DARIAH-EU

DARIAH Winter School in Prague

Open Data Citation for Social Sciences and Humanities

24th to 28 of October 2016

Full conference synthesis

The DARIAH Winter School was organised as part of DARIAH's Humanities at Scale (HaS). It received the active support from: EU's Horizon 2020 Research and Innovation Program under Grant Agreement No 675570. It has been organised by:

- Huma-Num
- OpenEdition
- Charles University

Scientific committee

- Marjorie Burghart, CNRS
- Lucie Doležalová, Charles University
- Nicolas Larrousse, Huma-Num
- Mike Mertens, DARIAH
- Pierre Mounier, OpenEdition - EHESS

Participants

- Free Tessa, Netwerk Oorlogsbronnen (Network War Resources), Netherlands
- Gkioulekas Panagiotis, Alexander Technological Educational Institute (ATEITH), Greece
- Görögh Edit, University of Göttingen, Germany
- Gura Caitlin, Austrian Centre for Digital Humanities, Austria
- Kalin Cihan, Istanbul University Central Library, Turkey
- Kovalendo Kira, Institute for Linguistic Studies, Russian Academy of Sciences, Russian Federation
- Lemaire Sonia, OpenEdition, France
- Malínek Vojte□h, Institute of Czech Literature of the Czech Academy of Sciences, Czech Republic
- Maryl Maciej, Digital Humanities Centre at the Institute of Literary Research, Polish Academy of Science, Poland
- Messina Adele Valeria, University of Calabria, Italy
- Nashed Michael, Bibliotheca Alexandrina (Library of Alexandria), Egypt
- Orban de Xivry Mathieu, OpenEdition, France
- Paavolainen Maija, Helsinki University Library, Finland
- Papaki Eliza, Digital Curation Unit, ATHENA R.C., Greece
- Papastamkou Sofia, Maison européenne des sciences de l'homme et de la société, France
- Popovic□Petar, National library of Serbia, Serbia
- Reed Zsuzsanna, Central European University Budapest, Hungary
- Simon Zsolt, Eötvös Loránd University of Budapest, Hungary
- Starczewski Michal, University of Warsaw, Poland
- Tavakkoli Amirpasha, EHESS, France
- Treanor Mairead, Met Éireann (Irish Meteorological Service), Ireland

Coordination

Nathanaël Cretin, OpenEdition.

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

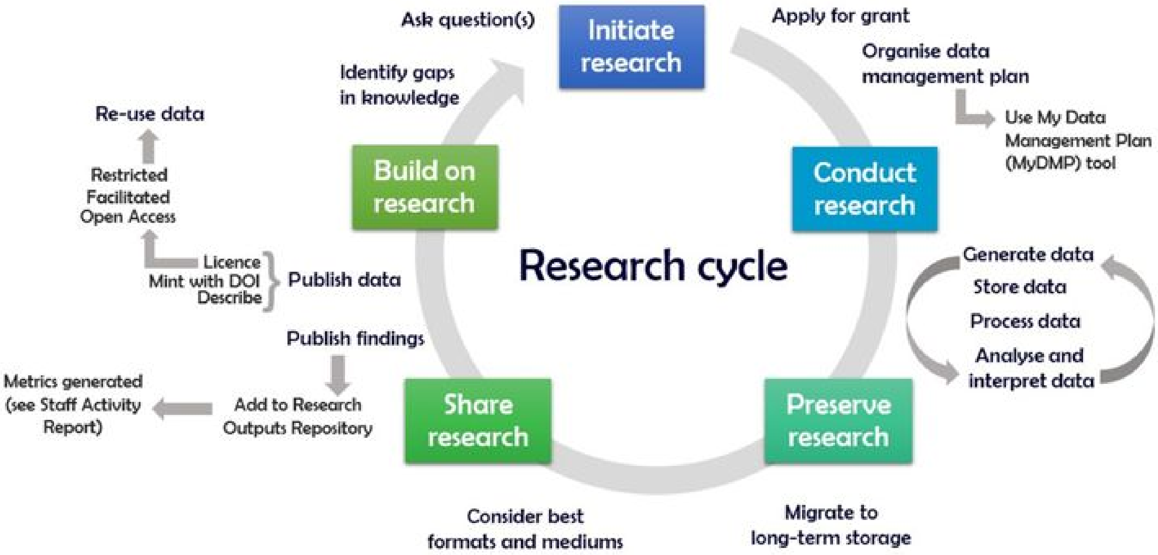
- Research methods as approach to make different levels of data: Observational data, Experimental data, Simulation data, Derived or compiled data.
- Input and output: For my own work, especially with PhD dissertation, but also with articles and report, there is an important distinction between input 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

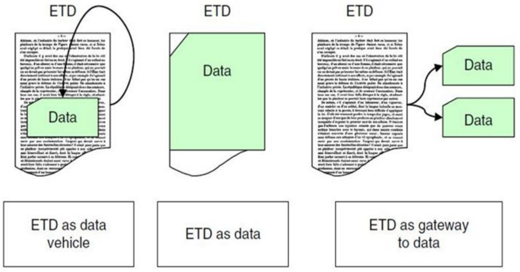

Research cycle
Initiate research
Ask question(s)
Identify gaps in knowledge
Apply for grant
Organise data management plan
Use My Data Management Plan (MyDMP) tool
Conduct research
Generate data
Store data
Process data
Analyse and interpret data
Preserve research
Migrate to long-term storage
Share research
Consider best formats and mediums
Publish findings
Add to Research Outputs Repository
Metrics generated (see Staff Activity Report)
Publish data
Licence
Mint with DOI
Describe
Restricted
Facilitated
Open Access
Re-use data
Build on research

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.

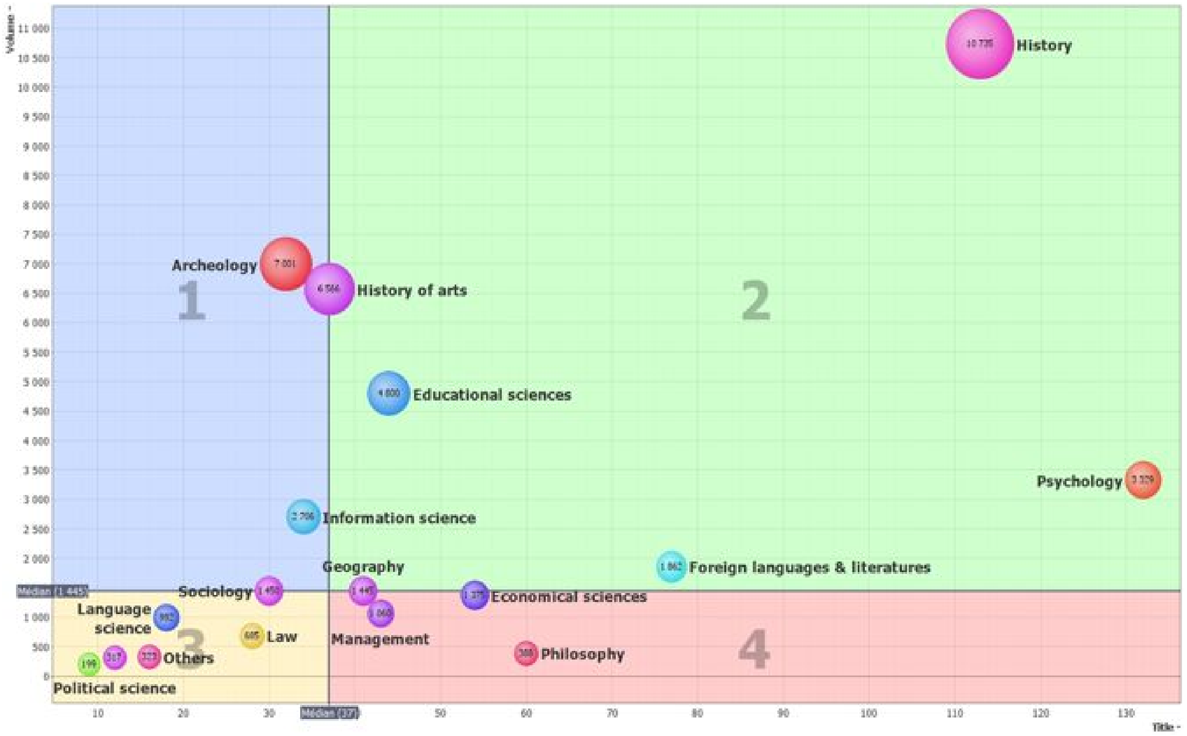
Data in dissertations

=> 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).

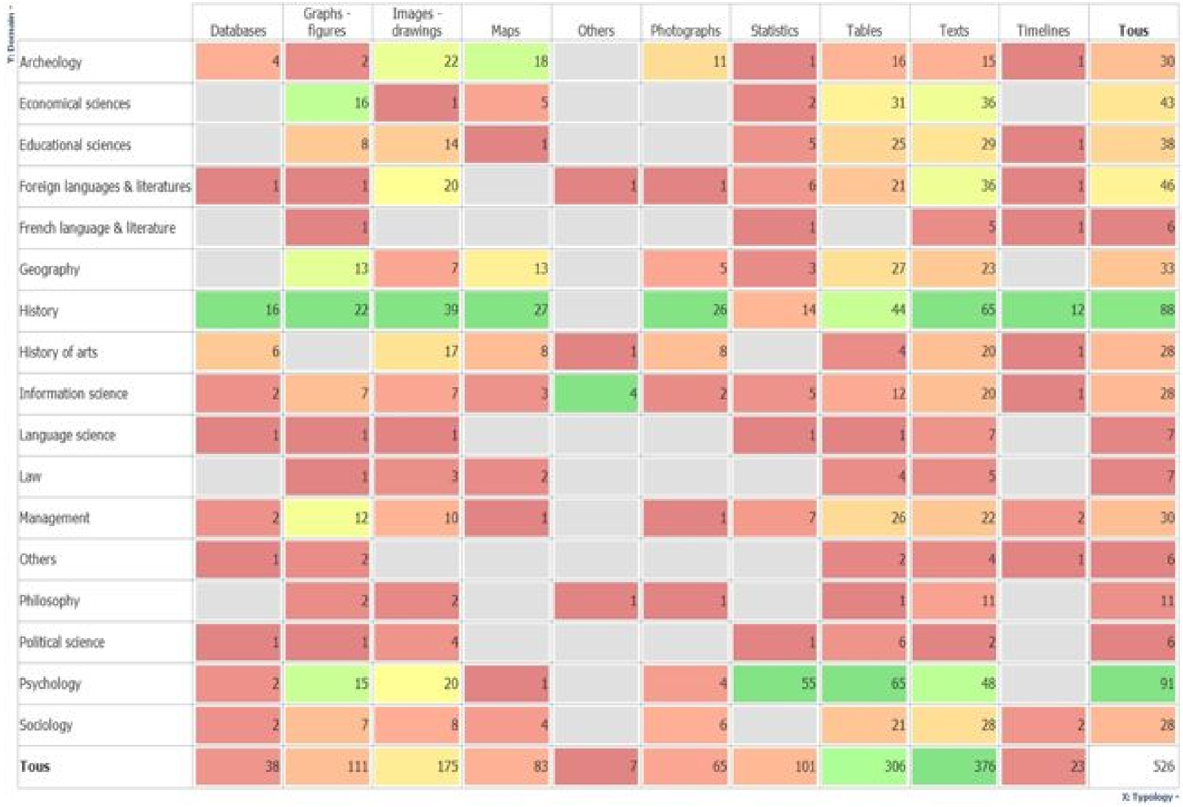


Disciplines and categories of data

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: Domain -	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

2-Open Critical Edition. The Missing Link Between Digital Humanities and Open Science

Marjorie Burghart (CNRS) is a medievalist and Digital Humanist, and Emmanuelle Morlock (CNRS) is a Digital Humanist and a research officer specialised in information architecture, and Research Data Management.
This session will consist in a general introduction to digital editions followed by a practical presentation of TEI. We tried to separate the topics and to have at the same time a complementary approach.

What is a Digital edition: Some Interesting Examples

- We can start with this Google Book. It is a scanned volume from a famous series of 19th century books, the *Patrologia Latina*, a major collection of latin works. These volumes are now available and 200 of them are on Google Books.

But is this is a *digital edition*?
=> Some answers: It is digital and it is an edition. Or it can be used as digital edition by scholars.

Is it used as an *edition in digital form*?
=> Different types of users have different definitions, but there are more and more strict criteria. People speak about *digitised edition* for such type of material. This is an *edition*, *digitised* (it became digital) and it is a little bit *critical*. But this is not what most people today define as a digital edition. There is more and more reflection on this topic. See for instance RIDE: A review journal for digital editions and resources, from a German center based in Cologne. This journal produces a *review of digital edition projects* with great criteria for reviewing. It shows the state of the art criteria for the best practices in digital editions. One of them is the distinction between *digitised edition* that is just a scanned book put online and the *native critical edition* which is meant to take advantage of all the perks of the Internet connected data. There are a lot of criteria to achieve but it is more of an ideal to reach, to tend towards, that can be used as a benchmark for projects.

- Another example is this text from the Corpus Corporum: It is in fact the same text as the digitised edition we have just seen on Google Books, but presented in a searchable corpus of Latin text.

Is it a *digital* or a *digitised edition*?

=> It is actually debatable because it is only the text of the 19th century edition which has been OCRed and put online, there is no extra work. So it is a *print-born edition* which has been *digitised* into text mode, *structured* and put online, but there is no added value beyond that, except the fact that it is easier to search and that you can reference it.

- Edition of the poems of Anne Finch

It is the digital archive of her work, a very nice work: the layout is really pleasant to read, there are a lot of features, you can access all versions of the poems, you can access the sources material, images, etc.

Is it a *digital* or a *digitised edition*?

=> This is a *digital edition*: it is born-digital, it takes advantage of more than just full text because you have different media which are linked, etc. but I think that if you run the RIDE criteria on this archive, it would not get the highest score, because it is a limited in the way you can take advantage of all the data that have been gathered. For instance, I have performed a search query on the corpus and all I get is some kind of interface. When you do a Google citation search on a website, it is the same thing, it just searches for a string of characters, you have no combination. So it is very basic in the way you can take advantage of the material. You cannot download the source work if for instance you wanted to integrate these poems into a corpus of poetry from the same period. It is a very valid scholarly work but it is "*self-contained*", not connected with the outer world. It is digital but not open yet. It is open access: you can access it, everyone can access it, there are no barriers; but it is not yet open data: you cannot access the data underlying the scholarly work, reuse it and connect it to different things. This is a step further.

- Another kind of edition: Map of London

It represents a 17th century map of London which has been edited just as a text can be edited with interesting features: for example, you can highlight all the churches with an overlay layout on the map and you can zoom until reaching a single building, for example St Paul's Cathedral. You also have a text explaining what it is and you have a list of all documents in which St Paul cathedral is mentioned.
So you have, around the map, a library of edited documents with links between the map and the documents. Here you have the step further: linked data within the sub-corpora of the website, and you also have references of place from and to you can link from other projects. Here we enter the world of connected data: when you click on a link, you have the transcription of the document in which you find the highlighted term (that brought you here).

- Plaoul Commentary

It is an edition, the interface where you can read the scholarly digital edition. The author is Jeffrey C. Witt, a US medievalist. He has a vision: he sees critical editing as building a huge database with assertions on the editions and a lot of annotations. So he modeled precisely all these pieces of information and annotations and on top of that he builds printed editions, but also a workspace for an edition. It is a real complete environment and it is open because you can see the corpora and view the underlying data on Github. It is also a big database of precise data annotation with a service to query in this huge database: you can see relations and properties, in a formalised way. And that's not all, as you can also see the images of the manuscripts, it is important to note that he didn't digitise the manuscripts himself, and in his editions, he didn't have to keep a copy of the images on a server. With the properties of linked data, he just accessed the manuscripts images that are published as linked data by the institution that keeps the manuscripts. There is a protocol for images that is now widely used (IIIF). It is a new model that allows you to build your edition, your transcription, your view of this text on top of some images that you don't curate at all. There is a visualiser and you can also built, as a researcher, a critical editor. If the images are in different places, you can build on top of it your workspace, your interface, just with links. As previously said, this is a great workspace with statistical tools, for example we can have the frequency of use of biblical quotations.

Group exercise with a poem

North of Everywhere, Helen Mort:
http://www.manifold.group.shef.ac.uk/issue7/HelenMort7.html
=> Goal: Think how to approach this document if you had to make an edition of it, from your background: what would you consider important to underline, to be able to share it with other people so they understand the document and take benefit from it, with of course a particular attention at what strategies to open the data.

Group presentation & paperboard
=> Synthesis: Groups had different approaches but a lot can be connected together. What the text is and how it does function in itself, with context and metadata? Some others had already in mind how the edition will operate in a broader system with API, as a kind of technical functioning. We heard also about intertextuality and expression of the relations with other words. What is interesting is that at the beginning of the analysis, you have to take into account the ecosystem in which you will publish and what you want to do with it (maps, etc.).
The context of your aim influences the decision about the representation. The question is can we represent all that in a practical way? Of course we can, but in the economy of a project you have limitations (money, resources, time). So you will have to list all the possible features and make choices. You can have an Interface, a displaying device, on top of digital data organised as a system.

Text Encoding Initiative

How can the Text Encoding Initiative help to prepare digital editions and encode text? Critical editions are an important part of Digital Humanities and the TEI allows to encode a text and take advantage of this encoding with an attention to open data.

Critical edition

Digital Humanities are everywhere: you can practice them by making bibliographic searches on databases, on Google, etc. You can search manuscripts, read books in digital libraries, you can generate reports or use corpora to identify the sources of a text. You can use computer assisted collation or stemmatics. Collation is the process of collecting all the witnesses, manuscripts or editions of the work you are editing and to compare them to see how the text differs, its variance. Once the collation is created, you have to try to determine what is the "genealogical tree" of the witnesses of the work, to try and see which one has been copied on which one. This is called a stemma.

TEI

You can also use digital tools to structure and analyse the edited text with TEI as it is a common frame to analyse and structure text, especially text from the Humanities, from historical and linguistic sources.
The first reason to use TEI is that you have the TEI guidelines, you can share something that a lot of scholars used for the past 40 years, it is "battle-tested". It is not something you can think out yourself and decide it is ok for everyone else, you have to discuss encoding options with many different types of scholars from different fields to reach an agreement. The TEI helped to find common ground from different fields and scholars around the world, in order to share a same model of information for text. Besides, TEI makes it easy to differentiate between the aspect of a book and its analysis. This aspect is important, specifically if you are working with ancient documents, medieval or epigraphic, but also with contemporary digital-born documents. It allows to report that there are for example three lines in a particular place in a document. And it is also possible to add references to "Isabelle" for example. It is important to hit both sides of the document: the physical aspect but also the meaning. Finally, it is a good way to be completely free from proprietary formats (pdf, word). Otherwise you are completely tied to the format used for your file and you have no warranty that in the long term it will be preservable.

The Text Encoding Initiative is:

- Human-friendly rules for modeling the text: TEI Guidelines. In printed version, it is more than 1 000 pages because it covers a huge range of texts - you don't have to know everything if you want to do a particular work.
- Computer-friendly way to implement the rules of the Guidelines through an XML schema. Guidelines are written for humans and the schema applies the same rules as developed in the guidelines, but for computer programs.
- Community of users that can provide support in different ways. It can be advices, discussions about your own issues and also software that has been prepared for other projects but in a generic enough way to be useful to others, sharing the same formalism, the same TEI modeling. It saves time and gives a better insurance for quality of reflection.

XML

In a way, it is really close to html. If you look at the code of a webpage, you can see tags, etc. XML is basically the same principle, you have tags, except that html is a closed vocabulary and xml is not, it is extensible. The rules are stricter than in html, it has to be a tree structure. TEI XML has the advantage of full text plus a database, so you don't have to choose between transcribing on one side and creating a database on the other side. You can have both together if the data analysis links into the text. It permits to retrieve text mentioning data and vice versa, access data pertaining to the text.

