DARIAH Winter School in Prague

Open Data Citation for Social Sciences and Humanities

24th to 28 of October 2016

Session 1: Introduction

Introduction

Welcoming remarks

Lucie Doležalová (Charles University)

Lucie Doležalová who was the local organiser of this event has chaired this introduction. She works as Associate Professor of Medieval Latin at the Institute of Greek and Latin Studies of the Faculty of Arts, and at the Communication Module of the Faculty of Humanities, both Charles University in Prague. She is also a researcher at the Centre for Medieval Studies of the Academy of Sciences of the Czech Republic.

Mirjam Friedová (Dean, Faculty of Arts, Charles University)

It is really nice to see people from such different countries, even beyond Europe, being our guests and I hope you will enjoy the workshop and Prague. Charles University is the biggest university in the country and the elite of our universities. It has over 50 000 students and 17 different faculties. I represent one of the schools behind the organisation of this event, the Faculty of Arts, and we are very happy and proud about it. As I have worked with medieval manuscript myself, I have a sense of what digital humanities could be about and have been about, so I am very excited that is happening right here.
Let me just wish you a very good school, a very good gathering, a very productive and inspirational, so you all leave Prague with new ideas and contacts, new possibilities for projects. Welcome again and thank you very much for coming.

Marek Skovajsa (vice-dean for research, Faculty of Humanities, Charles University)

I would like to join Professor Friedová in welcoming you at this event. I am from the Faculty of Humanities. I would to thank everybody involved who made this event possible with the partners and all the organisers from other countries. It is a pleasure for me to welcome here people from many European countries and countries from outside Europe. I will give regards from the dean of my faculty who unfortunately is not able to be here. As a public university, we are trying to combine both tradition and the most advanced approaches and technologies. I think it is a very important thing to develop digital technologies and infrastructures in the Art and Humanities. I am sure that this Winter School will contribute to strengthen the foundation of the digital humanities in Europe. To conclude, I wish you a very pleasant stay in Prague and thank you for coming.

DARIAH introduction: Issues and digital turn

Emiliano Degl'Innocenti (DARIAH-IT)

I will present DARIAH in Italy and with an overview of the wider perspective of the European landscape of the digital humanities in which DARIAH is involved.

Digital scriptorium

We cannot avoid anymore dealing in a serious manner with the complex framework of the *digital turn*. We are now moving from what was called during the Middle Ages the scriptorium, a sort of vertical environment where all research work was carried on to something that we call the digital scriptorium, some kind of digital environment where all this complexity is reflected and contained.

A complex digital ecosystem

Europe has a long established tradition of digital Arts and Humanities research. There are a lot of projects and infrastructures within the Esfri landscape. We are really proud of it but we also know that there is a certain lack of connection and sustain with the results of those efforts and projects. We need to address this complex situation on the long run. This digital ecosystem is vast and rich; it includes a lot of high quality digital objects, ranging from text to databases, from digital images to a great number of digital tools that support various basis of daily research work. But it is still fragmented and you will experience a lack of data and tools interoperability issues. So there is a real need to find a more interconnected and interoperable digital ecosystem.
Another issue is the need to bridge the gap between two kinds of cultures we are dealing with: we need a certain level of communication and collaboration between the humanities and other branches of scientific research. In most cases, we are dealing with more or less the same object, for example with the applied cultural heritage, the restoration, the preservation of artefacts, etc. It is obvious that there is a huge gap and a very different concept of what a digital object is, the terms of its production, the collection of data, the management of data and also validation of results. So, we are not able to communicate across those two environments.
Furthermore, we still lack a common epistemological methodological background as well as a common set of standards and framework to evaluate the results, tools and products of digital humanities projects.
To be synthetic, we need to reduce the fragmentation of this digital ecosystem. We need to develop an efficient vision of data lifecycle and a sustainable data management plan. We need also to develop a broader framework for permanent research identification and preservation, and move forward with a Linked Open Data strategy in order to make this landscape more interoperable and interconnected. Then we should also bring bridge between the tangible and intangible cultural heritage branches.

