure 3: PhD training and ETD/data workflow

2 - Technical solution and education are related (figure 3). Data literacy (DMP, RDM) is more and more considered as good scientific practice, like usual scientific skills (sampling, statistics etc.). Our experience and conviction is that it is not enough to develop RDM tools if we do not teach scientists how (and why) to use them. Therefore, available infrastructure will shape the content of the training program (e.g. for file formats or dataset structuration); but then again, the discussion and feedback from the training program contribute to the further development of the campus-based RDM solutions. Our approach can be considered as an organizational learning process on the campus.

3 - The project is run together with the academic library research support team. However, it is NOT a library project. The library staff is part of the project, with specific tasks and skills. Their contribution is essential for the success of the project, and they are of course members of the project steering committee. Yet, from the beginning the project was designed as a research project with a doctoral training program, under the responsibility of the graduate school, with a scientific project management held by our research laboratory, and under the political and strategic leadership of our academic research department. There are at least two reasons: legitimacy (in the sense that scientists have everyday experience with RDM); and the fact that scientists do not usually consider RDM as a "library affair" but as part of their daily research work with other scientists and technical staff.

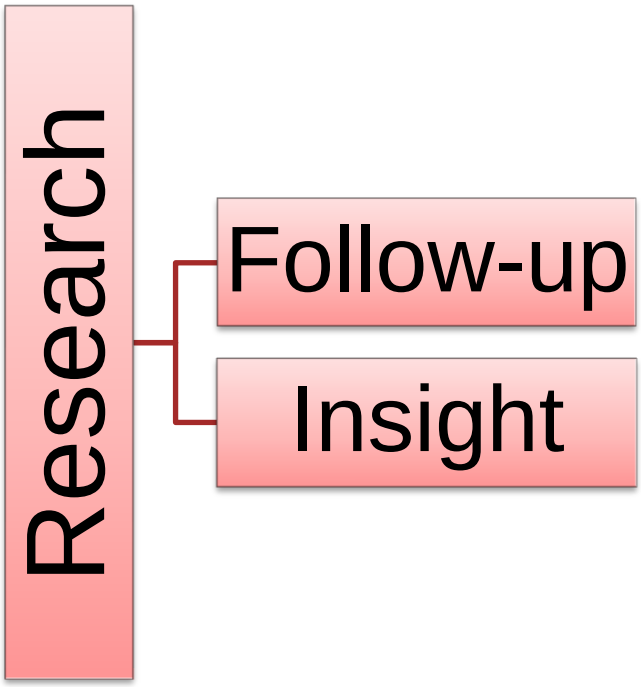


Figure 4: RDM related research

4 - The project is not limited to education and workflow. As said above, its main character is research. We can distinguish two different levels of RDM related research, i.e. follow-up and insight (figure 4). Follow-up means monitoring and assessment of the training program, including feedback and continuous adjustment; and it means evaluation of uptake and usage of the new data workflow. Insight covers a larger range of data related topics, such as type, format and content of datasets, the impact of data on format and content of ETDs and the text and data mining of dissertations and data. At present, we are working on the first issue, together with partners from different universities and institutions. In the future, we will focus on the second issue and prepare, together with German colleagues and ProQuest, an international research project on new formats of PhD dissertations.

References

AWRE, Chris et al. Research data management as a 'wicked problem'. *Library Review*. 2015, 64(4/5): 356-371.

CHAUDIRON, Stéphane et al. *Livre blanc sur les données de la recherche dans les thèses de doctorat.* Villeneuve d'Ascq: Université de Lille 3, 2015.

KINDLING, Maxi et al. The Landscape of Research Data Repositories in 2015: A re3data Analysis. *D-Lib Magazine*. 2017, 23.

NEUROTH, Heike et al. (eds.). *Digital curation of research data. Experiences of a baseline study in Germany*. Glückstadt: vwh, 2013.

PROST, Hélène, MALLERET, Cécile & SCHÖPFEL, Joachim. Hidden Treasures. Opening Data in PhD Dissertations in Social Sciences and Humanities. *Journal of Librarianship and Scholarly Communication*. 2015, 3:eP1230+.

SCHÖPFEL, Joachim et al. Open Access to Research Data in Electronic Theses and Dissertations: An Overview. *Library Hi Tech*. 2014, 32(4): 612-627.

SCHÖPFEL, Joachim, PROST, Hélène, & MALLERET, Cécile. Making data in PhD dissertations reusable for research. *8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic*.

SCHÖPFEL, Joachim, & PROST, Hélène. Research data management in social sciences and humanities: A survey at the University of Lille 3 (France). *LIBREAS. Library Ideas*. 2016, 29:98-112.

SCHÖPFEL, Joachim, PROST, Hélène, & REBOUILLAT, Violaine. Research data in current research information systems. *CRIS 2016, St Andrews, 8-11 June 2016*.

SCHÖPFEL, Joachim, KERGOSIEN, Eric, & PROST, Hélène. « Pour commencer, pourriez-vous définir 'données de la recherche' ? » Une tentative de réponse. *Atelier VADOR : Valorisation et Analyse des Données de la Recherche, INFORSID 2017, 31 mai 2017 Toulouse (France).*

TENOPIR, Carol et al. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE*. 2015, 10(8):e0134826+.

VOMPRAS, Johanna, & SCHIRRWAGEN, Jochen. Repository workflow for interlinking research data with grey literature. *8th Conference on Grey Literature and Repositories, National Library of Technology (NTK), 21 October 2015, Prague, Czech Republic*.

WARD, Catharine et al. Making sense: Talking data management with scientists. *International Journal of Digital Curation*. 2011, 6(2):265-273.