For medieval writing

Diplomatic edition: it follows strictly the aspect of documents, for example you do not expand abbreviations, you respect the layout of the document, etc. Otherwise you have the Transcription for research purpose, where you can expand the abbreviations for a better readability.
Here you can have both, a versatile document, a record of all the data about the aspect of the document, the diplomatic view; and a record of all the analytic data. It is possible to have two views of a document, one is a diplomatic view where you can see which words were abbreviated and what is the expanded form; and another view where you can see what are the sections of the document, like chapters, that strictly define certain parts. It is interesting for researchers to be able to search and extract different types of parts from a corpus.
Here, the reader also has options! Classically, the editor makes all decisions once and for all. Now you can have a system allowing users to choose their options. They might be interested in mixing different kinds of visualisation with scripts that produce a webpage that readers can use.

Inside TEI

The key idea is that it is not just TEI or just XML, it is a family, a constellation of technologies that work together to work some magic in the end. It starts with XML that has the role of *representing* the text. With XML you describe your data with *tags* that can be qualified with *attributes* and you have to produce a *tree structure*: one *root* for your document and this root has several *children* which may also have children. This is the only golden rule of XML.
An example of XML source:
It begins with a declaration and then starts with the XML itself. The root of the tree structure is <text> and the children of this root are <p>.

```
<?xml version="1.0" encoding="UTF-8"?>
<text>
  <p n="1">I am reading a book by  <persName>Jack London</persName></p>
  <p n="2">I live in  <placeName>London</placeName></p>
</text>
```

- Controlling the text: TEI schema

The TEI provides the rules for structuring the document beyond the rules of XML. This schema is the implementation of the TEI guidelines, from a human-readable version. There are several TEI XML schemas and people can create sub-schemas based on the TEI, but only using a sub-part of the TEI. They can extend the TEI if they want, for example, to make an edition of a musical text, they need another format description of the music model. The model can be extended or reduced.

- Displaying the text: CSS and XSLT

You can display what you have encoded and transform it into something else. CSS is a web language that you apply to XML pages. XSLT is more developed and powerful, it can transform data in a web page for example or in a different type of XML, or extract and transform data from your XML into RDF or JSON and share it.

- Querying the Text: XQuery and XPath

Defined by W3C recommendations: XPath is commonly used within XSLT and XQuery is more complex and more powerful as it allows to query data and structure together, so you can extract all the words in a particular part of the document, for example.

Open science

In fact, the main goal is to prepare data before publishing it in a way that machine can understand.
=> Opening the principles of open access, of openness to the whole cycle of research.
To explain this, we have to see some principles: *semantic web* and *Linked Open Data*.
Then we will come back to TEI to see how to interconnect TEI files with this web of data that are linked and exposed by machine.

- Giving access is not sufficient to research data and publication.
- *Open Access*: Free and persistent access to research data and publications.
- With *Open Access*, it is more about an access for the reader. So when you have a huge volume of information, how can you read it?
- *Open Data:* Files made publicly available by official organisms for re-use.
- *Open Process*: Right to openly observe the underlying data and workflows of research project.
- Openness also influences research as way of improvement as the underlying data are accessible. As we saw, if we just show you the result of an edition, you don't really understand what is at stake, what is the work of interpretation that has been done. In order to validate the scientific work on an edition, you also have to look in the underlying data. The workflows are also very important to be documented.
- *Open Science*: Free and persistent access to research data with the right to observe openly these data with digital tools.

=> Open Science = Open Access + Open Process

It means that it is not only the readers that can access the research and analysis but also machines. To do that you need data to be expressed in a particular way.
The difference with TEI and the schema is to know the meaning of the tags, a machine can parse it and build an interface, but the machine has to know the schema. And it is not always the case because the schema is inside the edition, even if we have standardisation, it is not enough to have a broader interpretation of the data. TEI allows to have data and with our interpretation.

Semantic Web

The *semantic web* is like a parallel web that differs from the original web by the kind of knowledge presented and accessed.
The knowledge found on the semantic web is *formal* knowledge with:

- a machine readable notation

- a formal syntax
- a formal semantics with inference mechanisms

The Semantic Web started as a vision by Tim Berners-Lee and became true via *Linked Data*.

=> Open Data + Linked Data = Linked Open Data (LOD)

The idea is to share machine-readable and interlinked data that are on the web with two aspects:

- A language aspect: how to interpret data that are in documents or in web pages
- Interoperability aspect: how to understand all this without referring to a schema

So it is a system to identify resources where everything is a resource.

Linked data

Design principles for sharing machine-readable interlinked data on the Web:

- Name resources with unique identifiers (URIs)
- Use the architecture of the web to get some information about theses resources (http)
- Use a standard model to give information about these resources (RDF)

RDF: Resource Description Framework

It expresses information about identified resources with very simple sentences and composed of three elements:

- a subject: identifying the resource
- a predicate: identifying a property of the subject
- an object: identifying the resource linked to the subject by the property

Ex. Helen Mort (subject) --- is the author of (predicate) --- the poem “North of everywhere” (object)

The result of the aggregation of triples is a graph and the specificity of this information model is that:

- relations are part of the data
- each triple is autonomous, complete, persistent
- a distributed model

TEI to LOD

TEI explicates the data but not exactly the relations. The relations expressed in the hierarchy.

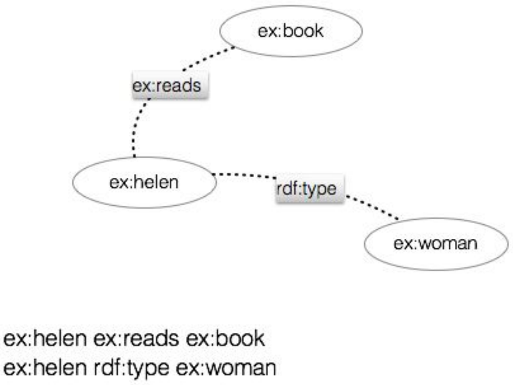
In the metadata, you have the title statement (titleStmt) and an author with a reference to the URI of the dbpedia page of Helen Mort. Dbpedia is the database made with wikipedia articles and facts extracted and transformed into a database which is accessible by humans and machines:

```
<titleStmt>
        <title>North of Everywhere</title>
        <author ref="http://dbpedia.org/resource/Helen_Mort">Helen Mort</author>
</titleStmt>
```

It is possible to extract the triples: the text represented in the TEI document has a title; the title of the text is “North of everywhere”, Helen Mort is the author of the text.. You have to select what could be interesting for others and express it in the RDF formal language to expose it and to make it available. You can also have “Hermaness” as the English (attribute) name of a place; this place is identied by the URI http://dbpedia.org/page/Hermaness, the longitude of this place is “60.837222”.
=> The sum of the triples produces a graph and the “magic is also done by the XSLT”

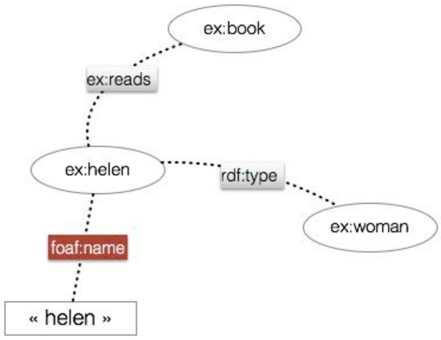
Step of conceptualisation: it is a point of view on the reality, ex: two resources: Helen (a woman) and a book; relation: she reads the book.
Step of language selection: where the URIs comes into play; pairs of resources are connected by the binary relation they belong in: ex:helen ex:reads ex:books; Unitary relations are connected to a class: ex:helen rdf:type ex:woman

A set of RDF triples is a graph



ex:helen ex:reads ex:book
ex:helen rdf:type ex:woman

Literals to associate a natural language fragment to a resource



Linked Open Vocabulary
Foaf (friend of a friend): It is a vocabulary that gives a property to the described relations between person. There is common vocabulary to prepare possible relations that some users will then activate, the catalogue of open vocabulary: https://lov.okfn.org/dataset/lov/. On Helen Mort's page on dbpedia, there are information on triple you can find with for example: sameAs. You can find further information with the Linked Ancient World Data Institute.

DBpedia Demo: Basic exploration of a RDF graph with simple SPARQL queries

Two simple sentences or assertions:

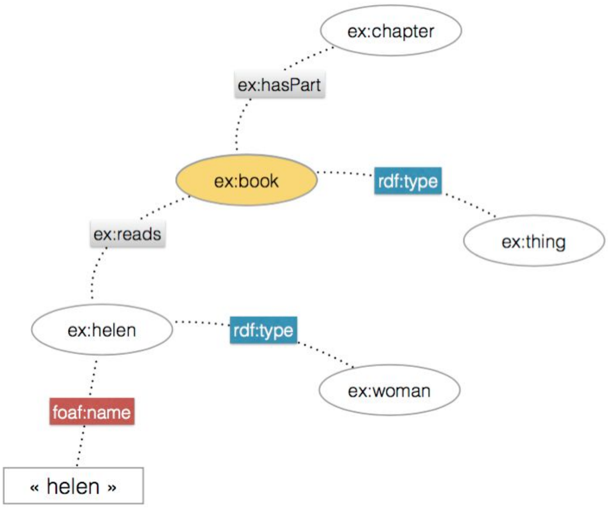
- "Helen" "reads" "a book"
- "Helen" "is" "a woman"

In RDF, with the prefix "ex:" we have our identifier:

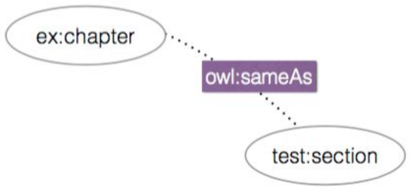
- ex:helen ex:reads ex:book
- ex:helen rdf:type ex:woman

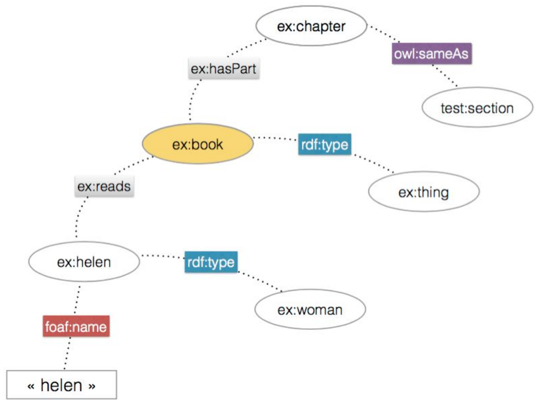
=> subject, object and the relation

It is here a unitary relation, this means that it is the class of the resource. The three elements are resources.



Here you can express that it is the same thing and make the aggregation function. This is why in linked data publishing practices, it is highly recommended to be generous and to try to align with "sameas" as much as possible your data.





SPARQL with DBPedia

DBPedia is the RDF graph built extracting all the data that is curated in Wikipedia. When you are human you see an html page, when you are a machine you see a RDF file for the same information. There is, I guess, a duplication of the database that you can directly query with a dedicated interface with the SPARQL language.

Simple Protocol And Query Language A query has a structure:

- SELECT distinct * == select all resources
- WHERE { } == the query
- LIMIT, GROUP BY, ORDER BY...

Simple queries 1

Find resource with the English label « Prague »
Answer:
http://dbpedia.org/resource/Category:Prague ===> http://dbpedia.org/page/Category:Prague
http://dbpedia.org/resource/Prague ===> http://dbpedia.org/page/Prague
With this you find the name of the resource and can use it for further queries.
Find all the properties of this resource
Find the types of this resource
Choose a type (ex. "?o"" for object)
Find the resources with the type

Simple queries 2

select distinct * where {?s rdfs:label "Prague"@en} LIMIT 100
select distinct * where {<http://dbpedia.org/resource/Prague> ?p ?o} LIMIT 100
select distinct * where {<http://dbpedia.org/resource/Prague>rdf:type ?o} LIMIT 100

select distinct * where {?s rdf:type <http://dbpedia.org/ontology/PopulatedPlace>} LIMIT 1000

This is a good to explore material when the relations are precisely defined. And you can start building a database without having in mind the whole schema, it can be flexible and adapted.

RDF is much more fact oriented and TEI is more precise to express subtle documents and gather a lot of precise annotations and distinctions. But they can work well in collaboration.

One of the key stakes of Linked Open Data is the quality of data and TEI is really good, it can be like a database where you keep all your scientific information and then extract some datasets in RDF or other language, in a continuous work of repackaging your data for different purposes.

CORESE

- Simple inference in action with Corese, a Semantic Web Factory (triple store & SPARQL endpoint) implementing RDF, RDFS, SPARQL 1.1 Query & Update, developed by INRIA: http://wimmics.inria.fr/corese
- Tutorial: http://wimmics.inria.fr/node/34
- Linked Data Navigator using Corese and SPARQL Template Transformation Language: https://corese.inria.fr/

What is the best way to share a good body of generated RDF?

EM: The best way is to find the appropriate data repository, one that is certified (Data Seal of Approval) and expose it here. If you want people to actually use it, I would do a **data paper** explaining concisely where the data come from and the context (technical but not only) that other researchers would need to use it for another research. All things that are obvious must be explicated. Like this, you delegate the stewardship of the dataset and you give all information and associated metadata.

Conclusion

Knowing that it will influence the way you prepare you text with TEI and at the same time, it opens to the notion that these **triples are not a technical thing, it is an editorial thing**. You have to decide which are the **interesting triples in a text and for a community**. **This is a new task of the publisher: to design. If you consider publisher or editor as a designer of information artefact, these RDF exposition of data must be editorialised and designed.**

Useful links

- Dbpedia example: http://dbpedia.org/resource/Helen_Mort
- DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia: http://svn.aksw.org/papers/2013/SWJ_DBpedia/public.pdf
- https://lov.okfn.org/dataset/lov/
- http://www.foaf-project.org/

- http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=terms
- Using SPARQL to access Linked Open Data from SSH perspective: http://programminghistorian.org/lessons/graph-databases-and-SPARQL#searching-rdf-with-sparql
- To see more refined uses of sparql queries in combination with nice displays for the result, watch that youtube video about wikidata (16 min): https://www.youtube.com/watch?v=1jHoUkj_mKw

Contact

Marjorie Burghart & Emmanuëlle Morlock, CNRS

Marjorie Burghart is a research fellow at the CNRS (French National Center of Scientific Research) and she is working in the CIHAM UMR 5648 research center in Lyon and is specialised in medieval history and computer science. She is an elected member of the board of directors of the *Text Encoding Initiative* (TEI) consortium, the scientist in charge for the EHESS partner of the Erasmus SP+ *Digital Edition of Medieval Manuscripts*, and also the scientist in charge for the EHESS partner of the DIXIT (*Digital Scholarly Editions Initial Training Network*) Marie Curie european project. She has published several papers and softwares, and is involved in differents projects of electronic edition of medieval documents in TEI format.

Marjorie Burghart's website: http://marjorie.burghart.online.fr/?q=en

Email: marjorie.burghart@gmail.com

Emmanuelle Morlock is a digital humanities research officer at the French National Center for Scientific Research (CNRS) and currently works at HiSoMA, a research center dedicated to Archaeology and Philology of the Ancient Worlds. Her main mission is to assist researchers in their application of information technologies and solutions for scholarly editions of ancient texts and inscriptions. Her activities include project ownership assistance and technical implementation of online publications (metadata modeling, definition of encoding strategies, TEI framework implementation, information architecture and digital curation of research data). She was educated in France where she studied French literature and received a Master's Degree in Information Science and Documentation from SciencePo Paris.

Twitter: @emma_morlock

Email: emmanuelle.morlock@gmail.com

3-Data Management Plan

Marie Puren & Charles Riondet, INRIA

Data management can offer many advantages, like higher quality data, increased visibility and better citation rate. In this approach, research data is an asset and a resource that can be shared with mutual benefits for the person who share the data and the person who collect the data. Yet, the Open Science movement implies radical changes for many researchers.

What is research data?

We can find a simple definition on the website of the Boston University Libraries: “Data are distinct pieces of information, usually formatted in a special way”. But it is difficult to clearly define “research data”, because research data is challenging:

- there is no consensus on the definition;
- it varies according the discipline;
- it differs according to the research funder.

The University of Bristol define “research data” as follows: “Research data is created as a direct result of ‘doing research’. It excludes teaching materials and administrative documents (such as job descriptions, emails or financial reports). Research data comes in an endless variety of formats”. For the Boston University Libraries, research data is "data that is collected, observed, or created, for purposes of analysis to produce original research results”. These data can be: observational, experimental, generated from test models (simulation), derived or compiled (like text and data mining), reference or canonical (for instance, gene sequence data banks). Therefore, research data can adopt multiple forms like: text or Word documents, spreadsheets, laboratory notebooks, questionnaires, videotapes, photographs, slides, samples, databases, methodologies, output for analysis software, standards, etc.

According to these definitions, “research data” could be defined as:

- data that help to do research;
- data that could be collected, created and analysed;
- data that come in multiple formats.

The term "dataset" is used to describe a collection of research data: “A digital dataset might comprise a single element [...] [or] a collection of related elements” (Oxford Research Data Website). Thus, a dataset is a compilation of research data. It could gather together data in a single document - like in a CSV (Comma-Separated-Values) for instance - or a series of data.

A new model of openness for research data

The movement for Open Science promotes a new model of openness, with an important impact on research data - in particular on data sharing. The main aspects of openness are: availability and access, reuse and redistribution, and universal participation. This new model tries to gradually replace traditional ways of thinking in the international research community.

Further information:

- Andreas E. Neuhold, work based on “The taxonomy tree”, FOSTER (Facilitate Open Science Training for European Research)
- Michael Nielsen, *Reinventing Discovery: The New Era of Networked Science,* Princeton University Press, 2011.

Generally, we consider that Open Science rests on six main pillars:

- Open Data
- Open Source
- Open Methodology
- Open Peer Review
- Open Access
- Open Educational Resources

The Open Data movement fosters initiatives to open data, which means that “anyone can freely access, use, modify, and share for any purpose” (Open Knowledge International, “The Open definition“). In this new framework, “It has become increasingly apparent that scientific data should be considered a product in much the same way journal articles or conference proceedings are [...].“ Felicia LeClere, “Too Many Researchers Are Reluctant to Share Their Data”, *The Chronicle of Higher Education*, 2010.

Supported by European and national initiatives

In 2013, the Pilot on Open Research Data (ORD Pilot) announced the European engagement to open research data in the *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020* as part of the Horizon 2020 Research and Innovation Programme. The Pilot “aims to improve and maximise access to and reuse of research data generated by projects for the benefit of society and the economy”. Two types of data are concerned:

- data needed to validate results in scientific publications;
- any data considered valuable by the project.

These data have to be made available for other researchers, industries and citizens. Nevertheless, if the research will be jeopardized or if intellectual property and personal data will be threatened by making data open, projects are allowed to opt out.

So, “the ORD pilot applies primarily to the data needed to validate the results presented in scientific publications. Other data can also be provided by the beneficiaries on a voluntary basis, as stated in their Data Management Plans.“ (*H2020 Programme Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0, 26 July 2016, p.3.)

In July 2016, the ORD Pilot has been extended to cover all Horizon 2020 funded projects. In this updated version of the *Horizon 2020 Programme Guidelines*, open access becomes the default setting for research data generated in Horizon 2020 (*H2020 Programme Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*, Version 3.0, 26 July 2016, p.8). This European program is particularly inspired by Anglo-Saxon policies. For instance, American funding agencies like the *National Institutes* *of Health* (NIH) since 2003 and the *National Science Foundation* (NSF) since 2010, fostered by the American Government, impose to their funded researchers to subscribe to their research data sharing policies on open access. In United Kingdom, most of the funding agencies require that funded researchers made their research data openly available, with the help of dedicated structures such as the *Joint Information Systems Committee* (JISC) and its *Digital Curation Center* (DCC), dedicated to data management in the UK.