FAIR principles

The principles develop by the FAIR approach are Findability Access Interoperability and Reuse. Those items should be the keys for this evolution because we are now in the situation where almost the total amount of available tools, datasets, and everything needed in order to move forward the research agenda is in a digital format, is within the digital landscape. So we are now really facing two challenges: moving traditional research, preserving all the needed content that are important for the scholarly community into this new digital framework. This means that we need to promote and support the data intensive research implements. There is no accepted definition for this data science terms but what is clear is that we are all facing a *data deluge* and we should be able to select, to create new approaches and to try to move from traditional research path to innovative research path by combining computer hacking, better analysis tools and a sort of problem solving attitude.

DARIAH

DARIAH is an ERIC, which is an acronym that means *European Research Infrastructure Consortium*. It is a great tool available for researcher in order to solve, or at least to address from political and infrastructure perspective, the various points of data use in this context. To address all this issues and problems, DARIAH started in 2006 and became in 2014, after a *preparatory phase*, an ERIC, an effective research infrastructure. Now DARIAH is in its *construction phase*, this means we have to select carefully all the priorities we want to work on in order to satisfy the needs of the research communities that are joining DARIAH.

Missions and priorities

- Enhance and support digitally enabled research,
- Promote cross utilisation among different disciplines in the Arts and Humanities landscape,
- Offer services and activities that are centered, not on technology, but on research and the needs of the research communities.

We have a number of different disciplines that are represented in DARIAH at various levels: from the scientific committee, to *Virtual Competency Centre* (VCCs) that are containers where all the scientific needs and issues are discussed in order to create outcomes from the political, from the infrastructure point of view and also where the actual research communities are represented as we continuously engage with scholarly networks, Cost actions, etc. We also set up working groups that are dealing with concrete research problems. DARIAH is not alone in the Esfri landscape, it is within a number of other e-infrastructures and projects, like CLARIN for example. There is in fact a constellation of different projects ranging from digital archives to aggregation of research, to archeology and other connected issues, etc.

Role of DARIAH

Within this vast landscape, where different actors are working together with the aim of reducing this complexity and fragmentation, the role of DARIAH in this *construction phase* is to make the dissemination of scholarly data in the Art and the Humanities more fluid. Fluid means to avoid not necessary transactions between data providers and the researchers. As

a result, users could waste less time doing something that is not research. So it means focusing on research rather than on technological transactions.

Data Charter Reuse

This will be carried on by supporting a second action. First of all, we are trying to build a framework called Data Reuse Charter to support those activities. We are now trying to organize those data reuse charter in different countries (Italy, Ireland, Spain, France, etc.) in order to make it concrete. We are selecting stable stakeholders ranging from the galleries, the libraries, the archives, the museums, research institutions in a bottom up approach to prioritize the needs of the research communities in order to produce some framework of licensing, about the possibility to use, to reuse, to make all the data more interoperable, by providing information to users and also trying to make clear the requirements coming from the users and form the data providers. Everything will be presented in a few months, as it is a currently ongoing process, in order to receive a feedback and and to evaluate the relevance of this agenda.

Pierre Mounier (EHESS, OpenEdition)

I would like to take the opportunity to thank warmly Charles University, the Faculty of Art, the Faculty of Humanities for hosting this event and to welcome us so nicely. And I especially would like to thank Lucie Doležalová and Marjorie Burghart for making this possible.

This Winter School

What is this meeting? When we worked with Nathanaël Cretin on the preparation of the event, we couldn't find a simple name for it: *Open Data Citation for SSH, DARIAH's Humanities at Scale Winter School in Prague, in Charles University*. It is like a chimera, made from different parts working together. Is it about open data? Yes, but not only. Is it about open access? Yes, but not only. Is it about humanities disciplines? Yes but not only. Is it about digital? Yes, but not only. Behind the organisation of this event, you have two French institutions: OpenEdition and Huma-Num. So, is it a French event? Yes, but not only. We are in the Charles University, in Prague, in Czech Republic, so is it a Czech event? Yes, but not only.