Direct benefits for researchers

This new model of openness aims to offer new resources that can be exploited by economy and by research. Sharing research data provides direct benefits to researchers, but, some of them are still reluctant to share them. As Felicia LeClere stated, "Data sharing is a bit like going to the dentist. We can all agree that it is a good thing to do and intrinsic to good scientific practice. In reality, however, researchers tend to view data sharing with a mix of fear, contempt, and dread" (*The Chronicle of Higher Education*, 2010). Fortunately, the situation is gradually evolving, but sharing (or not) rests most of the time on the shoulders of the researchers. Researchers need to be clearly aware of the benefits of sharing their research data:

- It fulfills requirements of:
    - Funders
    - Journals
- It increases research impact and visibility
    - By getting credit for research outputs
    - By boosting citation rate
- It saves time
    - By planning ahead the research
    - By being more efficient (data and methods already explained)
- It preserves data
    - By depositing in a repository
    - By making accessible unpublished data with a citable links. Videos, posters, full methods can be published and used with full citable links via permanent DOI
- It ensures higher quality data
    - Maintaining data integrity
    - Managing and documenting data throughout its life cycle will allow you and others to understand and use your data in the future.
- It promotes innovation and potential new data uses
    - By creating new collaborations between data users and data creators
    - By encouraging new research in a field.
- It maximises transparency and accountability
    - By allowing scrutiny of research findings
    - By improving and validating research methods
    - By reducing fraud
- It supports Open Access
- It helps less rich institutions and countries to do research
- It makes good science and contribute to scientific progress

Funders and research institutions can also take advantage of this model of openness as well:

- Maximising return on investment
- Reducing the cost of duplicating data collection
- Getting access to great resources for education and training

Why manage data?

For yourself:

- Keep yourself organized
- Control the various versions of your data
- Systematically control the quality of your data
- Make backups to avoid data loss
- Format data for reuse (by yourself or others)
- Be prepared: document your data for your own recollection and reuse (by yourself or others).

For funders, data are valuable assets and they are expensive and time consuming to collect. Data should be managed to:

- Maximize the effective use and value of data and information assets
- Be assured that the quality of data is continually improved (data accuracy, integrity, integration, timeliness of data capture and presentation, relevance and usefulness)
- Ensure appropriate use of data and information
- Facilitate data sharing
- Ensure sustainability and accessibility for reuse in science

Research data management

Definition

> "Data management is integral to the process of conducting research."
>
> University of Leicester, "When do you need to think about RDM"

Data management has to be seen as the baseline of the research lifecycle. In this regard, it should be designed as early as possible and evolve all along the research project. This practice allows researchers to plan and decide how they will "collect, organise, manage, store, backup, preserve and share [...] data during a research project, and after the project is complete". Good research data management is the "key conduit leading to knowledge discovery and innovation, and to subsequent data and knowledge integration and reuse" (*Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0, 26 July 2016, p.3).

Research data management usually involves:

- creating a Data Management Plan (DMP) submitted along with a research funding application to explain how data will be managed both during and after a project
- storing research data safely throughout a project and sharing with authorised colleagues
- at the end of a project, cataloguing data and making them available in a long-term repository

Research data lifecycle

Research data management takes into account that data has a longer lifespan than the research project that creates them. The research data lifecycle integrates distinctive feature by separating the research process into stages and by taking into account the expanded lifetime of research data. In this approach, at each stage of the research process must be implemented specific research data management practices. Five essential steps compose the research data lifecycle and cover the best practices that you should follow during a research project:

- Plan and fund
- Collect and Analyse
- Preserve and store
- Publish and Share
- Discover and reuse

Research data management services

Today the increasingly collaborative nature of research invites to develop research data management services or RDM services. Researchers need to exchange data and want to use effective systems to store, access and share data securely. In United Kingdom for instance, the *Engineering and Physical Sciences Research Council* (EPSRC) *Framework on Research Data* has stimulated the development of RDM services within many UK higher education institutions. Higher Education Institutions have a key role to play in research data stewardship. These data are an asset for institutions, bringing benefits and impact for the institution as much as for the researcher. In order to reach these benefits, effective systems and support services need to be in place.

Developing an RDM service in an higher education institution implies to:

- Identify areas of responsibility for university management, support, administrative services and researchers
- Have a strategy to develop coherently
- Develop an institutional policy
- Create a long-term business plan with objectives, predicted costs, resource deployment and anticipated benefits
- Provide guidance, training and support to researchers
- Create guidelines to help researchers when they apply for grants and have to submit an outline data management plan
- Make space for storage (safeguarded and accessible)
- Assess your research data
- Create “good” metadata by supporting mechanisms for registering metadata
- Be aware of compliance with institutional, national and international policies
- Guarantee the respect of legal requirements on privacy and confidentiality
- Track the impact of research data with metrics

Creating Data Management Plans (or DMPs)

The *Guidelines on FAIR Data Management in Horizon 2020* provide a clear definition of Data Management Plans: “Data Management Plans (DMPs) are a key element of good data management. A DMP describes the data management life cycle for the data to be collected, processed and/or generated " (*H2020 Programme Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0, 26 July 2016, p.4).

A DMP is a formal document that outlines what you will do with your data both during and after your research project. It describes the data you expect to acquire or generate during the course of a research project, how you will manage, describe, analyze, and store those data, and what mechanisms you will use at the end of your project to share and preserve your data.

Several funders now make data sharing mandatory and applicants must provide a data management plan.

DMP key features:

- It is a regularly updated roadmap
- It is a standardised document
- Its content varies depending on projects’ requirements and funding agencies’ requests
- It focuses on data and datasets collected, created, analyzed by the research project.

It is a deliverable of the project, but not a “technical” document:

- It materializes the data policy of a project
- It sums up the goals and actions that will be implemented

- It meets funder's requirements

Digital data requires an "active management", it means:

- an ongoing maintenance (backup, migration, conversion, etc.) all along the data lifecycle
- an action plan in terms of data quality, technical feasibility, financial viability

In this context, data management is not data stewardship and means optimizing resources for a specific purpose. It allows ones to:

- Identify and make visible the actions to be conducted;
- Plan key stages, deadlines and critical time periods.

=> This active management goes hand in hand with "digital curation". It is an ongoing process that requires time and resources and consists in selecting, preserving, maintaining, collecting and archiving digital assets:

"Data curation activities enable data discovery and retrieval, maintain data quality, add value, and provide for reuse over time. This new field includes representation, archiving, authentication, management, preservation, retrieval, and use" (Digital Humanities Curation Guide).

FAIR Data

In January 2014, researchers, professional data publishers and funding agencies met upon the request of the Netherlands eScience Center and the Dutch Techcentre for the Life Sciences (DTL) at the Lorentz Center in Leiden. They agreed to support a minimal set of principles and practices: "data providers and data consumers - both machine and human - could more easily discover, access, interoperate, and sensibly reuse, with proper citation, the vast quantities of information being generated by contemporary data-intensive science". They are the FAIR principles. According to these principles, data should be:

- Findable with descriptive metadata and persistent identifiers
- Accessible in that it can be always obtained by machines and humans upon appropriate authorization, through a well-defined protocol
- Interoperable by using open formats, common standards, documented data specification and consistent vocabularies/ ontologies
- Re-usable with clear rights and appropriate licence.

The European commission endorsed the FAIR principles and wish to see them applied in H2020 funded projects (*H2020 Programme Guidelines on FAIR Data Management in Horizon 2020*, Version 3.0, 26 July 2016).

H2020 framework requirements

Projects funded under the Pilot on Open Research Data were required to produce a first version of a DMP as a deliverable during the first six months of the project. At the research proposal stage, all projects submitting to "Research and Innovation Actions" and "Innovation Actions" had to provide a short outline of their data management policy, evaluated under the "Impact" criterion. Since July 2016, a revised version of the 2017 work programme extends the Open Research Data pilot "to cover all the thematic areas of

Horizon 2020” requiring all the Horizon 2020 funded projects to provide a Data management Plan.
A template is provided in the Annex 1 of the *Guidelines on FAIR Data Management in Horizon 2020* (version of 26 July 2016). More detailed versions can then be submitted as additional deliverables at later stages of the project but also when any significant changes occur such as the generation of new data sets or changes in consortium agreements. The first DMP must fulfill minimal requirements:

- A description of data to be generated or collected
- The standards and metadata that will be used
- The data sharing or how datasets will be shared
- The archiving and preservation: procedures which will ensure the preservation of data, including backup and storage.

Data Management Plans : How-to guide

Components of a DMP

There are five main categories of information that should be included in a DMP:

- Information about the data and its format
- information about the metadata content and formats
- policies for access, sharing, and reuse of data
- long-term storage
- budget considerations for data management (salary time for data preparation and documentation, hardware and software requirements, etc.)

Crucial points to address

- Responsibility
- Results management
- Back up plan
- Intellectual property rights
- Becoming of the data after the project (hosting in a long-term perspective, access policies, etc.)

Responsibility

It has to be addressed for each step of the DMP => Outline the roles and responsibilities for all activities: data capture, metadata production, data quality control, storage and backup, data archiving & data sharing. Individuals should be named where possible. For collaborative projects the coordination of data management responsibilities across partners should be expressed clearly. Data management is not just the responsibility of the researcher who has created or collected the data, various parties are involved in the research process and may play a role. It is crucial that roles and responsibilities are assigned and not just presumed.
Researcher is the DMP coordinator, responsible for the data and its description, but there are other actors:

- computer engineer (hosting, security, infrastructural aspects)
- Archivist (broad sense): interlocutor for data selection, standards choices, mappings, durations and technical solutions
- research staff designing research, collecting, processing and analysing data
- laboratory or technical staff generating metadata and documentation
- database designer
- external contractors involved in data collection, data entry, transcribing, processing or analysis
- support staff managing and administering research and research funding, providing ethical review and assessing Intellectual Property rights
- institutional IT services providing data storage, security and backup services

- external data centres or web archives that facilitate data sharing

Results management

Data Collection

Two steps

- Document the data creation process or the data collection process for existing data, and the methods of data acquisition.
- Characterization of the data:
    - Raw or derived,
    - Purpose,
    - volume estimation,
    - type (quantitative, qualitative, survey data, experimental measurements, models, images, audiovisual data, samples, etc.)

Datasets management

For each dataset, the DMP should give minimal information:

- Reference & name (Identifier for the dataset)
- Description (metadata of your data):
    - description of the data that will be generated or collected
    - origin (if collected)
    - nature & scale
    - whether it underpins a scientific publication.
    - to whom it could be useful
    - existence of similar data and the possibilities for integration & reuse.
- Which formats/standards are used for this data?

The DMP should contain rules, like a file naming system or a filing plan and people involved in the research should comply to the normative information.

Description and metadata

Is the data understandable by an outside researcher?

The actual description of the data differs from the dataset management on the targeted audience. The audience of the latter was fellow researchers in a project, funders. The former audience is other researchers that will reuse your data.

Documentation and Metadata

Metadata is data documentation. It includes contextual details about data collection and any information that is important for using and understanding the data. A DMP should express if the metadata is:

- Machine/human readable
- Standardized: DublinCore, DataCite Metadata Schema, homemade format
- Automatically or manually captured
- Stored in databases, text files, or as headers in your files (Cf teiHeader)

- Created with controlled vocabularies or any internal conventions

Example: DataCite metadata standard

Datacite is a consortium of several libraries and research institutes that provide Persistent identifiers (DOIs) for research data and a metadata format to describe them:

- Identifier
- Creators
- Titles
- Publisher
- Publication Year
- Resource type
- Format
- Subjects
- Languages
- Version
- description

Formats

Open + Interoperable + Well spread in the research community => Standard

This is applicable both for your data and your metadata: using standardised and interchangeable or open lossless data formats ensures the long-term usability of data. For example, .csv and .txt are non-proprietary and are likely to be readable in the future, regardless of software availability.

But, researchers are strongly encouraged to use community standards to describe and structure data. To help researchers' finding their way in the data formats jungle, we are happy to announce the next release of the Parthenos Standardization survival kit developed by INRIA, as "a comprehensive online environment aiming at providing basic information, documentation and resources concerning standards applicable in a wide scope of digitally based humanities and cultural heritage research activities." The idea is to gather in one place useful information created by researchers for research project about sharing good practices, guidelines, pieces of code, as a single environment, as a part of the big Parthenos infrastructure which goal is to foster communication and collaborative work between digital humanities researchers.

Backup plan

The DMP must contain information about the storage conditions and the backup procedures of the data during the research. In particular, DMP readers should know how eventual incidents are anticipated.
Technical information are strongly recommended, like the frequency of backups, number of copies, crypting solutions, restoring plans, etc.
Of course, these questions are related to financial and human resources questions: server space costs must be evaluated. The responsibility of the tasks must be identified.
=> Storage = Budget + Anticipation

Selection and Preservation

Which data are of long-term value and should be retained, shared, and/or preserved? Some selection criteria:

- Anticipate the futures uses and reuses
- Legal or policy aspects
- Potential value
- Ratio cost/benefit

Datasets and the associated metadata, software and algorithms used might have to be preserved, for example, the European Code of Conduct for Research Integrity demands to archive primary and secondary data for a « substantial period » (European Science foundation, 2011). In general, any raw data should be kept with any data products that were particularly expensive or time consuming to obtain should be preserved. You should then find out archives or data centers that are commonly used in your discipline. Data centers usually last longer than lab or personal websites. Besides, your data management plan should describe what data transformations and formats need to be preserved to ensure future usability of your data. Finally, you should identify the person who will be responsible for maintaining contact information with the data center, it is especially important if there are restrictions on data use, for instance a requirement that potential users contact the data collector before reusing data.

Data repositories

The appropriate solution to store your research data is, in many cases, a data repository, which provides (in theory) sustainability. Many exists like EUDAT, Nakala, Re3data, Zenodo. Important criteria are:

- Data available for reuse (Harvesting, API, etc.)
- Citability
- Visibility
- Transparency
- Links to papers
- Preservation

It is possible to choose a repository according to quality criteria, certified by the Data seal of approval. First created by the Data archiving and Networked services (DANS), in the Netherlands, the certification process is now an international board (mostly European) that gives a seal to repositories based on quality criteria:

- The data can be found on the Internet
- The data are accessible (clear rights and licenses)
- The data are in a usable format
- The data are reliable
- The data are identified in a unique and persistent way so that they can be referred to

Data access and sharing

A major requirement for any DMP is the description of how data will be shared. As explained before, most of the funding institutions asking for a DMP have in the meantime a specific policy regarding data access and sharing. Therefore, the DMP should gather all information about:

- access procedures and policies
- embargo periods (if any)
- outlines of technical mechanisms for dissemination & necessary software and tools for reuse
- definition of access (widely open or restricted to specific groups)
- data sharing mechanisms (underlying data of a scientific paper, data paper, research data repository, project website, ...)
- if the dataset cannot be shared, it should be explained (ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related)
- Unique identification of the data and its producers : "Where possible, contributors should also be uniquely identifiable, and data uniquely attributable, through identifiers which are persistent, non-proprietary, open and interoperable (e.g. through leveraging existing sustainable initiatives such as ORCID for contributor identifiers and DataCite for data identifiers)." (Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020.)

Underlying issues are ethics and intellectual property rights, and the potential constraints for the data reuse must be identified as well: cite the right holders or the way to contact them, explain the ethical issues possibly encountered (consent, privacy, sensitive data), ... Note that it can have consequences on long time preservation: For example, patents data should be stored indefinitely.

Share research data with a data paper

Amongst the classical ways to share the data, an interesting one for researchers is the data paper, which means that the research data is editorialized and can be published in itself. A data paper is a scientific publication whose main goal is to describe a dataset or a group of datasets, more than analysis or research results and to give access to the described data. So it can be based on your DMP and makes your data more accessible for potential reusers. It makes it human readable and it can provide citation and peer review. Examples of the *Journal of Open Archeology Data* (JOAD) and *Research Data Journal for the Humanities and Social Sciences* (RDJ).

Publish your sharing policy with the Data Reuse Charter

An initiative by DARIAH, Parthenos and other partners and it has been presented by Anne Baillot. It is another online environment as a work service where you create a profile (researcher, institution, cultural heritage institution or laboratory, another body that use primary or secondary cultural heritage data, data hosting body, etc.) and state your policy regarding the reuse of your own data.

The benefit of this charter is obvious: you have no more case by case agreement, no more blurry conditions; all is clear and set in one single place. Some basic features are:

- Register according to your personal or institutional profile
- Get in touch with cooperation partners and collections relevant to your activities
- Gather information on relevant topics such as licensing
- Gain visibility and recognition in the international research ecosystem
- Provide an opportunity for cooperation, retrieval of new collections because any institution, etc.
- Emphasize the important notion of citation (reference to the origin or the owner of data and provide sustainable tools for the further citation of your own data).

Conclusion

Making a DMP is defining how the data, within a project, will be:

- Described
- Shared
- Protected
- Preserved

A DMP contains:

- A data lifecycle description (including long term preservation)
- A data description
- A description of the data policy
- The associated costs

A DMP helps at secure and perpetuate data and is above all a way to see clearly a project's organization, on the data side. => Very strategic, but not technical

When?
Before the first data are created and Regularly updated
Why?
Funders wants it and it is a research good practice
Who?
A team work

A research good practice

A DMP formalizes inside a unique document a set of elements and information useful for the project monitoring and for a good management of the results. Its practical benefits are:

- Better understanding of the data
- Long-term research is easier
- Underlying data is more accessible
- Research more visible: better citability
- Save you time
- Allows to focus on research, increasing efficiency
- Prevents problems in understanding data and metadata in the future.

- Data are easier to preserve and archive
- Benefit for both yourself and others in your field. It might prevent duplication of scientific efforts to re-collect your data and it can lead to new and unanticipated discoveries you might not predict.
- Useful for PhDs: good practice, key data available, thesis and underlying data

=> optimization, "profitability" & perpetuation

The DMP Aide-mémoire

- Is there a model required by the institution/funder?
- Who will contribute to the DMP (team members, partner's projects)?
- Who can help (documentation professionals, IT, etc.)?
- Who will use the DMP?
- Use of an online tool?
- Come quickly with a first version
- Updates: required and/or desirable milestones
- Final version
- Identify datasets

Appendix : DMP tools

There are several tools available for helping the creation of data management plans. Two of the most commonly used are DMPTool and DMP Online. Both operate as “wizards” and provide prompts for the user to fill out in order to create their data management plan. You can save your plan, print it, or export it to your computer. It includes also templates for H2020 projects.

DMPonline Exercise

1) create an account
2) choose a model
3) create and share a plan
4) identify a dataset
   - definition criteria of a dataset
   - reasoning of the decision (reproducibility, cost, etc.)
5) others datasets? (granularity, strategy and concrete practice, impact)
6) commenting fonction
7) export

Contact

Marie Puren and **Charles Riondet,** Ph.D., are junior researchers in Digital Humanities at the French Institute for Research in Computer Science and Automation (INRIA) in Paris, members of the Alpage laboratory (INRIA – Paris Diderot University). As collaborators to the PARTHENOS H2020 project, they focus their research on the development of standards for data management and research tools in Arts and Humanities, and they currently work on the creation of a Data Management Plan for this project.

Marie Puren also contributes to the IPERION H2020 project, especially by upgrading its Data Management Plan. After being a lecturer and a responsible for continuing education projects at the Ecole nationale des chartes, Marie Puren has been a visiting lecturer in Digital Humanities at the Paris Sciences et Lettres (PSL) Research University. Her main publications belong to fields including intellectual history of the XXth century, French studies and digital humanities. Marie Puren has been awarded a Ph.D. in History at the Ecole nationale des chartes – Sorbonne University. She holds Master's degrees in History and Political Science from the Institut d'Etudes Politiques de Paris, and in Digital Humanities from the Ecole nationale des chartes.

Charles Riondet, History PhD and archivist, is also involved in H2020 EHRI project as a metadata and standards specialist, with a focus on archival metadata (EAD, EAC-CPF).

Twitter: @puren1406 & @charlesriondet

Email: marie.puren@inria.fr & charles.riondet@inria.fr

4-Persistent Identification

Persistent Identifiers (PIDs)

Ondřej Košarko, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

Why PIDs?

- Making data available but it then requires referencing and interlinking
  - resources
  - semantic definitions
- Reproducibility and reuse of results: collaboration on data can improve it or at least verify the results.
- Location on the web is not a good identifier. It will stop working. You upload your data and get a link to it, but the problem is that URL or links will stop working - I am not even saying that they *might* stop working, it will stop working, it is a matter of time.
  - The content can be changed at will: When you share a link, the target of the link can be changed, not necessarily to cheat others, but to improve data and correct it but this has an impact for a paper written years before for example. In this case it is not possible anymore to access what was there.
  - Names can lead to certain expectations: URLs often contain semantic words that can lead to some conclusions, like the "final_version.pdf" and "final_final_version.pdf" extensions

Good identifier

- Persistence
  - Longevity: in an imaginable future, we should still be able to get sense of what the attached idea was.
  - Commitment of involved actors for the future.
- Uniqueness of the identifiers: You should not be able to change it in the future, otherwise nobody can reproduce your work and it might even not be findable anymore => One ID for one "object" and ideally one "object" has only one ID (but that can't be guaranteed across multiple PID systems or even in some systems on their own)

PID systems

There are in fact different systems because people have different idea of what persistent identifiers should do beyond the identification, the persistence and the uniqueness.

- URI/URN (IETF standard)
- Handles (DOI)
- PURL

- ARK
- Info-URI
- XRI

Different concepts

- Naming schema, like URNs

urn:oid:0.9.2342.19200300.100.1.3
Eventually http://oid-info.com/get/0.9.2342.19200300.100.1.3
Urn:oasis:names:specification:docbook:dtd:xml:4.1.2
But it might not land on something that you consider as an authoritative resource if you are not even sure that this is a real identifier, because it might not be easy to find.

- Resolution system

It provides a way to give you a location, like web driven proxies, a mean to see the resource that is assigned the identifier. It means that if your browser is enabled with a DOI or Handle plugin, you could click the DOI and find the resource. If not, you can still use the proxy that is provided and that does the resolution. So when you click a link, you end up on the resource.

  - DOI Handbook: doi:10.1000/182 or http://dx.doi.org/10.1000/182
  - Mark Twain's sketches, New and Old: hdl:loc.gdc/scd0001.00162117695 or http://hdl.handle.net/loc.gdc/scd0001.00162117695

- Services around DOIs

With DOIs, you have an infrastructure built around like Crosscite does, it allows to get a citation just by providing a DOI, so you don't have to download the resource. See Crossref's Auto-Update for ORCID records.

- Parts/fragments

It is useful for continuous data: some DOI providers also offer identification for fragments of a described resource. If you have PID assigned to an audio document for example, one approach is to create another identifier for a specific part is this audio file.

- Costs and ease of use

With handles, when you are on the provider side and you want to start providing handles to your users, (you are not using someone else repository) it might take some time because you have to communicate with the global handle registry and pay some fees (50$ a year). With DOI, the pricing policy differs, if you want to attach PID to a great number of documents, DOIs might come expensive.

Versions and PIDs

The rules are up to the PID providers.
If PIDs should provide uniqueness, how can you assign one PID to two different versions of the same document? One objection might be: how do you keep track of the versions? Where is which version? Which is the newer? You can keep it in your metadata, use the DublinCore relation “isreplacedby” and provide information about the other resource that was previously used. But with PIDs you might not always get what you expect:

- What is a substantial change
- If the point of PIDs is persistence and uniqueness, shouldn't new versions “automatically” get different PID?

- The versions can be linked in metadata

Granularity
The question is to know if there is some minimal size from which you should assign a PID or not. If you are working on textual resources, you need to refer to one particular character and that resource because it is some medieval text and it is the only appearance of the character you have found, so it makes sense to assign a PID to it. Basically, the PID system does not limit what the objects are, but the provider might because of the rules that can be set up in such a way that you won't fulfil it. With repositories, if you are uploading a file that contains a character, it might be odd but still if you are able to provide descriptive metadata, it makes sense.

Persistence
It is the idea that PID is persistent, not the resource itself. The resource is made persistent by being added in an archive that guarantees that it will take care of it for years to come. PIDs provide a way to make the resolution possible, not the resource itself. In certain cases it makes sense to withdraw a resource, the PID should remain as the descriptive metadata (and explain the withdraw: claim not true, institute disappeared, etc.).

Shortref.org: **(Moving away from repositories to) Citing arbitrary views of data**

How to cite the use of your research data on one specific aspect which might be representative of the phenomena but not for the complete picture => Pictures are better than words. For example, if the dataset and the querying service are online with such an URL: https://lindat.mff.cuni.cz/services/pmltq/#!/treebank/pdt30/query/IYWgdg9gJgpgBAEgK4AcUwE5wFwF44DaAUAM7AC2MIALhjFcCYgEYzUDu9YANEaJLEQAbCJyx5CcALrSA3H3DR4CVhy458BOAEI4wMAGMYJahCzI0mXgAo41s7AwgUdI7CYX0GPmChwHmCAAZhBCIuwe4ZgAIERmdgFOIWGiHqhePn6Jzq4w7sKiMXDR0kREQA/result/svg. You can use a shortener even if it might have the previously mentioned issues. Shortref.org service (handle) is part of LINDAT/CLARIN, the Centre for Language Research Infrastructure in the Czech Republic.

Metadata and url:

- Assign a PID (handle) to an url, or make it usable - even if the service is down
- Possible to change the location: We keep a track
- Additional metadata
- Healthcheck on the location

For users:

- Fill the form: URL & metadata
- Save the token for future updates
- Use the handle as reference

For services:

- Use REST API: provide metadata and URL, store the token

- Show the PID
- Just a click away

Overview:

1. Creation
2. Monitoring: Health check, the locations are periodically polled. After certain number of failures, an error page is shown: http://shortref.org/resource_down.html#hdl=11346/FFF-TN7H
3. Resolution

Contact

Ondřej Košarko, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University

Ondřej Košarko is a programmer working at the Institute of Formal and Applied Linguistics (UFAL), Prague, Czech Republic. He is one of the developers behind LINDAT/CLARIN repository. The repository is based on DSpace and has been modified to meet the needs of CLARIN centers. This modified version is now deployed in several member institutions. He is also responsible for parts of shortref.org, a tool to ease persistent data citation, and various other bits and pieces like this guide for choosing the adequate licence.

Institutional websites: http://lindat.cz & http://ufal.mff.cuni.cz/
Email: kosarko@ufal.mff.cuni.cz

Canonical Text Services

Christopher Blackwell, Furman University

CTS Implementation has been developed through the Humboldt chair at the University of Leipzig. This presentation will be an implementation of the values that Ondřej Košarko just articulated.

CTS is a protocol for identifying and retrieving passages of text by means of machine actionable canonical citation. This is something we have developed for the Homer multitext project since the 2000. In this context, CTS is not necessarily traditional citation, but unique, unambiguous citation values that are independent of technology or format allowing to cite a physical book, a digital text in TEI-XML or not. Ideally a citation should capture the semantics of the text, what we call the citation hierarchy and the bibliographic hierarchy.

I am a classicist and a lot of our texts have been read and deeply cared about for a long time and they have good traditional schemas of citation, which map nicely into the digital realm, for example Homer, Iliad, book 1, line 1. Some texts don't have a traditional schema of citation at all because they might be 20th century texts or they have a traditional schema of citation that doesn't work well in the digital realm.

CTS consist in two parts: the probably most important is URNs, the identifiers, machine actionable citation for identifying passages of text. There is also a CTS service protocol: request and response for retrieving information about a passage of text; http using CTS URNs.

CTS is based on a model of text that is an ordered hierarchy of citation objects, we call this OHCO2, pronounced "ochio 2". CTS is not a cataloguing application, it is not an editing application, it is not a search engine, it is not a browsing, reading or commenting application, but we have found it to be useful as a component in all of this things. We found CTS URNs to be extremely useful for expressing the results of automated textual analysis and also the work of human editors. CTS is not limited to TEI-XML text, it is not limited to digital text; CTS URNs can identify passages of text in physical volumes. CTS can't say everything about the history, nature and meaning of a text, which is why we write scholarships on commentary, it is just for identifying and retrieving passages of a text.

Data model: Ordered Hierarchy of Citation Objects

According to this model, a text consists of citation objects. In a traditional format, each of this is a precise identification of a citation object. The original OHCO goes back in 1990 and consider a text as an ordered hierarchy of content objects. It immediately caused violent civil wars and people started to fight about what was content. For example, is markup content? Is whitespace content? If you have two whitespaces in a digital text is that one content or two?

Our approach was to back it up and think in term of citation objects. In CTS, the citation object has textual content and it can be mixed content or just plain text. By separating the textual content from the citation object, we have a lot of flexibility. We can have textual content itself or embedded with some kind of analysis in our text. We can add it without altering the citation and the text doesn't even have to look like text. For example, in one digital exemplar of one of Iliad text, you might not be interested in the content of the Greek corpus but in metrical values.

Ordered hierarchy of citation object
It is ordered because you read from the beginning to the end, the sequence matters.
It is a hierarchy, text may be organised by containing elements, ex. Iliad book 1 contains 611 citation objects, which can be poetic lines. The hierarchy may be only one level deep or many levels deep, different sections of a text may have hierarchies of differing depth. CTS is good at this and we use XML, so we know the citation objects, headings and paragraphs, etc.

Text in a bibliographic hierarchy
CTS also put text in a bibliographic hierarchy which is a sort of FRBR (Functional Requirements for Bibliographic REcords). In CTS, text belongs to text group that may be author or not. Notional works become real when you have versions of them. The Iliad is an abstraction, but you sometimes want to talk about real things, like post translation of Iliad, specific editions and texts of the Iliad. A text may also be an exemplar, which is a specific instance of a version, for example Thomas Jefferson's personal copy of the edition of Iliad, in which he wrote notes. In a digital realm, we see the idea of an exemplar as a specific text derived from a version. So, if we do diplomatic edition of the Iliad, which is a *version*. From that, if we produce a normalised version, we would call that a *digital exemplar*. It has a clear *relationship* to a version, it depends on it.
The CTS URNs are our effort to capture all of this semantics and bibliographic hierarchy and citation hierarchy as precisely or imprecisely as we need to in a machine actionable citation. This *arbitrary identifier* happened to be the numbers that the TLG canon of Greek authors and works used. In this authority list, TLJ0012 is homeric epic, TLJ001 is the Iliad, and a particular version MSA, and then citation book 1, line 1. You can then mix and match the components of this, ex.: Homer, Iliad, no version, 1.1. So this is the Iliad in general book 1, line 1, with no specific version in mind. It identifies any version of the Iliad that has a book 1, line 1. There is a manuscript in Venice that begins with book 16, this doesn't cite that manuscript, but any version out there that does have whether in English or French, etc., that has a book 1, line 1. We can express *ranges*: from book 1 line 1 to book 1 line 10. There is no reason to believe that it will results in 10 citation objects, they look like numbers but they are just arbitrary numbers. Besides, you can have *mixed range*: from book 1 line 600 through the end of book 2.
With CTS and this *sub reference*, we can identify more specific strings of text within the citation element and, at least in the Homer multitext implementation of CTS, we don't retrieve on this. If you put this on the CTS server and with a "get passage" request, you will get all of book 1 line 1 and then it is your problem to find the first instance of the string meaning in it.

CTS Service

It is an http request, you have a URL to the service. Request equal "get capabilities", will return a catalog of text that the service knows about and can offer. And we have schemas that define how that catalog works. These are the three most important CTS requests with what you can get everything done:

- get valid reff
- given the URN: will give every valid citation define by the URN. This is useful in cases when you have fragmentary text, ex. manuscript that begins with book 16 and other text where you need to know where the citations are.
- get passage given the URN: (also get passage + get first reference)

Text for CTS

Any text uses pieces that can be identified by Canonical Citation CTS compliant. TEI-XML works great, it is a little complicated because you have to do some processing, but it works great. The Homer multitext with our work starts with TEI-XML and we process them into RDF statement. The complete expression of OCHO2 model book 1 line 2 of our edition of the Iliad has a series of RDF statements. We are confident on the data model because we regularly start with XML and bring it to RDF and back into XML fragments for serving. This kind of round tripping suggests that this model is actually capturing the semantics of the text.

The simplest possible CTS text will be two column tab delimited/separated values files where you have an URN and text content. The URN captures the citation hierarchy and you have an ordered hierarchy of citation objects and this will be a good CTS text. Over the years, we have implemented CTS on a lot of ways, google app, engine, existDB, etc.

Homer multitext implementation of CTS

It is a downloadable virtual machine. The fundamental difference between our implementation and what Matt Munson is going to show you, it that is works with TEI-XML and keeps the file as XML that allows constant integration that way; our involves RDF backends, so it is two different approaches.

Recommended reading

Amy H. Blackwell & Christopher W. Blackwell, Hijacking Shared Heritage: Cultural Artifacts and Intellectual Property Rights, 13 Chi. - Kent J. Intell. Prop. 137 (2013). Available at: http://scholarship.kentlaw.iit.edu/ckjip/vol13/iss1/6

Contact

Christopher W. Blackwell holds a B.A, summa cum laude from Marlboro Collect in Vermont, USA. He holds a Ph.D. from Duke University, where he was the William H. Willis Fellow in Classics. Since 1995 he has been on the faculty of Classics at Furman University in South Carolina, USA. He served as Chair of the Classics Department for 14 years, until 2015, and is currently the Louis G. Forgione University Professor. Since 2001 he has been Project Architect, with Neel Smith, of the Homer Multitext, a project of the Center for Hellenic Studies of Harvard University under the editorship of Casey Dué and Mary Ebbott. With Smith, Blackwell is co-creator of the Canonical Text Services protocol and the CITE Architecture for identification and retrieval of scholarly resources by canonical citation in networked environments. Blackwell has led several digitization projects and has

collaborated with scholars in the U.K., Italy, Germany, the Netherlands, Greece, and Croatia. He has published two books on the history of Alexander the Great, and articles on topics in Classics, Computer Science, Intellectual Property Law, and Botany.

Academic website:
http://www.furman.edu/academics/classics/about/Pages/FacultyandStaff.aspx
Email: christopher.blackwell@furman.edu

Workshop: CTS with CapiTainS, Hook(Test), Nemo, and Nautilus

Matt Munson, Humboldt Chair of Digital Humanities, University of Leipzig

HookTest, Nemo, and Nautilus Setup

1. Instal Docker up and running
2. Instal Hooktest, create a virtual machine
3. Nemo nautilus

This repository contains the documents that will be edited during the CTS Workshop at the DH 2016 Conference in Krakow. Nautilus is a python based CTS API using XML TEI files following CapiTainS guidelines. Nemo is a user interface, starting to grow as a CMS (Content Manager System), which reads its core data from standard CTS API calls (and so is compliant with any standard CTS API normally) and is starting, as of the beta of 1.0.0, to accept annotation resources such as treebank and images in the form of plugins. Hooktest is a software developed to make unit tests on XML repository regarding CapiTainS Compliances.

CapiTainS compliance

(urn:cts:gerLit:ger0001.ger001)

1. Rename the files (ger0001.ger001.opp-{ger,fre}1.xml)
2. Create Repository structure (data/ger0001/ger001/)
    a. Run HookTests
3. Create the __cts__.xml metadata file for the text group ger0001
    a. Run HookTests
4. Create the __cts__.xml metadata file for the work group ger001
    a. Run HookTests
5. Add URN to ger0001.ger001.opp-ger1.xml
    a. Run HookTests
6. Add line numbers
7. Add refsDecl
    a. Run HookTests
8. Check out Nemo!

Hooktest is a continuous integration environment for text. When you do code development with collaborators, you know the continuous integration is where every time you make a change with the code it runs certain test to make sure you didn't break anything. And this is basically what Hook is for text. Every time we make a change to a text, it will test the text against whatever we want to test it against.

Run Epidoc test on the Goethe repository: ./epidoc.sh DH2016-master: Some tests are passed and other failed. The results are exported as Result.json & results.html.

The webpage tells you have two files and none passed (the tests). Then, add metadata file into the repository, description of the text groups (author) and of each work level, describe as precisely as you can what you actually have.

Citation schema

Poem: citation level & line citation level. Make sure that the citation scheme is correctly described and encoded. The citation scheme for a poem is by line, so this poem has 30 lines=> add a number to each lines. Add to the TEI header where are the citations located. Tell the XML parser how to find these two (levels of citation): In the encoding description, in the epidoc, you have a reference declaration, and we say it is CTS and then you have two patterns: one for the line and one for the poem and Xpath tell where these levels are located.
In Kitematic, you can get the Nemo Nautilus, which is the CTS server and it shows that we have one collection (German Literature), within it we have one author, etc. You can also make CTS API, see the documentation.

Some Requests

- .../api/cts/?request=**GetCapabilities**
- .../api/cts/?request=**GetValidReff**&urn=urn:cts:gerLit:ger0001.ger001.opp-ger1&level=2
- .../api/cts/?request=**GetPassage**&urn=urn:cts:gerLit:ger0001.ger001.opp-ger1:1.2
- .../api/cts/?request=GetPassage&urn=urn:cts:gerLit:ger0001.ger001.opp-ger1:1.2 **@mit**
- BUT NOT .../read/gerLit/ger0001/ger001/opp-ger1/1.2@mit

And you can also refer to a specific word or letter. It will lookup and return the whole line (to get the context of the occurrence). This allows you to refer to extremely specific parts of a very specific text. Then you can add annotation to this very specific word in this very specific edition that is in this very specific place. If someone wants to look at your annotation and see your work, they will be able to go back to your text and see exactly what you were talking about.

Useful links

- http://capitains.github.io/pages/tutorials
- https://github.com/Capitains/docker-hooktest
- https://github.com/Capitains

Contact

Matt Munson, Humboldt Chair of Digital Humanities, University of Leipzig
Matthew Munson received an MA from the University of Virginia in Religious Studies, his thesis studying the use of the Greek word for law (νόμος) in the letters of the Apostle Paul. Before joining the Digital Humanities Team, he worked at the Scholars' Lab at the University of Virginia and in the DARIAH project at the Göttingen Centre for Digital Humanities at the University of Göttingen, Germany. He is currently working on his PhD in Theology in Leipzig studying the automatic extraction of semantic data from biblical texts and the automatic tracking of semantic drift between corpora.
Academix Website: http://www.dh.uni-leipzig.de/wo/team/

Email: munson@dh.uni-leipzig.de

5-Evaluation, Acknowledgement and Credit Circulation

Open Peer Review

Julien Bordier is a sociologist, an independent scholar who works for OpenEdition on an experiment of open peer review and open commentary on a corpus of pre-publication in French.

Introduction

Peer review is what ensures a scientific publication to be a real scientific publication. Traditionally, it is said that everybody has to be masked to make clear and objective review of text. So, a peer reviewer is like Spiderman. And Spiderman says that “With great power comes great responsibility”, because it is a great responsibility for a reviewer to evaluate a text as it will determine if the pre publication will be published or not. The point of *open peer review* is to *open the process* and that means that nobody is anonymous anymore. This practice is developed in the perspective of open access but one can imagine a publication which is not in open access but which also practice open peer review. In the experience we implemented with OpenEdition, everything happened online: the prepublication was displayed on a scientific blog, hosted on Hypotheses. Our hypothesis when we experimented was that we could have a better level, a better quality of scientific communication if everything is open and accessible to everybody. The idea is to make possible conversation between the authors and the reviewers. I will no longer follow the metaphor but you have to know that when Spiderman wants to make a kiss to Marie Jane Watson, he takes off the mask!

Presentation of the experiment

Protocol of the experiment with two different branches:

- Open peer review
- Open commentaries

The report is called: *Open peer review: from an experiment to a model A narrative of an open peer review experiment*. It is available here in open access: https://hal.archives-ouvertes.fr/hal-01302597
You can annotate it with the tool we used for the experiment: hypothes.is

This was implemented on a famous journal called Vertigo, hosted on Revues.org. It is a French-speaking journal based in Montreal about environmental science and ecological issues. We implemented the two different protocols. The first one was open peer review protocol: we published online, on the journal's blog, five pre publications that was submitted to the journal, we found reviewers and asked everybody to know if they agreed to disclose their identity and to make the process *visible* online - of course they accepted. In a practical way, you have the blog post, which is the pre publication and the comment of the post are composing the evaluation report. So, it is quite identifiable and relevant because it is blog form and people are getting used to it. This was the first and main part of the project and we were very ambitious about it and, afterwards, it worked quite well.

We also implemented another protocol, which is open commentary. At first, this second protocol was not for us so appealing but it revealed itself to be very interesting. In this second protocol, the journal selected some articles to run the experiment after receiving a lot of propositions and especially articles that were not really well written. The idea was to publish online texts that has a real scientific interest but that was not really well formalised and to ask the community, not just the designed reviewers, to comment and to help the author improve the quality of their pre publication. This second protocol appeared to be really relevant and it became a very helpful insight for the authors engaged in this process.