Integration

As Emiliano Degl'Innocenti explained, it is about integration.

- Integration between Humanities disciplines which are very fragmented.
- Between Humanities and digital. We know that the articulation and the integration between those two fields is not easy.
- It is also about integration in Europe, between the European countries and between its scientific communities. I would like to put some stress on that because we all know what is going on in Europe, so I think that it is our political responsibility as scientists, as scholars and as humanists to work together more tightly and to enhance our collaborations, because I think Europe needs that, particularly for science and culture.

What is it about?

- First, it is about Digital Humanities as a way to integrate the emerging and powerful digital field and the traditional humanities disciplines. Tradition and innovation. We will see how concretely this integration will work on the subject we are going to work on.
- It is also about Open Science which is often a buzzword, but we should make it more than a buzzword and make it concrete. For most of the people, open science is open data plus open access plus citizens engagement, but it is not just an addition. Real open science is the integration of that. And this is exactly the point of our Winter School: to integrate open data and open access to make it meaningful and useful for the citizens.

An anecdote

Now, I would like to conclude with an anecdote: last week, I was at the Pubmet conference in Zadar, Croatia. It was about publication and metrics in the open access framework. There were people from a very nice open access journal. They were speaking about how nice it could be for them to enhance their publications from the traditional print publication towards publication with some additional material, multimedia materials. It is particularly meaningful for many disciplines, in fact for all humanistic disciplines. Someone asked them how they could imagine the way there could be an interaction between traditional publication and multimedia material. They said: “Ok, here is how we imagine that: we have the text in a pdf and we have links inside the pdf pointing to those multimedia materials that the author could put on its personal website”. I think we have a lot of work ahead of us! This anecdote introduces the importance of this Winter School for our communities and I would like to thank a lot DARIAH which is the overarching institution for organising this meeting.

The status of data: What is data?

Joachim Schöpfel, Lille University

It is with a great emotion to be in the oldest university of Central Europe, the center of culture and science of this part of Europe, it is great to teach here with you.
In this session, we will talk about research data, not only on social sciences and humanities, but in publication and I will present you what we are doing at Lille University. I am working in SSH and information and communication science. My speciality is scientific information. I am German, I have a PhD in psychology, but I am working, teaching and doing research in information science for more than 25 years in France. Terminology, categories of data, critical issues and data in publication is the heart of my presentation. It is not about research data management and I will be extremely short about data journals because it is part of another session.

Terminology and categories

The US government defined data in a broad way as “Recorded factual material commonly accepted in the scientific community as necessary to validate research findings”. The University of Edinburgh has another one: “Re-usable research results, collected, observed or created for purposes of analysis to produce original results”.
In fact, definitions are more about functions (validation, reuse, innovation) and types (not by nature). The question is to know what is information, numbers, facts?
=> Research data refers to information, in particular facts or numbers, collected to be examined and considered and as a basis for reasoning, discussion, or calculation. In a research context, examples of data include statistics, results of experiments, measurements, observations resulting from fieldwork, survey results, interview recordings and images”, for H2020 program.
“Data are like cows. If you look them in the face hard enough they generally run away” (adapted from Dorothy L. Sayers).

General typology

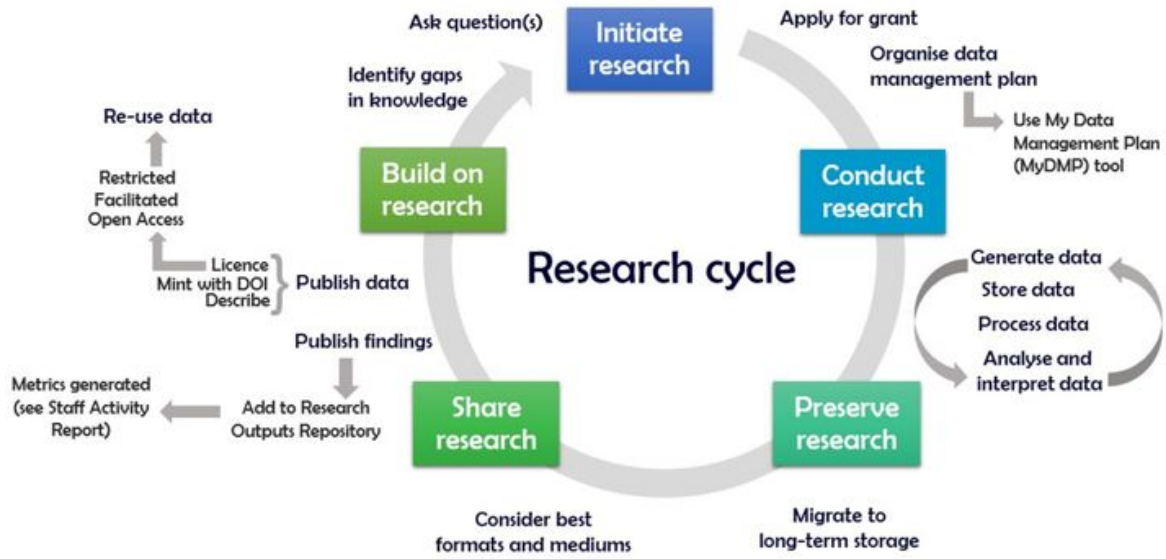
- Research methods as approach to make different levels of data: Observational data, Experimental data, Simulation data, Derived or compiled data.
- Input and output: For my own work, especially with PhD dissertation, but also with articles and report, there is an important distinction between input and output with two categories: data collected and used for research and data produced within research: primary data (collected) and secondary data (produced).

You can find a lot of categories with these two examples that are not specific to SSH, but are multidisciplinary.

- From re3data (REgistry of REsearch Data REpositories): archived, audiovisual, configuration data, database, image, plain text, raw data, etc.
- From HUB (Humboldt University in Berlin) who made surveys a few years ago: observations, experiments, surveys, etc.

- Lille study: We conducted a study in SSH, with PhD dissertations and with scientists and students in general. It gave us two different lists: survey data, texts, spreadsheets, databases, multidimensional visualisations and models, audio recordings, maps, software, etc. The most important formats of data produced by SSH scientists on our campus are mainly texts, spreadsheets, excel files, statistics, timelines and databases.

Link with disciplines: Data and publications

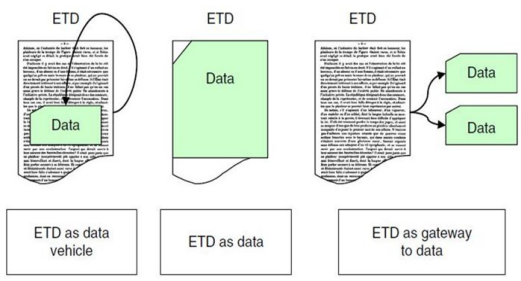


Description of research data management related to research cycle
Source: http://guides.library.unisa.edu.au/ResearchDataManagement

As a researcher, I would like that my own research, or the research from my colleagues, would be like this, as a cycle, a straight process, but of course it is never the case. This is a model, intellectually satisfying, not always real. It is meant to make understandable some aspects. This ideal research cycle is related to data management, as a kind of umbrella concept for many different things, from backup to indexing, sharing and making data reusable. For the end of the research cycle [the left side of the schema], it is interesting to see the publication of data and the links between publication and data.
Elsevier tried to draw different levels of this relationship between data and publication from data published in a research article enhance data explanation in supplementary files, data referenced in research articles and available in repositories, and data publications describing available data, especially data journals.