The technical implementation is the editorial environment made by Revues.org (where the journal is hosted) and also by Hypotheses.org (where the journal has its blog). The point was to make both of them communicate and it was important to develop something like that because they represent a legitimate space where scholars recognise and regard.

The traditional review process is a wheel of evaluation: the reviewer makes his remarks about the text, usually in a pdf or a doc file where he inserts comments within the document. Our challenge was to find the right tool to make reviewers annotate the text. We decided to hypothes.is.

Results of the experiment

Open peer review is subject a great discussion and a lot of people think that it will not work while others think that it is the future of scientific publication. I personally don't know but what I saw as a sociologist, playing as an assistant editor implementing the experiment, is that there was a lot of enthusiasm around it. I think that it is the first important point because it means that the scientific community is quite ready to open the evaluation system and especially for publication. Times are changing, like the Bob Dylan's song. Anonymity used to be important in the 70's for example, women scholars asked for it in the review process in order to be fairly evaluated.

Another interesting result is that when you read literacy about open peer review, you have a kind of myth saying that open peer review is quicker than traditional review, but from my perspective and what we experimented, it is not true at all. One can think that as it is on a computer it goes quick and easy, but it is not the case, it is the same as in traditional peer review. Open peer review does need human mediation, it is not robotic or cybernetic peer review. It is still difficult to get reviewers, to get people to follow the deadlines and so on. And it is even more important for the open commentary protocol because the point was just to help the author, so scholars does not have this habit, they usually do it during a seminar but not for publication. It was quite difficult to explain the process and its issue. It was also difficult to get people from outside of the scientific community to comment the text. It is interesting because it means that we still have a lot of work to do about how to get a better integration between scientific community and the rest of society.

Overview of the results

- Enthusiasm
- Need of human work
- The experiment worked because of the integration to a relevant environment (between revues.org and hypotheses.org - seen as a legitimate publishing environment in the French speaking area). For example, Mickael Bon is trying to make an open peer review journal platform, in STM, called Self Journal of Science, but he has a lot of difficulties to get scholars publishing on his platform because it doesn't have the legitimacy.
- In open peer review, reviewers and authors are able to talk to each other and it worked quite well. I do not have empirical results to present now because our corpus was small: it was just five texts open peer reviewed and five texts opened to commentary.
- As a sociologist, I asked questions to all the protagonists of the experience: did you search the person you were doing the review? Of course, everybody did it! And of course, even in classical [closed] review, everybody try to do it and to know who is reviewing and who is the author. And of course, it is really to easy know and that is why classical review is just a myth that now should be opened.
- People did get information about each other and knowing what each one is working on, you can feel how the remarks are legitimate and true. I think that even for the journal or the editor, it is very important because you can have a more fluid conversation in the exchange of information.
- At last, in the community, it shows who is actually working and who is actually doing good work about reviewing. To get a comparison, the text that was openly peer reviewed was also traditionally reviewed, it was a double blind review. The result is funny because in the traditional review, the reviewer just added a line on each section; whereas in the open review, reviewers make efforts to be clear and understandable.
- Besides, it is not only the journal that can judge the level of the evaluation, it is the whole community.

Limits and further potentialities of the experiment

- In order to get users to annotate the text, we asked them to use hypothes.is and it is clearly not ergonomic because you have to ask the reviewer and the authors to create an account on hypothes.is and to install it on their web browser. I had phone calls from older scholars who had some difficulties to use it. -It is not a judgment- but when you have a discussion with someone who do not make difference between Google and its mailbox, when you have to get him using an annotation tool which is not implemented on the platform he uses, it is really difficult. At the end it worked. This is reminder for the idea that you do need this specific human work around open peer review - you have to get people really doing.
- Hypothes.is will be implemented on OpenEdition Books! I think it is a good start, but I think that hypothes.is has some problems, for example on Vertigo Journal, on one

of the text, when you open hypothes.is the annotation refers to another webpage in German which has nothing to do with the displayed text. This is a big problem.

- Potentialities about credit circulation: it makes visible the work of everybody involved in the review of pre publication. It is a very important part of a scholar but it is an *invisible task*. Our commitment in open peer review with this experiment is to make possible to credit everybody who worked on the text. In the way we did it, reviewers are credited by citation, so once the text is published in the journal, you can find the name of the evaluators. So if a scholar wants to refer to it, there is an URL that you can use for example in your curriculum vitae. For the moment, it is just *bricolage* or *craftwork*, it is not yet real digital publishing; an important issue would be to make the name of the reviewer or commentators indexable. It would enable when you use a tool of indexed authors and find what an author wrote, what he reviewed, what he commented.

Conclusion

With this experiment, we tried open commentary and open peer review on pre publication, before the document is really published. We just tested it this way but every form are hybridable: you can open texts to commentaries post publication, you can also close or disclose some parts of an evaluation, it only depends on the agreement you have with reviewers and authors. I think that editors should keep hands on this to find the right policies, the most relevant for their field of research and their publication policy.

Practical Session

- Install hypothes.is on your web browser
- Annotate parts of the report for the European Commission published on HAL: https://hal.archives-ouvertes.fr/hal-01302597
- Give feedback: ergonomic, how you feel with the tool, any particular problem?

Contact

Julien Bordier, sociology PhD, independent scholar, editorial adviser, works on public-space issues. He conducted the open peer review experiment for OpenEdition / Centre pour l’édition électronique ouverte.
Twitter: @julbordier
Email: julien.bordier@openedition.org

Simplifying License Selection

Ondřej Košarko, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

Neglected task

- Research(ers) rely on availability of data & software, but license clarifies what you can do with it
- Authors hold exclusive rights: unless they “give up” the exclusivity, the data are not “usable”; That means assigning a license

Attributing a license to your data is a necessary task as research itself relies on some data. For example, we use data produced or collected by colleagues and we run experiments on them, we evaluate what other researchers did, we use the tools they produced in order to move the field further. Still, without a license that would clarify what you can actually do with the data or software, you cannot work. Of course you can do whatever you want with what you found on the Internet, but the problem is that you shouldn't do it this way because you don't know how the data was collected, if there are private information, if data are reliable, etc. The only person who should know it is the author of the data and it should be the only person that has the right to allow derivative works with the data and tools and publish them. In order to make it available for a wider public, he has to give up on some of his exclusive rights by providing a license that gives either individual persons or in better case the general public the right to operate or to modify the data. As seen during the previous session about open revisions or open reviews, it is necessary for the good advancement of the research to have access to the underlying data. I called this a neglected task because not all of the authors have attributed licences to their work.

Resignation

Some of those reasons might be:

- Not distributing the work
    - Law is scary/boring/complicated
    - License texts are long and hard to read

Without making a license statement, it is like hiding data on your hard drive thinking that you don't need to care about it.

- Not attaching a license
    - I did not assign a license so everyone is free to do as they wish, right?
    - No one reads them anyway

If you want people to do whatever they want with your data, the way you obtained the data allows you give them this right, it better to tell it. There are several licences that say it in a few words.

- Using a license that is familiar with
    - Good attempt, but GPL with data?

You might hit the right one or not, but some of the well known licenses might not be the best solution, for example licenses that are good for softwares are not adapted for data. If it relates to softwares, executables and libraries, if you attribute it to data, what does it actually mean as there is no executable, no library. It equals to not attributing any license because the user cannot get any special right.

Data and software

- Overall different beasts
    - Regulated by different directives
- Different caveats
    - Software: libraries versus executables
    - Databases
- Different licenses [for data and for software]

There are such differences that it is in fact regulated by European directives. On data, you are often guided by the regulations for databases or collections of work. For software, you might find distinction between library and executable.

Making it easier

One thing you can do each time you want to release something is to ask the person in charge what you can precisely do and what you cannot. You can also build these questions into a process, a tool.

- Get together with people who know this matter
- Prepare a tool/process to help choose
    - Filtering based on requirements
- Capturing differences of data and software
- Adhering to limitations, according to the way you obtained the data
- Promoting open access, as free as possible

When you turn this question into a process, you do one more thing, you can push people in certain direction: advocating open data and open access licensing if the strings attached allow it.

A few tools

- http://licentia.inria.fr/
- http://wizard.elda.org/index.php
- https://tldrlegal.com/
- https://ufal.github.io/public-license-selector/

GPL license: it is nice because you can modify but you have to disclose source code and instal instructions. There is also a quick summary of the provided full text and this can be a collective process with for example the change sets, you can go through the changes of the summary if there are any. There is also a section for comment where people can talk about some not obvious interpretations of the license. Terms of services are even worst than licensing: they are longer and the language is really complicated. When working on the term of services, we should take care of the last changes. If you are using it, you should choose the latest version. It is supposed to be for users, because it explains the framework in which you are allowed to use it, but it is written for lawyers.

Tl;dr legal

Tl;dr stands for *Too Long Didn't Read*, it is used on the Internet either to state that you didn't read it or when you ask for a summary. This perfectly grasps the licenses issues, terms of use, community driven summaries (textual, can/cannot/must), tracked changes, verified content [by a legal expert]

Public License

It grants certain rights not to one particular user, but to the general public (everybody). You can configure the tool to allow modification or not (non derivative is not open data compliant).
According to the permissiveness, you can find open data compliant licenses. From the software development perspective, there is something that might happen is that you might find a combination of licenses that are incompatible: one licence can require you to distribute all the work and another one not to make derivative, it is conflicting. So there is also a licence compatibility issue and it depends on an interpretation that can be different from the official one provided by the authors of the licenses.

- A lot to choose from:
  - 78 OSI approved licenses for software
  - Creative commons, Open data commons, etc.
- Open access
  - Not all public licenses meet requirements of Open data/open access [Open data compliant or endorsed by open data]
  - http://opendefinition.org/licenses/

TOOL: Public license selector

- Tool
- Github
- Pawel Kamocki (legal expert specialised in IP and personal data protection), Pavel Straňák and Michal Sedlák
- Distinction between data & software
- User friendly, asking questions and providing with comprehensive explanations
- Promote open licenses
- Licenses not compatible with the answers are removed from view
  - Question about licenses already used
  - Table of compatible licenses
- Licenses ordered based on openness

How to attach a license?

- Some licenses tell you to insert it prominently in all relevant locations. But, is the download page enough? At the beginning of each file? What about textual only files, how to make the distinction between the text of the license and the text of the file?

- A solution is to mention your license in your metadata [but it doesn't apply to the metadata themselves-descriptive metadata are not copyrighted]
- Download license in RDF (Turtle syntax) to attach it to your data

Resources

- Public license selector: https://github.com/ufal/public-license-selector
- Article about the public license selector: http://www.lrec-conf.org/proceedings/lrec2016/summaries/880.html
- Open Source Initiative: https://opensource.org/licenses/alphabetical
- Creative Commons: https://creativecommons.org/share-your-work/
- Open Data Commons: http://opendatacommons.org/licenses/odbl/

Contact

Ondřej Košarko, UFAL

Ondřej Košarko is a programmer working at the Institute of Formal and Applied Linguistics (UFAL), Prague, Czech Republic. He is one of the developers behind LINDAT/CLARIN repository. The repository is based on DSpace and has been modified to meet the needs of CLARIN centers. This modified version is now deployed in several member institutions. He is also responsible for parts of shortref.org, a tool to ease persistent data citation, and various other bits and pieces like this guide for choosing the adequate licence.
Institutional websites: http://lindat.cz & http://ufal.mff.cuni.cz/

Email: kosarko@ufal.mff.cuni.cz

Evaluation of the SSH and the evolution towards open science

Ioana Galleron, European cooperation in Science and Technology, COST action

General outline

- Evaluation systems in Europe and the place of the SSH
- Challenges of SSH evaluation
- A new player in the field: European Network for Research Evaluation in the SSH (ENRESSH) (COST Action 15137)
- Gathering data about the SSH: main problems
- Some observations about the impact of the open science trend

Evaluation systems in Europe and the place of SSH

They can be classified into four criteria (Geuna et al., 2001).

- Evaluation performer: national, regional or institutional player
- Evaluation purpose: funding allocation or strategy formulation
- Criteria for evaluation: there are many but four main groups: quality, quantity, impact and utility.
- Methods: bibliometrics (impact factor), scientometrics (also takes into account the research environment, size of teams, etc.), peer-review, peer-review supplemented with bibliometrics/scientometrics (*informed peer-review*)
- An interesting dimension in classification terms should be who is evaluated (individuals, teams, institutions?)

There is a lot of discussion around classifications of systems because for example some people think that a funding system is not an evaluation system, it is just a performance based system.

To summarize, evaluation systems can be placed somewhere between two "typical" models:

- evaluation conducted ex-post (after the research), performance-based, indicators, peer review.
- evaluation size based, where allocation of funds or strategy will be formulated looking at the teaching volume, students, the staff, etc.

The general trend is obviously towards the first type performance-based resource allocation system because of the new public management policies and everybody looks for accountability and value for money.

What about SSH evaluation?

When you look at how it is specifically done in SSH, the image is more blurred. We conducted a survey about practices before the beginning of the action to understand the system applied to the SSH. We had 43 participants involved in research evaluation from 25 European countries answering about:

- the level of the evaluation protocol (national, regional, institutional)
- disciplinary differentiation

- who is evaluating
- object of evaluation
- purpose (funding/strategy)
- methods
- timeline
- transparency
- costs

Even if the interviewees had a good knowledge of evaluation, they often had no clue about how to answer our survey; and this is already informative about the situation of the SSH evaluation.
=> Good degree of agreement amongst respondents from the same country about who is evaluated, the methods applied and the link between evaluation and funding.
=> But there is a lot of disagreement or even misunderstanding about **terminology**, ex. evaluation/assessment, "excellence", etc.
=> Evaluation is mainly done nationally and linked to a performance-based funding system.
=> National publication database about SSH production in 13 countries
=> Respondents signal the existence of an SSH specific evaluation in 14 countries (but with a low degree of agreement about the existence of specific methods of evaluation for the SSH).

Methods

- Peer review
  - Most countries use peer review to evaluate SSH
  - Only 9 use informed peer review
  - In most countries, peers apply criteria but there is no agreement about criteria among participants from the same country
- Bibliometrics/scientometrics
  - It seems to be the main method in six countries
  - There is no agreement on the kind of data used to evaluate the SSH

Transparency

Respondents from 14 countries consider their system to be transparent regarding the criteria applied for SSH evaluation whereas 11 consider it opaque. There is a lot of disagreement about this dimension.

Challenges of SSH research evaluation

- Scholars don't want it because of academic traditions. For example, looking at book reviews one observes the tradition of courtesy, of collegiality, you don't say bad things about your colleagues; ideology, some scholars will refuse any form of evaluation as being representative of a relation they reject between individuals and the State; fear, based on real concerns about what is coming out with the evaluation.
- Managers (and decision makers) don't like it because there is a lot of disagreement amongst scholars and no policy maker is happy with an exercise that creates tensions in a field which he/she needs to deal with

- Research Evaluation don't know how to do it, because of the shortcomings of bibliometrics, the problems of peer review and the diversity of the SSH.

Shortcomings of bibliometrics

Ill adapted to the SSH

It has been repeatedly proved that bibliometrics are ill adapted to the SSH:

- specificities of Lotka's distribution; it says that out of a population of scholars, you will have a smaller proportion publishing 3, 4 or 5 articles than the proportion of scholars publishing just one article. We have some data from Italy where it appears that this needs field adaptation when applied to the SSH.
- Also, the Bradford's law is not working because "no core literature in a field can be identified" (Nederhof et al., 1989). Journals in the field are classified in three tiers: the first is the one who will publish the most important papers in a discipline; the other papers, less important, less visible will be published in a second tiers; the third tiers will be with everything else. This law is useful when you are trying to spot a core group of journals where the most innovative ideas from a discipline are published, but as it happens, nobody ever managed to find a core group of journals even in English Language and Literature for example.
- SSH publications are poorly covered in major international databases (WOS, Scopus). There is an under coverage of SSH scholar published books and the coverage of journals is biased with regards to the "language, country, publisher size and age" (Hicks, 2011); you have much more chances of being in WOS if you are an ancient journal than if you are a young one.

War on JIF

Tensions about the Journal Impact Factor concerns all the sciences because:

- It concentrates on journal articles to the detriment of a much more diversified research output landscape
- It is a very approximate proxy of quality: it is not because it is published in a prestigious journal that it will be excellent science. Students in MIT fabricated a scientific paper with a copy and paste method and they had it accepted in a prestigious journal. It is not a guarantee of quality.
- We also know that it is conductive to multiple controversial behaviours (parroting (researchers repeating the same kind of research, sending very little differentiated papers to different journals based on the same research and the same funding, just presented otherwise, to boost their impact and citation factors), psittacism (a sort of "I quote myself"), parochialism "I quote my dear friend X and they will do the same in turn", etc.).
- It under evaluates and under represents the outreach of a publication. This has been demonstrated since we have altmetrics, this is the publication of open access that allowed us to see that there are other parameters for impact: views, downloads, shares, comments on social medias.

Bibliometrics applied to the SSH

Czech example: Malek et al. "System of evaluation of research institutions in the Czech Republic" (2014). It is a performance based research evaluation system based on points that determine money allocated to a university out of the points they can collect over a given period of time. 61% if you are publishing something in a journal in WOS, books give only 9% of the points. The rationale behind it is that a paper indexed in the Web of Science has a minimum of 10 points but if you look at the formula it is multiplied by a factor. So the idea is that one paper published can easily bring 300 points while if you publish a paper in SSH journals listed in ERIH, it can bring only 30 points, so there is difference 1 to 10 between the two. So it is incitative to publish papers in certain journals than books, even if they are more adapted to the research conducted.

Problems of peer review

- Blind or not, prior to publication peer review may be anti-innovative and can lead to gatekeeping; we have numerous cases of blocking innovation through peer review
- In small countries or disciplines, the pool of evaluators may not be sufficient and this happens easily in Europe because we have a lot of *demographically* small countries.
- Better to it internationally, but criteria and expectations are not the same. A very distinguished Czech professor evaluating a piece of research in French will have the same criteria as an English one? Does the title of professor mean the same thing in all European countries? Even the perimeter of the SSH is not the same when looking at different European countries.
- It is time consuming and the cost may exceed benefits when you put into place huge campaigns of evaluation for an entire system or a discipline.

SHS specific biases

- There is a lack of transparency about the methods and criteria, the selection of reviewers, the treatment of conflicts of interest
- There is a low degree of organisation and quality control over peer reviews
- And there are acute intra and interdisciplinary conflicts about quality.

Survey

We organized a survey in 2014-2015 about peer review in (prestigious) publishing houses, within a project subsidized by ANVUR agency for research evaluation in Italy.
Amongst the questions asked:

- Has the PH a scientific committee assessing the book proposals?
- Has the PH a blind review system?
- Does the PH provide referees with an assessment sheet or guidelines for the evaluation of book proposals?
- Does the PH reject negatively-reviewed book proposals or asks for revision that take into account the reviewers' reports?

We selected publishing houses with specialized series in philosophy, history, literature, languages and linguistics.

- More than 250 publishing houses contacted (100 in Slavic area, 96 in UK and USA, 61 in Italy) => 54 answers
- Up to 9 reminders, high level of opt-out for numerous questions.
- Italian PH: 25% declare not having a scientific committee; more than 33% do not practice blind peer-review; when a peer-review is in place, 35% affirm not using an

assessment sheet as a guidance for peer –reviewers; only 2 PH communicated their assessment sheet.

Peer review in SSH journals and publishing houses:

- There are huge discrepancies with regards to review practices (length, argumentation, style) from one discipline to another. This is a very difficult information to access but we manage to build out a small corpus with the reviews we had in hand from participants to the project. Reviews are sometimes done in one word ("bof" as a peer review evaluation for proceedings to an international conference, see Anne Baillot's paper, Peer review says "bof"). On the other hand, we had 10 pages of observations over an article of 15 pages. So, differences are really huge.
- National incentives and authority involvement are needed to gather a more accurate picture
- When such national incentive exists, it leads to interesting initiatives, for instance the Flemish initiative in Belgium: they decided that the situation is not acceptable, that they are willing to accept books for evaluation but only books published in publishing houses having thorough peer review procedures; whatever procedures, but they have to be thorough, easy to monitor and demonstrate that a certain peer review took place at a certain moment. And they award a label to publishing houses that put into place such mechanisms.

Example of the peer review in the French assessment exercise:

- analysis of 104 reports of evaluation of SSH research units (all research units in two regions, Bretagne and Rhône-Alpes)
- evaluated period: 2004-2008
- conducted using corpus linguistics methodology and tools (Atlas.Ti and AntConc)

Question => How do official criteria for quality are translated into this reports?

Official criteria (AERES)

The words associated with good research are:

- New (original, breaking through, generates new patents, methods, norms, etc.)
- Partenarial (multidisciplinarity is encouraged, as well as extra-academic cooperation).
- Impactful (in the academic community: citation indexes, number of thesis, etc.)
- Useful (to the economy; to the society)
- Recognised (by peers: publications, selection as speaker, leadership, membership; by others: expertise, rewards)

Also, good research in SSH is published in certain journals (« périmètre de scientificité »). It is not really clear if it is an added criterion, a specification, or the only criterion of quality. But what is [scientific] quality?
In practice, reports shows that research appreciated as being good is not individual, the group research must have a thematic coherence. So, in France, individuals work under three constraints from:

- The institution, because they are strongly invited to be a member of an established research group;
- The Research unit, where people are incited to be in conformity with the group
- The discipline, since it is necessary to conform both in choice of research group and research production to the expectations of the CNU (Conseil National des Univesrsités, national council of universities)

3rd Challenge: Diversity of SSH

SSH is a general umbrella for a very contrasted landscape with regards to the publishing habits and underlying representation of quality. Traditional classification in STEM and SSH disciplines are not verified when we are looking at the publication habits.

- Mutz et al. "Types of research output profiles: A multilevel latent class analysis of the Austrian Science Fund's final project report data", 2013. Latent class grouping of publications: it is not unexpected but you have new insight, for example economy is closer to computer science and to mathematics, whereas everybody puts economics into the SSH. So what are we talking about when we designate 'the SSH'? => We are putting in the same basket very different kinds of fruits.
- Another project at ETH Zurich, funded by CRUS (Rector's conference) in Switzerland in 2012. They conducted repertory grid interviews with 21 scholars from 3 disciplines: German literature studies, English literature studies and art history. They observed many of the discrepancies, differences in view mentioned above.
- Ochsner et al. "Four types of research in the humanities: Setting the stage for research quality criteria in the humanities", 2013: 4 types in the SSH in terms of perception. In terms of perception of quality, 7 scholars and we have 4 representations of what quality is.

COST Action 15137: ENRESSH

European Network for Research Evaluation in Social Sciences and Humanities

- Started April 2016 > End March 2020
- NOT a research project: coordination of existing research
- 33 European countries involved and observers from South Africa, Mexico, Moldova

Objectives

- improve the understanding of how SSH fields generate knowledge; because basis of an evaluation is to know what scholars mean, what are the pathways towards producing something.
- to observe what kind of scientific and societal interactions characterize SSH;
- to observe patterns of dissemination and quality representations in the SSH.

Work groups

- WG1: Conceptual frameworks for SSH research evaluation
- WG2: Societal impact and relevance of the SSH research
- WG3: Databases and uses of data for understanding, monitoring and evaluating SSH research
- WG4: Dissemination
- A transversal special interest group for early stage researchers

Gathering data about SSH research

Some people say that SSH scholars publish books. Ok but is it true? Because the evidence so far gathered shows that the SSH scholars start to publish more journal articles than books.

Therefore, we look at:

- Full bibliographical coverage: we want to have visibility of publications not indexed in WoS or Scopus
- Not for citation counts: monitoring and understanding the system
- Our focus is on metadata rather than on datasets and publications.
- Successful development in countries where full coverage is part of a funding related evaluation system: eg. Norway and Belgium (Flanders)

Countries where data are gathered about SSH research:

- Belgium (VABB-SHW)
- Scandinavia: Norway (CRISTin), Sweden, Denmark, Finland (KOTA)
- Czech Republic
- France (RIBAC: only for CNRS researchers, not counting the universities (2/3))
- Italy (CINECA)
- Lithuania
- Latvia
- Poland
- Portugal
- Spain
- Switzerland
- UK (RIN)

Important differences of coverage, methods, categories.

The leader in the field is Norway with CRISTin: Current Research Information System In Norway. They have given the possibility to the scholars of declaring every kind of output, not only WOS. CRISTin: principles behind the use of institutional data in a national information system:

- Completeness: All scholarly publications should be included
- Transparency: Every institution can see and check all other institutions' data. The national database is also online and open to society at large.
- Participation: The indicator is developed and maintained in collaboration between the institutions and the authorities
- Multiple use of the data: CV's, applications, evaluations, annual reports, internal administration, bibliography for Open Archives, links to full text, etc. An important point is that data are imported (from ISI) and they can add data (about other publications). Scholars themselves can check if the number of outputs, not only publications, is correct and add whatever is not there.

Another interesting initiative is VABB-SHW from Belgium.
Engels et al., “Changing publication patterns in the Social Sciences and Humanities, 2000–2009”, 2012

- Creation of an “authoritative panel” to select publications (other than indexed in WoS) to be covered by the database
- Five categories of outputs: different coverage and philosophy (a) articles in journals; (b) books as author; (c) books as editor; (d) articles or chapters in books; (e) proceedings papers that are not part of special issues of journals or edited books
- Four conditions: (a) be publicly accessible, not necessarily in open access (b) be unambiguously identifiable by ISBN or ISSN number; (c) make a contribution to the development of new insights or to applications resulting from these insights; (d) have been subjected—prior to publication—to a demonstrable peer review process by scholars who are experts in the (sub)field to which the publication belongs. Peer review should be done by an editorial board, a permanent reading committee, external referees or else by a combination of these.

Challenges

There are countries where full coverage starts to exist, but issues are still faced..

- Gaining sufficient political support and funding for achieving systematic data collection in all European countries
- Interoperate RIS, in spite of differences in scope, degree of exhaustivity, typology

Beyond publications

- Criterium of societal impact brought to the fore the question of how to document engagement with society.
- What place for research data?
- What about submitted/ funded research projects as an indicator of activity and excellence? They are proof about scientific activity, even if rejected, time spent in the preparation should be taken into account.
- Flanders: published papers > 4p.
- Lithuania: book = 40000 characters * field coefficient (SSH=8); if you publish 10 pages less, then it is not a book!

Incomplete reporting and auto-censorship related to different factors:

- Technical barriers: HAL, RIBAC vs. Research gate, Academia
- But incomplete coverage in RG, Academia, Google scholars, etc. (almost same coverage biases as WoS)
- Increase of researcher’s workload (double, triple declaration), not interoperable
- Lack of institutional incentives
- On-line CVs: exclusion of “not prestigious enough” outputs

Various typologies

- euroCRIS (CERIF specification):
- comprehensive, but incomplete, for example prosopography, footnotes, glossary; excavation report (a very specific document in archeology) as "report"?

- debatable: PhD Thesis/ doctoral thesis authored book (what is an ‘unauthored book’?)/ monograph (in France it would be a book written by one author but in the UK can be a book written by up to three authors) Book = ISBN or not? If you are an African scholar and you want to publish in Ghana, there is no ISBN.

Beyond typologies:

- Genre analysis reveals huge discrepancies between products from the same category: i. e. bilingual abstract and keywords are NOT an universal feature
- Quotation and bibliographical habits are not the same in the various SSH disciplines (“art of the footnote”)
- Absent metadata: eg. institutional subsidies and their uses. This is also a potential indication of quality.

SSH evaluation and the open access

Positive

- Stimulates the production of metadata (with the above-mentioned problems of standardisation/ mapping)
- Modifies research practices (more cooperation: intra, inter and international cooperation; more articles than books) > changes the symbolic weight of outputs, and even types of outputs to be taken into account in the evaluation. Open publication result in more cooperation, it is linked.
- Changes some evaluations habits (post-publication evaluation)
- Imposes new metrics (altmetrics rather than JIF) but open access still has a limited influence because perceptions remain biased towards hard copies of books (see interviews conducted in IMPRESSH project, France, 2013)

Less desirable effects

- Large offer of “predatory open access journals”
  - Stimulates fake productivity
  - Lowers quality checking (“we publish within a week”)
- Costs of open access
  - puncture already limited budgets;
  - pay capability vs. quality?
- A model to be found for books, proceedings and chapter of books.

Contacts

Ioana Galleron is a Senior lecturer in French language and literature. Her research interests are the French theater of the 17th and 18th century, as well as the evaluation of research in the SSH. She is involved in several projects of electronic edition of plays (see http://www.licorn-research.fr/Boissy.html), and in a research group of the consortium CAHIER, dedicated to computer-assisted literary analysis. Since April 2016, she is the Chair of the COST Action CA15137.

Institutional websites: www.enressh.eu & www.evalhum.eu
Twitter @Enressh & @IoanaGalleron
Email: ioana.galleron@evalhum.eu

6-Case Studies

OpenEdition: Towards a European infrastructure for open access publication in humanities and social sciences

Pierre Mounier, EHESS & OpenEdition, France

OpenEdition

OpenEdition is a public infrastructure based in France since 1999. It is dedicated to SSH open access publication and scholarly communication. We are supported by 4 higher education and research institutions: CNRS, Aix-Marseille University, EHESS and Avignon University. It is exclusively dedicated to the dissemination of Humanities and Social Sciences research in *open access*. We think that dissemination on the web, not only on the Internet, is really important because it is a way to encourage and foster uptake for this kind of content in the different communities, scientific or not.
We are based in France but we are working to make our platforms and infrastructure more international, so we are working with an international network of partners in different European countries and even outside Europe. For example we have a partnership with Torino University in Italy to develop a specific program: OpenEdition Italia. We also have a partnership in Lisbon for the development of Portuguese content; another one in Germany with the Max Weber Stiftung; in Spain with the Uned University; and we also have partnerships outside Europe: In Canada with the Public Knowledge Project (PKP) which develops the Open Journal System (OJS) and in the US.

OpenEdition is an infrastructure and also a portal: openedition.org. This portal gives access to four platforms. We created and designed a platform for each type of document or information we want to disseminate:

1. Revues.org: it was the first platform created in 1999 and it is dedicated to journals. On this platform, we disseminate many open access journals from different countries and from all disciplines of Social Sciences and Humanities: history, philosophy, sociology, anthropology, geography, etc. Even though we are based in France, we are working with scientific committees in the different countries who want to disseminate their journals in any language on the platform. So the platform is multilingual: French, Italian, German, English, Spanish and Portuguese.
2. OpenEdition Books: It is a platform we have set up four years ago to help academic publishers disseminate open access their catalogue of academic books. It is multilingual too and at the moment we are working with approximately 70 publishers.
3. Hypotheses.org: This platform is dedicated to academic blogging. It is very different from the previous platforms where the disseminated materials are peer-reviewed. Hypotheses.org is for direct communication -there is no peer-review- but it is still interesting because researchers, librarians, research teams, institutions, projects, laboratories can disseminate information about what they are doing. It is completely different from the publication of a book for example, it is in fact the result of several years of work before publication. The content of the book is certified to be good quality, but it publicly appears many years after the research has been done. With a blog, it is the opposite, as soon as a researcher or a research team carry on a

research, then they can disseminate their information by themselves about what they are doing at the very moment they are doing it. So access to information is much quicker and more direct. Plus, as it is blogging, you have a commentary function, so the readers can directly comment on the blog and then you can have interactivity, some sort of a scientific discussion between authors of the blog and the readers. It is actually really complementary to the two firsts platforms.

4. Calenda.org: it is simple but really useful, it is a scientific agenda for SSH. It means that every organiser of a scientific event (workshop, seminar, conference, call for paper, etc.) can disseminate on this platform the information. DARIAH is one of the main partners of the platform, for example all events organised by and though DARIAH are disseminated as well on this platform.

Some figures

Revues.org hosts at the moment 437 journals, 100 600 documents on open access. There is a wide diversity in terms of languages, represented countries and disciplines.
For the books, as it is a younger platform, we disseminate at the moment around 3 000 books and 2 500 are open access. It is growing rapidly as we are going to work with more and more publishers.
For hypotheses.org, when we opened the platform in 2008, it was a bet because we didn't know if there would be any uptake in the community around this form of new communication. We had some professors saying that it would never work because "*blogs are for kids or teenagers*", "*not for academic content because it is not peer reviewed*". But we open the platform to see how it goes and it was a very good surprise because we now have more that 1 600 living blogs producing or having published 15 200 posts in open access. So, the uptake was enormous in fact and the most interesting thing about this platform is that the scientific community invented its own usage of the platform - we didn't anticipate it. For example, we didn't think people organising a seminar could use it for a seminar to announce new sessions, to record and spread sessions, to publish the bibliography of the seminar, to continue the discussion on the blog, etc. And it is the same with research projects, some ERC and ANR projects opened a blog to disseminate information about their project. Some of them opened a blog even before the project started. You can easily open a blog and start publicising information about an idea, maybe gathering new partners and then applying for funding. It is not compulsory to have funding to open a blog, you can open it before and it can be a good element on your file when you apply for fund.

Overview of visibility gained through the platforms for the disseminated content

This year, we are going to have a little bit more than 60 millions visits on the platforms, that is approximately 30 millions unique visitors. Those platforms are also meant not to be read and used only by scientific community and specialised scholars, but also by citizens, society at large. There are a lot of example of how those contents, like a chapter on a very narrow focused research monograph, are shared on Facebook or cited on twitter by academics and non-academics. This is the point of open access and of the web, because it can help improve and foster up taking by the community of this kind of content. The readership is worldwide in fact and disseminated in many countries.

Digital edition

What is *digital edition* and what is the difference between a *digital* and a *digitised* edition? For us at OpenEdition:
Digital publishing means to publish on the web, not putting some pdf files on a server, but to *convert those files and content and put it on a nice layout to be read and accessed in a web browser*. The goal is give access to a very complex information stored inside SSH publication. With an article, you have a lot metadata, a very complex text, highly structured, you have many information to display, so it is a real work to have it and to present it in a nice and usable way on a web page. As a platform, a digital edition is to give access to the same text in many different formats because we know that the readers are disseminated around these formats. Some people prefer to have a pdf file, other prefer web pages or epub format, so we do it and people can have their content on reading devices, softwares, smartphones, etc. To produce a digital edition, you also need a digital publishing way of working: you also have to work particularly with giving access to the data stored inside a text. For example, the images which are accessible in a web page are also accessible in full resolution with metadata attached. But you can also embed a flash file, for example an interactive map, so the reader can show or hide different levels of information on this specific figure, or even to embed videos.

Collaboration inside the DH community

To carry on this publishing functionality, we have created and developed **Lodel**. It stands for LOgiciel D'édition ÉLectronique and it is published on Github as an open source publishing software, like OJS, but different. It is a CMS based on XML TEI. The content published on OpenEdition Books and on Revues.org is converted into TEI XML and stored into our databases. We can then automatically generate, from the XML, the HTML, ePub or pdf file. We use a TEI than what was previously presented because we do not encode primary material, we encode publications. We use a subset a different tags, not all tags from TEI, only a limited number. LODEL works by converting content. As we work in the humanities community, we know that most of the authors work with text processing softwares like Word or Libreoffice. So we help editors to structure the word file they receive using styles (titles, normal, citation, etc.). They structure the information into the word file according to our own schema and they can upload the word file onto the publishing server. Then LODEL transforms the word file into XML. After that, you have dissemination through different formats.
LODEL => Structuring the information of front, converting it into XML and publishing in different formats and into different channels.
LODEL takes into account all the specificities of the humanities content, so it is not like Wordpress or OJS, which are not specifically meant for SSH. LODEL is meant for this specific content and that is why we are inside the humanities community. For example, it generates identification for paragraphs, it manages footnotes and complex structuration of a text, it can manage bibliography, it can manage multiple indexes of keywords, geographical data, chronological data, hierarchical, ethnic populations, it is multilingual and attribute DOIs.
We are finally also part of the DH community as we publish the Journal of the Text Encoding Initiative.

R&D: OpenEdition Lab

- **Bilbo** is an automatic annotation software that can parse a text, detect and automatically recognise bibliographical references inside a text, analyse those references to divide between the name of the author, of the publication and then query the CrossRef database to see if there is DOI attached to this reference. At the end, Bilbo transforms a plain text bibliography, which is the norm in SSH, and then add a link to the bibliographical reference, so the bibliographic is no more a plain text but hyperlinked and it is a little bit more useful for the reader. We want to add this functionality into the footnotes and the next step will to try to detect fuzzy references inside the text to see if there is an online reference available to be linked to. This project received a Google Grant.
- **Agoraweb** is an automatic detection of book reviews. The idea is that inside publication and blogs, you have book reviews, authors writing articles or blog posts, which are in fact reviews of published books. In journals, it is easy to detect because there is a specific section "Book reviews". But in blogs, it is much more difficult to detect because a blogger never says "*this is a book review*". So you have to parse through whole blogs inside hypotheses.org to automatically detect if blog posts are book reviews or not. After that, with Bilbo we can see which book is reviewed and then make a link between the book and the reviews. The final aim of this project is to gather on OpenEdition Books the book by itself and at the bottom of the page all the available book reviews published in different venues, in journals, in blogs, or anywhere on the web. This is a really useful information for the reader. Demo website of the OpenEdition Reviews of Books: http://reviewofbooks.openeditionlab.org/
- **Open Peer Review**: an experiment proposed to journals which are traditionally [anonymously] reviewed: Julien Bordier, 2016. « Open peer review: from an experiment to a model: A narrative of an open peer review experimentation ». <hal-01302597>.

You can find further readings here: https://lab.hypotheses.org/bibliographie

OPERAS

Our very new initiative is Open access Publication in the European Research Area for the SSH (OPERAS). The idea with OPERA is to be able to work at European level. We try, with our partners in Europe, to extend this network and to make everybody work together to set up an infrastructure for open access publication in SSH. We defined inside the network some common goals, the idea is that the players in this field of open access publication in SSH are very small and fragmented, so we need to gather all those players to adopt common standards for example or to share research and development costs (because it is expensive). It is better to do it all together and to share the results, to identify and adopt best practices, to assess sustainable economic models and to advocate for open access in SSH. For now, we are 19 partners from 10 countries (Germany, UK, Netherlands, Spain, Portugal, Italy, Croatia, Luxembourg, Greece and France) and we are open to extend this network to other countries. Our first partner is the Association of European University Presses (AEUP): an association that gathers main University Presses at European level. The idea is to set up a cyber infrastructure for open access publication and more specifically for books, because in SSH most of the books are not in open access and not even digital. So there is a lot of work to make books *accessible, disseminated, visible, indexed and used on the open way*.

HIRMEOS

This project's name stands for High Integration of Research Monograph in the European research area for SSH (HIRMEOS). It has been submitted last year to the Horizon 2020 framework and it is going to start in January 2017. The idea is to gather 5 open access book publishing platforms (OpenEdition Books, Ubiquity Press, OAPEN, The Gottingen University Press and EKT platform) and to implement 5 sets of service at the same time:

- Identification at all level on 5 book platforms: DOIs, ORCID identifiers for the authors and FUNDREF to ease indexing to OpenAIRE.
- Certification of quality provided by the Directory of Open Access Books (DOAB).
- Implement automatic recognition service on named entities inside the full text of a book for places, names, periods, dates and maybe topics. Then it will be given back to the platforms to enrich their index or to link to DBpedia for example.
- Implementation of an open annotation service: the reader will be able to annotate line by line the full text of the books in open access. The reader will also be able to answer to the annotation with a *forum feature* inside the annotation service: when someone annotates a line or a sentence, then it is published on the web and someone else can answer to this annotation and the author of the annotation can answer to answers and so on. The idea here is to rise the uptake of this content by developing conversational features around the books, giving possibility to the reader to discuss about a book on the same website where the book is disseminated.
- Develop book metrics: implementation of Altmetrics for the books. Altmetrics are a new impact metric, invented by PLOS, to measure the impact of an article by counting not only the number of downloads or views on a platform, but also the numbers of shares on the social media such as Facebook and Twitter, in order to aggregate that and to make a new metric to measure impact. Annotations will be counted by the Altmetrics for the book metrics, because the number of annotations attached to a book is also an impact measure.