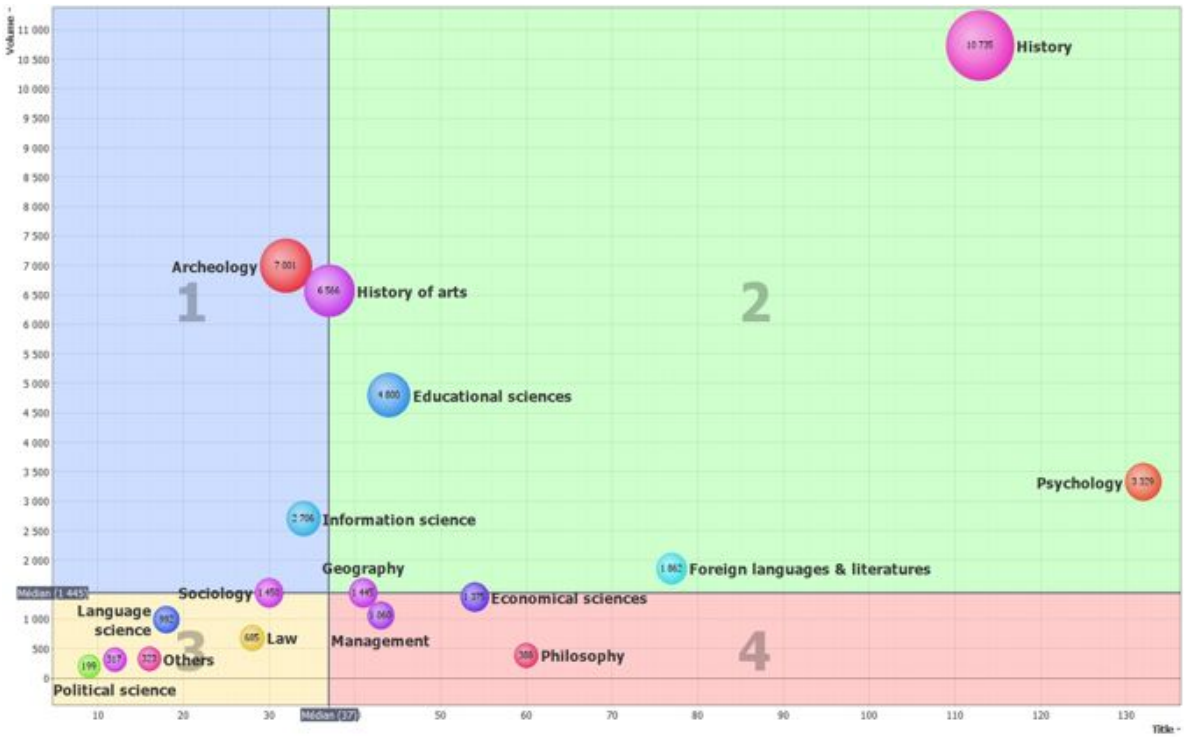
Data in dissertations

=> From our work on PhD dissertations analysis, publications, documents like PhDs, reports, etc., can be considered as data vehicle (as supplementary material), gateway to data (when publication contains links to data, integrated or not in the text), but also data sui generis (exploited as primary data source for TDM).



Disciplines and categories of data

In our study conducted with University of Ljubljana in Slovenia, we evaluated the research data included in PhD dissertation (approximately 800 PhD dissertations in SSH). It allowed us to illustrate that the volume of data (in pages, vertical axis) and the number of dissertations with data (horizontal axis) is very different, at least in our sample, between the disciplines.



For example, in History you have many dissertations with many data included, in Psychology you have many dissertations with less data included, in Archeology you have less dissertations but many datas included in the dissertation, etc. A great emerging question for us is how to make those data available, because they are often not reused after PhD. So we need to train students to share and to make them reusable.