That is why we work on this project with different partners: for example we work with ORCID to attribute ORCID IDs, with CrossRef for the DOIs, with Hypothesis for the annotation plugin service, but also with Huma-Num, with DARIAH, with OpenAire for indexing our content, etc.

Contact

Pierre Mounier is deputy director of OpenEdition, a comprehensive infrastructure based in France for open access publication and communication in the humanities and social sciences. OpenEdition offers several platforms for journals, scientific announcements, academic blogs, and, finally, books, in different languages and from different countries. Pierre teaches digital humanities at the EHESS in Paris. He has published several books about the social and political impact of ICT, digital publishing and digital humanities.

Associate Director for international development OpenEdition
Coordinator of OPERAS: http://operas.hypotheses.org
ORCID: http://orcid.org/0000-0003-0691-6063
Twitter: @piotrr70

Email: pierre.mounier@openedition.org

Czech Literary Bibliography

Vojtěch Malínek, Institute of Czech Literature, Czech Academy of Sciences, Czech Republic

The Czech Literary Bibliography is a basic infrastructure for interdisciplinary research into literary culture of Czech lands with tradition since 1947. It is one of the largest and more opened research infrastructures for individual national literatures in Central Europe. All of our data are fully available online. We are supported by the Ministry of Education, Youth and Sports of Czech Republic and we are now included in the Czech Republic Roadmap of Large Infrastructures for Research, Experimental Development and Innovation. The hosting institution is the Institute of Czech Literature of the Czech Academy of Sciences.

The information sources we are processing are:

- set of bibliographical and other specialized databases (biographical base Czech Literary Figures, literary prices, book editions, etc.)
- documents collections and card index catalogues

Basic numbers

- over 2 000 000 bibliographical records (articles, conference proceedings, books, etc.)
- nearly 40 000 biographical entries about authors and literary scientists
- over 1 500 newspaper and journal titles processed
- data instantly available without any limitation online

Chronological range

Our bases cover the whole modern Czech literature, which means from 1770 to present, and data are continuously added and updated. Processed documents are mainly in Czech of course, but also in German, Slavonic languages, etc.

Range of disciplines

- Czech literature and literary studies
- other national literatures and associated humanities and social science disciplines (theatre studies, history, philosophy, linguistics, journalism, etc.)

Geographical scope

Concentrated on Czech lands, but we are working with bohemical literature published abroad, in Europe and USA for example.

Average usage rate

150–200 accesses per day

Main activities

- Processing of databases
- Digitisation and software development

Retrospective Bibliography of Czech Literature□

It is a large card index compiled from the 1950s to the 1990s that covers a chronological range from 1770 to 1945. This card catalogue consist of nearly 1.7 million records, with 530 titles of newspapers and journals mainly in Czech and German. Thematically, articles of following types are processed: Czech and world fiction, translations, journalism and specialist literature. The catalogue is organised by authors, reference, subject and identification section.

RETROBI software: http://retrobi.ucl.cas.cz/

During the “Retrospective Bibliography of Czech Literature 1770–1945 card index catalogue digitization project“ (2010-2012), we have developed a software for digitization and online presentation of card indexes. We have scanned and prepared OCR transcriptions of all of the cards. This allows full-text searches in text representation. We have also developed a tool for editing of text data available: it means that data are available for any user for corrections and editing (example). Cards could be corrected or rewritten as a whole or semi-automatically structured. For registered and skilled users, tools for large-scale editing of a chosen database field are available. RETROBI system enables complex queries, variable export options and offers administrator interface with several advanced functions, etc. Since 2012 it was used for digitising of 3 others large card catalogues in different institutes of CAS. RETROBI software was created with the kind support of the Ministry of Education, Youth and Sports as the result of the “Retrospective Bibliography of Czech Literature 1770–1945 card index catalogue digitization project“ (VZ09004), completed between 2009–2011 under the INFOZ programme.

Contemporary Bibliography (since 1945)

The Contemporary Bibliography is completely processed in database form. We are now finishing the conversion of older datasets into nowadays standards. So it now meets common standards for contemporary librarianship and bibliography (MARC21 exchange format, RDA cataloguing rules, Aleph software etc.). Everything is fully integrated to the national research information exchange networks. We are using various persistent identifiers.

Conclusion

This quite strongly structured data can be used for statistical and quantitative analysis of the literary field defined by the needs of researchers with regards for analysis of bibliographic data.
For example, you can search for the total number of records per year, it can put light on the way how political conditions can influence the number of records pro author, magazine etc.: after the of the Second War, the total number of records rapidly increased but after communists take power in 1948, the number significantly falls down, with its lower level in 1952.
You can make comparison of selected authors and analyse its connections to social contexts and measure the impact on their ranking with bibliographical data used in a quantitative and statistical way.
On the background of the data from RETROBI, the Top 10 Authors with highest number of records for the period 1770-1945 can be easily shown. The most “productive” of them was Arne Novák Arne with 7 370 occurrences.

Contact

Vojtěch Malínek, Institute of Czech Literature of the Czech Academy of Sciences

Institutional website: http://clb.ucl.cas.cz/

Facebook: http://www.facebook.com/ceskabibliografie
Email: malinek@ucl.cas.cz & clb@ucl.cas.cz

Turning the Polish Literary Bibliography into a Research Tool: Challenges, Standards, Interoperability

Maciej Maryl & **Piotr Wciślik** (Institute of Literary Research of the Polish Academy of Sciences)

Polish Literary Bibliography

The project we are currently working on is quite similar to the one described by Vojtech Malinek, Czech Academy of Sciences. Let me start with the use of the data, i.e. with some examples of research we would like to conduct on our data once the project is completed: The Polish Literary Bibliography – a knowledge lab on contemporary Polish culture.

Data-driven research into literary culture

The goal of our project is to make our bibliography not only easily accessible and searchable but also capable of serving as a discovery tool for researchers. We would like to enable them to perform similar tasks as in those studies which used rich bibliographical data for exploration of literary culture:

- Analysis of literary locations, e.g. comparison of the novels set in rural areas with the urban ones over time (see: Jockers 2013:45).
- Gender distribution of authors as compared with their coverage in the media. Katherine Bode has shown that since the 1990s more novels are being published by female authors, yet both critics and academic scholarship focuses more on male novelists. (Bode 2014: 132-134).
- Tracing co-publishing patterns as the source of the knowledge about the shape of literary networks. You can compare and interpret various affiliation networks from other countries (Long and So 2013 a: 150; 2013 b: 274).)
- Franco Moretti's analysis of the length of the titles in Victorian novels shows how the genre became standardised (Moretti 2013: 183). Moretti also traced the patterns in a genre lifecycle on the example of Victorian novels (Moretti 2005 15-19)
- Some existing bibliographical infrastructures already make some research tools available like exporting queries into csv format for further analysis (Australian Literary Bibliography) or tracing relationship between authors in the form of a network as in the case of Women Writers (NEWW).

In order to perform such research, we need reliable data, but our problem is that bibliographical data was collected not for research purposes but as reference material. So the data is prepared in a way that allows people better access to knowledge rather than statistical inferences. For instance, lots of information is preserved in unstructured annotations, which are difficult to be transformed into a database. Our challenge is to work with data collected as a reference resource, not research material□with the goal of forging a database ready to answer unpredicted questions, i.e. questions which were not taken into account when collecting the material.

PBL: Polish Literary Bibliography

The PBL is an annotated bibliography of Polish literature that has been published since 1954 as a project that is a comprehensive register of all articles, notes and other materials concerning Polish literary life. This can include different types of records: authors, works, reviews, articles, contests and awards, TV, radio, theatre and cinema adaptations, events, exhibitions and semantic annotations (e.g. ’literary theory’; ’exile literature’; ’psychology’; ’education’; ’censorship’), etc.
The Polish Literary Bibliography as a research tool on contemporary Polish culture is a□3-year project funded by the National Programme for the Development of Humanities (2015-2018) and is developed by Institute of Literary Research, PAS in cooperation with the IT partner: the Poznań Supercomputing and Networking Center (PSNC).

Goals

- Creating a database of 4 million bibliographic records on Polish literature (1939-2004)
- Integration of heterogeneous datasets (retrodigitisation of available data in printed format)
- Integration with Linked Open Data Cloud
- Data mining and visualisation tools for modelling processes of literary life□□

Challenges of ontology and retrodigitisation

Ontology
We apply schema.org ontology (with its BIB extension) in order to connect our data with Linked Open Data cloud. We would like to offer rich searching features that enrich an answer with additional information next to the results (cf. Google rich snippets). So we have to make connections between existing data and to retrieve automatically such information from the LOD cloud.

- Choosing the right data standard: Schema.org instead of bib-dedicated ontologies and data models (FRBR, RDA, BIBFRAME) because
    - it is not well equipped to handle theatre, cinematographic, radio and television instances of literary works (except FRBRoo) (=different events in the database, different ontologies)□
    - It is too complex to handle by metadata producers (including FRBRoo)□
    - Validity still TBC by community of practice

- Whereas Schema.org offers
    - a widely-used Internet standard
    - a robust model to handle PBL data model
    - not unprecedented in the library domain, there is a bibliographical extension with a careful list of ontologies elements and relevant vocabulary.

Process:

- Remediating the input tool (and habits of teams working with previous database)
- Converting the existing online database (1988-2002)
- Retrodigitising printed volumes (1944-1987; WW II; Bibliography of Samizdat).
- Integration with other bibliographies: we are figuring out how to map our entities, our bibliographical system onto schema.org ontology in order to ensure further cooperation with other national or regional bibliographies

Retrodigitisation of a huge collection in regard with the structure of the database
Challenges:

- Technology: Parsing & Lemmatization
- Tracing methodological inconsistencies
- Time factor
    - Subject-classification (new approaches to literary studies, e.g. gender studies)
    - new forms of literary life, e.g. online literary life
    - changing political geography: Yugoslavia and Soviet Union are no longer existing entities

Research challenges: local specificity

- Data-collection methodology has changed over the years
    - Selection of writers
    - Unrecorded debuts
    - Data censorship
- Geography
    - Changing borders
    - Literature in Exile
    - Polish literature = literature in Poland or in Polish?

- Shape of literary life: Official versus Underground publishing

Our overall goal is a conscious research into literary culture based on bibliographical data, which could be formulated as following: in order to come up with right research questions, one has to be sure to know how the data were selected, collected, censored, abridged, standardised, annotated, digitised, corrected and linked together.

References

Bode, Katherine. 2014. Reading by numbers. Recalibrating the literary field. London: Anthem Press.
Long, Hoyt and Richard So (2013a) ”Network Analysis and the Sociology of Modernism” boundary 2 40(2):147-182
Long, Hoyt and Richard So (2013b) ”Network Science and Literary History” Leonardo 3(46), 274-274.
Jockers, Matthew L. 2013 Macroanalysis. Digital Methods & Literary History, Chicago: Chicago UP.

McCarty, Willard. 2008. "Modeling in Literary Studies" A Companion to Digital Literary Studies, ed. Susan Schreibman and Ray Siemens. Oxford: Blackwell, http://www.digitalhumanities.org/companionDLS/ (28.02.2016)

Moretti, Franco 2013 "Style, Inc.: Reflections on 7,000 Titles" Distant Reading, London: Verso.

Moretti, Franco 2005 Graphs, maps, trees. Abstract Models for Literary History, London: Verso.

Pacek, Jarosław, 2010, Bibliografia w zmieniającym się środowisku informacyjnym, Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.

Shea, Christopher. 2008. "The geography of Irish-American lit" Brainiac [Blog], http://www.boston.com/bostonglobe/ideas/brainiac/2008/07/matthew_j_jocke.html (28.02.2016). □

Woźniak-Kasperek i Ochmański, red. 2009. Bibliografia : teoria, praktyka, dydaktyka : praca zbiorowa, Warszawa: Wydawnictwo Stowarzyszenia Bibliotekarzy Polskich.□□

Contact

Maciej Maryl & **Piotr Wciślik** (Institute of Literary Research of the Polish Academy of Sciences)

Institutional website: http://chc.ibl.waw.pl/en/projects/pbl-lab/
Personal website: maryl.org
Twitter: @maciejmaryl
Email: maciej.maryl@ibl.waw.pl

Creation of Open Data Resources: Benefits of Cooperation

Kira Kovalenko, Institute for Linguistic Studies (Russia) & Austrian Centre for Digital Humanities (Austria) & **Eveline Wandl-Vogt,** Austrian Centre for Digital Humanities, (Austria)

I am a research fellow at the Institute for Linguistic Studies in Saint Petersburg and an invited researcher at the Austrian Academy of Science in Vienna, so I represent two organisation and I will be talking about cooperation between these institutions.

- Austrian Centre for Digital Humanities (ACDH), Austrian Academy of Sciences (Vienna), established in 2015: 1 department, 4 working groups, 50 researchers, director Dr. Karlheinz Mörth.
- Institute for Linguistic Studies (ILS), Russian Academy of Sciences, established in 1921: 6 departments, 120 researchers, director Prof. N. Kazansky

One of the biggest department of the Institute for Linguistic Studies is dedicated to lexicography. We create a lot of dictionaries, such as the dictionary of modern Russian language, the dictionary of the 18th and 19th century language and the Dictionary of Russian Dialects that started in 1965 and has so far 48 volumes published. It has more than 300 000 entries. The chief editor is prof. Sergey Myznikov, 8 compilers are working on it - and I am one of them. As a result of our discussion with Eveline Wandl-Vogt who is a member of the group compiling the Dictionary of Bavarian Dialects, we decided to combine our efforts on digitalisation of the dictionaries. Since then our plans for cooperation have enlarged, and now we have three common projects. The main aims of our cooperation are to:

- increase accessibility
- increase interoperability
- increase reusability
- enrich dictionary data
- interlink dictionary data
- create new workspaces
- open up dictionaries for research process and public curiosity

Projects

All started with the Dictionary of Russian Dialects and we decided to create an infrastructure for compilers and researchers. You can now see the dictionary *online* on the website of the institution and you will find almost all published volumes, but it is just a pdf format and of course we cannot correct the text or add new material to the volumes, etc. This is why we decided to create such infrastructure and we use TEI P5 to markup the dictionary. We also use Lemon model for searching information and for other technical issues we envisage Ontolex core model with extensions: *synsem*, *decomposition*, *variation* *and translation* and *linguistic metadata*. We are planning to have an infrastructure that we could add and from which we could extract what we need, automatically. It will be online, available for everybody and all users could have better access to the material.

Another project we are working on is the Russian Manuscript Lexicons infrastructure for researchers. Russian manuscript lexicons appeared as a new lexicographical genre in the middle of the 16th century and has been developed since that time. The infrastructure will include:

- alphabet arrangement according to their first apparition
- close connection to the [original] text
- more than 9 types
- from 700 up to 16 000 word entries
- foreign words from Greek, Latin, Hebrew, Church Slavonic, Ruthenian, Tatar, Arab, and German origin
- about 150 lexicons
- important source for historical lexicography

Approximately 15 of them can be found online on the National Library website, but in order to find something you have to look through a lot of pages and the process is very difficult if you need some particular words. During my PhD, I have manually copied some manuscripts representative for different types as a plain text, and now have them in text version; that is why it would be nice to have them in parallel text version and in facsimile version. We are planning to create such infrastructure where you could see this and then search necessary information. We are planning to use as well:

- TEI P5 to markup them
- cr_xq: a standards-based, fully configurable publication framework for XML data
- full-text search and field-specific searches
- synoptic view of facsimile and text
- computer-assisted collation and stemma creation
- facilitates creation of various indexes of tagged information in the ingested resources (lemma list, index of translations by language etc.)
- standards: METS, FCS-SRU, currently mostly used for TEI content (e.g. Austrian Baroque Corpus ABaC:us)

The last project is not started yet but it should begin next year, it is the Russian Plant Names in the Diachronic Aspect (from 11th to 17th centuries). It will be a database with a search engine. We applied for a grant and if we succeed, we would like to have such a database. We are a team of researchers from different background, we share interest for languages, folklore, literature, etc. We will use primary sources (manuscripts, printed books of the 11-17 cc.) and secondary sources (historical dictionaries, modern researches, etc.). So we are going to collect all this information and to create a database. We hope to have interoperability with the Austrian plant common names database (exploreAT!). It will also contribute to the project Biodiversity and Linguistic Diversity. In the end, it will be a collaborative Knowledge Discovery Environment. This database will include:

- Old Russian name
- modern Russian name

- foreign etymon (if loanword)
- type of representing a foreign phytonym (translation, transcription, transliteration, loan translation, hyperon)
- Latin name
- example of the use from text/dictionary
- bibliographical information (if printed: author, title, place of publication, date, page, genre; if manuscript: genre, author, name, location, page, etc.)
- use of the plant
- symbolical meaning

Then, it would be interesting to connect this database with international resources such as Europeana. You already can find some Russian names there, but not historical Russian plant names, so if we have such a database with material we have in manuscripts and old dictionaries included and enriched, it would be really interesting. It could help to have more modern and historical names represented in the international online resources.

Benefits of cooperation

- collaborative approach allows to establish sustainable workflows
- shared use of unified or de facto standards and infrastructures instead of starting creating new; develop new standards in a collaborative approach
- experimenting with new methods and emerging technologies
- gives a chance to open up new data
    - dictionaries
    - manuscripts (ease access to them)
    - lexical data
    - cultural data
- connection of data to the global resources (Europeana)
- more and better results for both cooperation partners
- sharing failures, risks and furthering learning and innovation
- improving competitiveness and visibility
- SHARING DIGITAL TRANSFORMATION AND INNOVATION

Contact

Kira Kovalenko, Institute for Linguistic Studies (Russia) & Austrian Centre for Digital Humanities (Austria) & **Eveline Wandl-Vogt,** Austrian Centre for Digital Humanities, (Austria)
Email: kira.kovalenko@gmail.com

Network of Dutch War Collections: pursuits and goals

Tessa Free, Netwerk Oorlogsbronnen, Netherlands

The Network of Dutch War Collections is a program of NIOD Institute for War, Holocaust and Genocide Studies, a research institute and a WWII-collection keeper in Amsterdam. The Network of Dutch War Collections is an independent program. It is facilitated by NIOD and a steering committee provides substantive direction. The goal on my organisation is to make scattered resources from and about the Netherlands in the Second World War digitally better findable and usable. In the Netherlands, there are approximately 400 institutes that keep documents about this period and there is a great diversity - from two papers to a couple of kilometers. Also, some documents are digitized and standardized metadated, and others are for example described on the personal computer of an almost retiring employee. We want to make 9 million sources findable and usable not directly for the public but firstly for the intermediaries: teachers, researchers, app or game builders, etcetera. So they can use the sources and reach the big public through their products (books, articles, classes, etcetera).

Scattered sources findable

An example: Westerbork memorial site was transit camp in the east of Holland. It is an important place of memory but a lot of documentation about it is scattered. Maps of the camp, extracts of diaries or movies are kept by different institutions and it is difficult to find all these documents in one place. We have different sources telling stories about one place but all kept by several institutes with different rules, different kind of publication, different possibilities to research it. We want to help all those different institutes by making the documents digitally better, findable and usable. For example, we can trace the story of a woman told by the different sources in different places.

Useful connected resources

By adding context, we connect resources. This is the most important thing for researchers (or other war source seekers) as it is the way you can tell the story of one person through all these different sources. Researchers are mostly looking for "*who*" (persons), "*what*" (themes), "*when*" (date) and "*where*" (places). So we are focussing on these questions to make the sources available through these four aspects. Besides making the resources better findable, we aim to make them better usable. By clearing property rights if possible [because some documents are copyrighted]. Or inform about these rights, so a researcher knows where to ask for publishing-permission.

Projects

- What: Build of a WW2 thesaurus implemented in different softwares
- Who: Personsportal, crowdsourcing
- Where: geocoded WW2 resources
- When: describing moments

R&D

- Open War Sources (Wikipedia)
- Pilotproject full-automatic access of the Central Archive of Special Legal Procedures (using OCR and NER (Name Entity Recognition) techniques)

Cultural change

Besides the projects, we work on explaining the importance of sharing, connecting and open publishing (if possible) to collection keepers. The Network keeps a War Collections-Portal where we harvest all the available resources. Nothing is hosted on the War Collections-website, it reflects information we harvest from the other institutes. Our work is sometimes uneasy to demonstrate because it is firstly a backside work and not visible at the first stages.