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

X: Typology -

=> On the one hand, different categories of data are not really related to one specific discipline. Each has a kind of discipline profile. Probably, the data type is more conditioned by tools, instruments, methods or procedures (surveys, experimentations, simulations…) than by disciplines. On the other hand, each discipline has a kind of specific data profile. So you can describe disciplines by the data and you can describe data by the disciplines. Even if it is not a big news, it is important to be aware of this if you want to work with data: you can't have a disciplinary approach only, you can't have a one size fits for all approach.

Data in dissertations: Issues

- Incomplete, inadequate or missing description: you cannot even understand the data provided by PhD students: data sets and individual data are not completely documented
- Missing organisation: data are not structured, not correctly presented, all is mixed up, information mash-up not suitable for further research
- Inadequate format: ex. Pdf: data and text are glued together instead of being separated and published in adequate files formats: not easy to reuse. Even if it is possible to get an xml file from a pdf, if you have database, it is better to produce it as a database, not as a pdf.

Enhanced articles with data

In SSH, it is not really easy to find such articles but what is interesting is the category in the description of each article: data availability. For instance, in the *Palgrave Communications*, I didn't find any article with data effectively available. On the other hand, there were several articles clearly mentioning that data are not available because they had to protect people

involved, in gender studies for instance, or with regards with privacy issues, confidentiality or some datasets were only available on demand: you have to ask the author, not the publisher, to get the data. The third category was the most important: data not available because the research did not produce any data. In fact, there can have some confusion here when author use collected data but do not mention its availability.

Some examples

- Reference: Prost, Hélène, and Joachim Schöpfel. “Les données de la recherche en SHS. Une enquête à l’Université de Lille 3.” Report. Lille 3, 2015. http://hal.univ-lille3.fr/hal-01198379/document.

It is hosted on the French National repository in France, HAL, which is organised by laboratory collection. You can deposit a spreadsheet as a complementary file to a publication, but this file is not described and not indexed, no documentation, no persistent identifier and you can’t search for it because it is considered as something complementary to the main deposit which is the report.

- Reference: Schöpfel, Joachim, Južnič Primož, Hélène Prost, Cécile Malleret, Ana Češarek, and Teja Koler-Povh. “Dissertations and Data,” 2015. http://hal.univ-lille3.fr/hal-01285304/document.

Another example from us, in a different way: a communication, a keynote from the research we did with the colleague from Ljubljana about research data in PhD dissertation. Again the keynote is deposited in HAL server in France and my colleagues did the same in Ljubljana in their institutional repository. But this time, we didn’t deposit our data with the report, but we submitted it to the Dutch National data repository (DANS EASY). A DOI is attributed and it points towards 3 files. It gives a link to the dataset on another repository with a DOI, a description and indexing of specific metadata for these files. And you can find it by searching on the web.

- Working Paper on Repec: data are included and deposited together, well disseminated: http://econpapers.repec.org/paper/kudkuiedp/0907.htm
- OpenEdition Books: http://books.openedition.org/ksp/244

Books are available with data in appendix: great and big tables with data. Data are here and they are waiting to be reused, at least used for validation or cross validation. I am sure that if you take contact with OpenEdition or with the author, you will get the access to the full data.

Publication as data

You can now stop to consider publication as a kind of binary object, with on the one hand text, information, conceptual information and on the other hand data, even if you don’t know exactly what is data.

- TDM on research publications: You can now grab data, dissertation for instance, do some data mining. We started with dissertation, others in chemistry and law in the UK. They applied text mining to law dissertation to get out expression, phrases, specifically in legal English, I think it can be useful for foreign language teaching and for translation. This is promising approach to this kind of documents. We try, with our colleagues in Lille and from other laboratories to do the same with Master and PhD dissertations, specifically with geographical names.

- Legal situation: Legal situation up to now in France wasn't really favorable for it in a legal way, but the situation changed now.
- Technical issues: You can do this with pdf, you have to transform it into another format (XML). It would be better to have another format.
- Impact on publication: Structure of the documentation, better understandable for machines. Content: if dissertation can be exploited and linked with data by text and data mining tools, what does it mean for writing dissertation? I suppose, we will not write dissertations as we did so far. And what is the impact of data analysis and tools on publication, on the writing?

Critical issues

We will now see some general issues about the complex reality of the relationship between publication and data

- Separation of text and data: available or not, included or separated, etc. Format: table, photos, tables included in the text, perhaps not tagged as such, difficult to know how to reuse it in a intelligent way, it would be better to separate. Related to dissertation, there are some projects and initiatives in Germany or in Lille to do this in relation to the deposit of data and the dissertation where data is separated: two different workflows, two different ways of processing, of indexing but the link is stable through identifiers and some central metadata.
- Metadata: there is an ongoing discussion about which level of metadata should be applied to research data. There is a debate between generic metadata and field, instrument or domain specific metadata. When it comes to evaluation of data, there is this strong push to have generic metadata (to be able to process in the same and compare metadata from different disciplines). Of course, each scientist will push forward the interest to have specific metadata, the best to explain the specificity of a research data set.
- Preferred formats from DANS: a list of different formats can be used to deposit data. The list is not closed, it is evolutive. For each type of research data, there are many different formats. This must be supported somewhere, someway.
- Persistent identifiers (DOI, ORCID): It can be a big topic when going through the literature. Today, the discussion is only about two identifiers (DOI and ORCID). Some people are processing the data with handle, other with different specific identifiers, but on the international level, when it comes to research data management, the consensus is on DOIs, especially in Europe, managed by the Datacite initiative.
- Altmetrics (DOI) and usage (low). Another issue it that it is not easy to get usage statistics of data sets. No uptake for depositing and sharing, no usage. Data usage is not very high. Altmetrics are an impact measure in social media. Many of these altmetrics are based on DOIs. So there is a problem: when you have no DOIs, you have no altmetrics! On the other hand, when you have altmetrics with documents and files, specific content and format, it is a lot of work, even manual work, to do this. So, for the moment, with altmetrics the only difference is that you don't need Scopus of Web Of Science, but you also have Twitter and Facebook, Mendeley, Researchgate, etc.
- Another issue with the use of data is with the continuum between backup and reusage, when you speak with scientists, most of them are very concerned with

backup, storage and preservation; when you speak with librarians and information officers, they are mostly concerned with reusage and sharing. On the other hand, there is one specific about quality of the site where the data are deposited. If it is for storage or preservation, sharing or reusage, you can do it on a personal website, on the website of your laboratory or your department - good repositories should have a minimum level of quality, a guarantee of long term preservation (5 to 10 years), metadata, identifiers, etc. Today, there are labels like the Data Seal of Approval and other for quality of data repository.

Disciplinarity

- Impact of disciplines: greater on profile than on specific data categories.

Often in SSH, I think there are more impact from methodology and instruments (like surveys) than from discipline. I think there are more similarities between survey data from sociology or education science than between education science and sociology.

- Evaluation: need a standard and generic approach: impact on merging together different disciplines on metadata and on the level of identifiers.
- When it comes to preservation and sharing, you have repositories, like HAL, Figshare and a lot of disciplinary repositories, with specific metadata, characteristics to handle the description.

Research evaluation

We made a research about how the evaluation system deals with research data as it has been developed more than 20 years ago. What we found and communicated was that research data are evaluated as research output, but is also an input!
=> There is a mix between primary and secondary data. And contrary to publication, this system does not evaluate quality or volume of data. It evaluates data management. For publication, the research information system takes into account the number of articles, the number of articles in high impact factor journals, the number of conferences, communications, etc. Regarding research data, nothing like this is evaluated, it is just evaluating if there is DMP, if there is a description and identification (yes or no), which metadata scheme is applied, if data are conserved somewhere and if there is a policy of sharing. So far, up to now, even what research evaluation does with research data is just to evaluate the announcement: “we will do this”. And there is no follow-up. The next step could be: “what did you do with your data?” because in fact it is not about having good data, many or small data, one spreadsheet or big databases, it makes no difference. If you compare this with publication, it would be as if research evaluation just asked if you put your book in the right shelve in the library. In fact, what is evaluated is not the work of scientist, but the work of data officer, information managers and librarians.

Legal issues

- Intellectual property: Career strategy & Publication
- Database protection (sui generis)?
- Third party rights: for example, what we found in dissertations, especially printed dissertations, a little bit older, is that many data are protected by third party rights. Students used it (photo, maps, etc.), put it in their thesis, disseminated in printed

format. If it had been disseminated on the web, there would have immediate problems!
- Confidentiality: Private company information & Corporate secrets
- Privacy Issues?

Political issues

All countries represented in this room have their own open data policy (data produced by public administrations should be disseminated openly, freely, without restriction to, not only to citizens, but to society and also to the corporate sector). On the European level, an open science policy has been formalized this year, with the reference document: Amsterdam Call for action on Open Science, EU2016.
Reference: Zaken, Ministerie van Buitenlandse. “Amsterdam Call for Action on Open Science - Publication - EU2016.nl.” Publicatie, April 7, 2016. https://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science.

On the one hand, the idea is that all scientific results should be freely available, from now on to 2020. On the other side, the word is not that it should be available, but “as open as possible” and “as closed as necessary”; which means that some parts of research will be open science and closed science, as before, will be some parts of the research.
So this concerns publication, with all the problems we have with the publishers and the green and gold open access, etc. But, for us here, the second important point is about data, not only about publication.
In my mind, I can understand the separation. These two points are related, not only because results are on the one hand the publication and on the other hand the data, but also because of the economic interest. For the dissemination, the publishers are ready for that. So, it is not only open science between scientists. I think scientists don’t really need this because we already work together in infrastructure and we have access to our own data. Today, the governments put the focus on societal impact, on dissemination of research results not only for citizens (transparency), but also and above all for the corporate sector (innovation, value creation).
Governments are above all concerned by Ebola, Zika or climate change, and they try to improve and accelerate the production and dissemination of research - scientists are expected to be more performant, more efficient, work more quickly and disseminate immediately their results to those who can transform this into product, drugs against Ebola, Zika vaccination etc. I think we should keep this in mind when we are speaking about open science. It is not only about sharing and people get friendly and work together; there are economic and societal interests well beyond the needs and challenges of the research communities themselves.

References

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.

Prost, H., Malleret, C., Schöpfel, J., 2015. Hidden treasures. Opening data in PhD dissertations in social sciences and humanities. Journal of Librarianship and Scholarly Communic MPation 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., Prost, H., Malleret, C., 2015. 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., Juznic, P., Prost, H., Malleret, C., Cesarek, A., Koler-Povh, T., 2015. Dissertations and data (keynote address). In: GL17 International Conference on Grey Literature, 1-2 December 2015, Amsterdam.
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., Chaudiron, S., Jacquemin, B., 2016. Dissertations as data. In: ETD2016, Lille 11-13 July 2016.

Contact

Joachim Schöpfel, Lille University

Joachim Schöpfel is director of the French National Centre for the Reproduction of Theses (ANRT) and scientist in information and communication sciences at the University of Lille (France). From 1999 to 2008, he was head of the library and document delivery of INIST (CNRS). He holds a PhD in Psychology of the University of Hamburg (Germany) and has signed several publications and communications on scientific information, documentation and job development, see CiteULike and the French national LIS repository. He is member of the editorial board, peer reviewer and evaluator of different journals, collections and organizations. His research interests are related to open access, grey literature, ETDs, open data, scientific communication and library development. He is member of the GERiiCO laboratory on information and communication sciences (Lille), the Council for Documentary Information of the Free University of Brussels, the International Advisory Board of the Project COUNTER.

Linkedin profile: https://www.linkedin.com/in/schopfel
Twitter: @schopfel
Email: joachim.schopfel@univ-lille3.fr