Contact

Tessa Free, Netwerk Oorlogsbronnen
Institutional website: http://oorlogsbronnen.nl/
Twitter: @tessafree or @oorlogsbronnen
Email: tessa.free@oorlogsbronnen.nl

Open Access Meets Productivity, "Scholarship, see effect of being an efficient source"

Adele Valeria Messina, University of Calabria, Italy

Case study: The Method of Online Academic Reviews and the Alleged Delay of post-Holocaust Sociology

At the beginning of this study there was a lot of confusion both about the use of EBSCO database and the delay of post-Holocaust Sociology. It was a research halfway and between Hemerographia and meta-Sociology.
This method has been chosen to verify this alleged delay of post-Holocaust Sociology in sociological literature through important indexes: the speed of publication and the scientific impact of research on academic public.

Beyond the digital research

How concretely open access answered to my own research question dealing with this sociological delay?
EBSCO hosts databases and open access to full text allowed to measure three important indexes:

- Productivity of sociologists
- (Their) Visibility
- (Their) Degree of Appreciation

It was then possible to address questions with regards with:

- how many written works a scholar has produced
- in which periods
- how many times the name of the authors appears in articles and reviews on EBSCO.

I also calculated the degree of appreciation in sociological works, thanks to the number of citations that academic environment has reserved for them. This method was important and useful because it allowed to verify this delay. "*Unknown papers emerged, unpublished reports and this fact cleared up doubts related to the question of the alleged delay of sociology*". By perusing the online academic sociological reviews (year by year) it was and it is possible to glance at and examine who promoted which research project and in which scientific reviews.

Three authors are special examples:

- Everett C. Hughes (1897-1983). It appears that Hughes wrote about the Holocaust in 1948, in a period very close to the World War II. In particular, he speaks about “banality of evil”, an important category or concept, that we know instead thanks to Hannah Arendt in the 1960s. And this fact emerged from articles, letters, personal writings found thanks to open access in full text.
- Talcott Parsons (1902-1979). He is a famous author in Sociology, but for his works about the destruction of the Jews is completely unnoticed.

- Paul Neurath (1911-2001). After the conflict he put into writing his personal camp experience, but at the end of the war no publishing house was able to publish his work.

They are good examples that highlight the importance of open data for SSH and how perusal of sociological reviews permitted to determine “who” promoted “which” research projects and in “which” scientific reviews. It was also possible to comprehend the rhythms and delays within the academic environment for political reasons and research funding. The academic reviews are important for the sociological field because it is based on and built by the scientific reviews and dissemination of works. The online academic reviews do not replace or supplant paper or printed journals but support them permitting to have a wider range of analysis:

- Without this scientific and scholarly literature available online, thanks to open access, I could not have verified this delay.
- And without the perusal of well-qualified online reviews, several sociological studies would have been forgotten.

The idea behind this work is to contribute to the digitization of *unknown documents and manuscripts* related to modern and contemporary history and critical thought and to address the necessity of their usage and employment into current research. One of the central goal is to host unnoticed texts in a semantic open access platform in order to support the circulation and connection of data.

Links between open access and Humanities

It might be in a text, defined as tool of democracy in the sense that it is an expression of ideas with a freedom of speech. It is possible to speak of the new so-called “digital democracy” with digital tools.
That is why it is important to share data beyond digital tools, to link knowledge, to highlight relationship, to cross the bridge, made of words and concepts, that exists between any written and its readers. Right, on the relevance of these bridges, made of paper, Alessandra Cambatzu has written a great essay.
Thus, any text speaks to reader. Any reader talks back to the text. And this meeting, written or digitized, is powerful.

Contact

Adele Valeria Messina, University of Calabria
Email: adelevaleria@gmail.com & avmessina.freeebrei@gmail.com

Case studies on digital content reuse in the context of Europeana cloud

Eliza Papaki, Digital Curation Unit, ATHENA R.C., Greece

The work to be presented here was developed in the context of the Europeana Cloud project between 2013 and 2016. The Digital Curation Unit, ATHENA R.C. was leading the work on "Assessing Researcher Needs in the Cloud and Ensuring Community Engagement".

Europeana cloud

Europeana Cloud was a 3-year project that started in February 2013 and ended in April 2016. Overview:

- Total Project Cost – 4.75m Euros
- EU Funding Contributing 3.8m Euros (80%)
- Matched Funding 950k (20k)
- co-funded by the CIP-ICT Policy Support Programme
- CIP-ICT-PSP-2012-6 - Project number 325091

Amongst its aims was to:

- build a cloud based infrastructure which would add new data to Europeana
- give solutions to content providers and aggregators to store, share and provide access to digital material in the area of cultural heritage more efficiently
- give researchers new services, tools with which they could access this work and share the content stored in the cloud - aiming for further strengthening its public impact.

Our work in Assessing Researcher Needs in the Cloud and Ensuring Community Engagement focused on:

- Developing an effective research content strategy for Europeana vis a vis Humanities and Social Sciences research communities and improving the understanding of digital tools, research processes and content throughout the research lifecycle.
- Engaging the Humanities and Social Sciences research communities in the use of Europeana as a valuable resource for Digital Humanities research.
- Managing Europeana Research, which aims at opening up cultural heritage content for use in research, by fostering collaborations between Europeana and the cultural heritage and research sector.

Methodology

In order to reach these aims we employed several methodological steps:

- Identification of research communities under the umbrella of SSH

- Web Survey to document the practices of researchers in Europe (tools, content and methods)
- Expert Forums which focused on the target audience of the project (documenting problems, gathering suggestions)
- Use cases on digital innovative tools (interviews were held with people employing tools, noting gaps and suggestions for further improvement)
- Focus Groups (discipline specific)
- Case Studies on particular research topics

Case studies:

Population Movement as a Result of Conflict in the 20th Century

This case study, which was conducted by Vicky Garnett of Trinity College Dublin, focused on conflicts of the 20th century:

- Greek/Turkish Conflict of 1920s
- Hungarian Revolution of 1956
- Yugoslav Wars of the 1990s

It imposed the following research questions:

- What was the international response to displaced population resulting from each conflict?
- How does this compare with the response to displaced populations in the 21st century?

Useful resources were found in newspapers, transcriptions and photographs but among the problems faced was that nothing could be found externally for the Greek and Turkish conflict and nothing online about the Yugoslav Wars.

Problematic resources within Europeana:

- Newspapers (at the time) were not properly searchable
- Metadata was often incomplete or inaccurate

Problems in accessing content beyond Europeana:

- Specifically for the Yugoslav Wars, the content is still sensitive and subject to
  - government embargoes
  - cultural sensitivity
  - lack of resources to maintain records

Considerations for Europeana

- This was a worthwhile study and there is much content beyond Europeana that can be potentially absorbed as part of a collection.
- Europeana needs to create stronger links to existing content and maintain that.
- Cultural sensitivities may be a problem in obtaining data, particularly for the more recent conflict of the 1990s

- Financial issues may also be a problem for maintaining records and content within local GLAMs.

Case study: Children's literature

This is a case study conducted by Eliza Papaki of Digital Curation Unit, ATHENA R.C. Definition of the topic: "*Children's Literature is (among many other things) a body of texts (in the widest senses of that word), an academic discipline, an educational and social tool, an international business and a cultural phenomenon*". Hunt, P. (2004). International Companion Encyclopedia of Children's Literature. Routledge.

Methodological steps:

- Matching Children's Literature to academic disciplines: History, Cultural History, Literature and Languages, Education, Library Studies, Philology, Textual Studies, Linguistics and Media Studies.
- Searching Europeana for related content resulted in: Text (1596) / Image (194) / Video (25) records ranging chronologically from 1450 to 2014 and geographically from all over Europe, mainly the United Kingdom.

Diverse content retrieved in Europeana can be considered as brainstorm of results.

Suggestions for promoting digital content reuse within Europeana

1. Enrichment of content in the topic of Children's Literature, even through potential collaborations with other digital libraries:
    - Gap in geographic coverage of available digital material
    - Collections as means of organising content
    - Easy access to references, databases, journals and books
    - Map of publications
2. Mobilising already existing associations of this area in respect to new digital resources, tools and services will increase the number of potential users and the community.
3. Further development of digital tools and services on such content gathers great attention among researchers who still employ more traditional, non-digital approaches in their research.

Findings from Europeana Cloud

- Different research disciplines use different types of data in different ways
- Data aggregation horizontal rather than vertical: Not useful for high-level, advanced research, but very useful for teaching
- Need for collection-level descriptions
- Need for user-friendly tools and services which will enable re-use of Europeana data

What is still needed

- Better understanding of how content and metadata is actually used, and its relationship with digital methods and tools

- Targeted outreach and engagement methods
- Empirical understanding of technical tools that will increase use of content and metadata.

Europeana Research

"*Europeana Research will help open up cultural heritage content for use in cutting-edge research. It will run a series of activities to enhance and increase the use of Europeana data for research, and develop the content, capacity and impact of Europeana, by fostering collaborations between Europeana and the cultural heritage and research sector. It will provide an important focus for the emerging communities of practice who rely on Europeana for their research, and support the European investment in digital cultural heritage*".

Europeana Research is currently an ongoing project continuing and further enriching work conducted within Europeana Cloud in improving Europeana for research use and by enhancing the network of the research community.

Contact

Eliza Papaki, Digital Curation Unit, ATHENA R.C.
Institutional website: http://research.europeana.eu
Twitter: @eurresearch & @DigCurationUnit
Email: e.papaki@dcu.gr

7-Data Journals & Editorialization of Open Data

Data Journals

Anne Baillot & **Marie Puren**, INRIA

What is a data journal?

Why is it called “journal” and to what extent is it different from a traditional journal? Why do we need data journals? What is advantage of publishing it in the traditional structure of a journal? => You can get some credit for it

This session will not be too technical but will try to reflect on what it means to have the opportunity to publish data papers, to construct data journals, what it means for the academic system in terms of recognition for digital research and for academic communication in general.

For the Australian National Data Service: Data journals are publications whose primary purpose is to expose datasets by providing the infrastructure and scholarly reward opportunities that will encourage researchers, funders and data centre managers to share research data outputs. Data journals have evolved from traditional journal model that describe datasets including supplementary material. Data journals have more in common than journals that publish articles or overlay papers that describe data but take the concept a few steps further.

As the primary purpose of data journals is to expose and share research data, this form of publishing may be of interest to researchers and data producers for whom data is a primary research output. It enables the author (or data producer) to focus on describing the data itself, rather than producing an extensive analysis of the data. Publishing a data paper may be regarded as best practice in data management as it includes an element of peer review of the dataset, it maximises opportunities for reuse of the dataset and it provides academic accreditation for data scientists as well as front-line researchers.

Data journals are nowadays well established and indexed, which is important for questions of credit, but until now data papers were mostly published in *mixed journals* - journals that have a separate section for data papers, in order to have journal articles and data papers altogether. The conclusion of the article “Data journals: a survey” in 2015 is that although there are platforms to publish data papers, they are still not open enough to foster data sharing and data reuse which is actually the point.

- “Scholarly publication of a searchable metadata document describing a particular online accessible data set, or a group of data sets, published in accordance to the standard academic practices.” Chavan & Peney, 2011 quoted in “Data journals: a survey”, 2015: http://onlinelibrary.wiley.com/doi/10.1002/asi.23358/abstract
- “This artefact is homologous with articles for traditional journals; it is expected to have an identifier and a content with title, authors, abstract, number of sections, and

references." "Data journals: a survey", 2015: http://onlinelibrary.wiley.com/doi/10.1002/asi.23358/abstract

My main thesis is that we tend to separate certification and evaluation from research itself, for different reasons like career pressure, the amount of scholarly publications, the development of questions specific to digital publication format and this leads to a deep lack of satisfaction from those who produce and disseminate scholarly knowledge. Since we won't be able to redesign the academic system in a quick and efficient way, we need to think of ways to improve the conditions which determine how we work and communicate the results of our work. This is the spirit in which this data journal model is being developed by DARIAH. This model is explicitly not purely research but at the interface of research and infrastructure - infrastructure is becoming more and more essential to the way we do research. And it is of crucial importance that researchers identify themselves with this kind of work at the interface of research and infrastructure.

Authorship

Do we still need peer-review? Data journals as a way of reconsidering our evaluation culture and our understanding of research

In this presentation, the idea is to give you a broader historical perspective on the question of authorship and try to identify systematically which aspects of peer review are misleading scholars and which aspects can be reappropriated in a more constructive way.
The core assumption of this presentation is that data can be a scholarly publication when they meet clear academic standards. This is one issue we encounter when dealing with inadequacy of our evaluation system is that it is author centered because it doesn't correspond to actual practices since scholarly work is hardly ever an individual endeavour. The concept of author, as it emerged in the 18th century, is mostly conceived to concentrate on one name, preferably a male name, all the authorship qualities. There are economical considerations behind this idea: big names are attractive and sell more than the mention of the actual contributors (copyist, lab experimenter, editor, publisher, etc.) will do. Also, copyright was conceived with this notion of single authorship which in turn encouraged single big name authorship practices. The opportunity to construct the publication system around a *dispatched authorship model* could have emerged with for example the European Republic of Letters.
If you look at the facts, there is probably no point in our publication history when the author who appears on the book cover was the sole producer of the content of their books. You can probably name isolated counter examples, but the general trend is that book production, especially literature and science production, is and has always been a collaborative phenomenon. We can even identify -with variable accuracy- the different spheres of influence (family, friends, lab assistants, publishers, etc.). We are aware that we have to decipher these modes of participation, but the knowledge of split text, book or scientific production remains some kind of hidden truth even if we know it. This awareness is not a major epistemological principle reflected at large in the humanities' understanding of authorship. The result is that literature history, and to a great extent also science and scholarship history, still live in the myth of the author, this great man.

Why is this a problem for *digital publications*?
Because part of the recognition we need has to do with split authorship or split producership. When it comes to digital publications, we are expecting something different, especially because the modes of cooperation don't obey the same hierarchy and rules than it is or was in the analog world. In digital publications, we don't want the publishers to appear separately anymore because we consider that design and funding is in the domain of scholarship, it doesn't have to be separated from the production of the work. Along with the designer, all intermediates (software designer, technician, etc.) also contribute to the final form of the publication that is offered to the reader. In digital publication, attribution and versioning are two key techniques which have always belonged to the core principles of IT archiving and publishing. The TEI has inscribed in the header the revision and version as a mandatory element for a good reason, and other elements such as institution and funding have a prominent place as well. It is the whole production context that is taken into account. The aim of such an inclusive understanding of text producing is not to make all of the instances involved accountable for the content in a legal sense, but to render the production context as extensively as possible. In other words, there are no technical challenges to the implementation of authorship distribution or split producership in the case of digital publications. There are, though, cultural issues: the change of mentalities that makes the bridge from traditional journal formats to data journals difficult to cross. Some data journals are consistently using micro attribution to address this issue. They name every participant to the production of a dataset by providing appropriate credits to each, by capturing their contribution. But this is not systematically implemented.

Two position papers: Reconsidering scholarly publications in the digital age

In two recently published texts by two working groups I am affiliated to, we listed the various possible authorship or contribution forms with the aim of showing the extension of this variety of functions in text production. We also insisted on the fact that digital publication can take a variety of forms (monographs, articles, edition, database, code, images, videos, etc.). It is not new, but it tends to show how narrow our understanding of a publication in the humanities has become in the course of the history.

The question of academic recognition is at the core of the debate in both papers:

- http://dhd-wp.hab.de/?q=content/empfehlungen_ag_digitales_publizieren
- https://www.merkur-zeitschrift.de/2016/10/24/siggenthesen

Additionally to the question of displaying various and complex authorship and contribution modes, there are two other aspects that make the implementation of data standards and any inherent certification even more difficult:

- Time machine problem: standards and evaluation criteria develop and change. This makes it difficult to attribute them for once and for all and to name them down in a manner that would be definitive.
- One of the most difficult thing to grasp for the traditional academic evaluation system is the fact that digital publication is hardly ever finished. Almost all of them are processual kinds of publications. Some hypotheses are only verified later and implemented in an update, new material is found, etc. There can be many reasons

why you would change a digital publication: emendations, enrichments, cross-checks, etc. The reactions that this processuality phenomenon provokes are not unanimous:

  - Some see it as a chance to publish editorial material progressively, arguing that there is no need to wait 10 years to publish results; instead of that, you can enrich and improve progressively.
  - Others find it hard to cope with the lack of liability inherent with this openness to change: what is the version of reference if you know that the publication is always going to change? How do you refer to that publication? Admittedly, tracking changes via log files and version history is not self-explaining: it has no equivalent in the print culture. So there is a question of scholarly culture and mentality that needs to be addressed specifically and that can't be changed at once.

=> The question of authorship is not just historical or secondary, it is really at the core of the whole academic system.

Peer Review conundrum

Pre-publication peer review was established at a point where it was not possible anymore to print everything. The analog production of all scholarly papers and books would have been too cost intensive. Nowadays, pre-publication peer review is considered on the one hand as the *best* way to evaluate good science, on the other hand as a system that has become *unreliable*. Peer review is taking more and more time as the number of scholars grows and as the concurrence increases in submissions. We have also heard that peer reviews are not really achieving their goal of generally contributing to opening up innovative research questions and answers. This question is regularly addressed in the Guardian Higher Education:

- https://www.theguardian.com/science/2016/sep/21/cut-throat-academia-leads-to-natural-selection-of-bad-science-claims-study
- https://www.theguardian.com/science/2011/sep/05/publish-perish-peer-review-science

Those of you who have received reviews after submitting a paper will know that quite a lot of the overall produced peer reviews consist in a reviewer being touchy because his or her work on the topic was not quoted. We have intrinsic problems with pre-publication peer review, especially because of its dominant position in the evaluation system and because it produces delays in the whole publication process without necessarily improving the quality of submitted papers. As editor for a journal, you have to wait a lot of time for the reviewers to accept to do the review, then you will have to wait for them to actually do the review and this is really delaying the publication of many journals; but on the other hand you know that reviewers have many other review requests pending.

As opposed to the *paper reality of the analog world*, there is no real room problem in the digital world. It doesn't matter if a paper has a predetermined amount of pages, because there is no need to calculate paper and binding cost. The argument is obsolete. Even if the digital production and maintenance of online publications is not at zero cost, institutional

repositories now exist for scholars and allow to make primary data and research accessible, readable, without any valid cost argument.
One model that counters this method is *post-publication review*. One advantage is that it is particularly relevant in the context of data journals as data or publications are being submitted and accepted for submission only if they already fulfill some basic editorial conditions of legibility and scholarship.
=> It means that you submit papers in a better quality if you know that they are consultable online before you submit them.
In this context, we still don't know what post-publication open peer review will bring in the long run, but it seems worth a try compared to the failure of pre-publication peer review we are now experiencing.
=> There is a clear gap between the reality of research, especially in the digital era, in terms of *temporality*, *contribution types*, *techniques available* to take all of these into account on the one hand, and the reality of the evaluation system on the other hand, which is slow, author-focused and in an authoritative position towards the research production.

Why data journals in the Arts and Humanities

This is precisely what we are trying to do with a workflow for data journals in the humanities, which is aiming at improving the recognition of the in-depth phenomena previously mentioned, especially in the case of digital scholarly editions. The initiative of the data journal as a structure comes from DARIAH-EU, it is supported by the French institution The Center for Direct Scientific Communication (CCSD) which hosts the episciences platform, and Inria. It is this infrastructure we are currently adapting in order to offer to the scholarly communities a data journal model in adequation with the reality of scholarship.
This project started under the codename "*living sources*" (one example), because it is based on the core idea that digital resources are processual - they keep growing and need to be re-reviewed along time. The concept was first developed at the Max Planck Institute (MPIWG) and has been since then claimed by commercial platforms such as scienceopen.
What matters is not only to emphasize the lively character of the process, but also the adequacy it wishes to generate, in the overall process of scholarship, between publication and evaluation. In this perspective, the role of the review is not to sort out the good from the bad for it to be published, nor is it to put a stamp on a digital publication. More importantly, the review is becoming an incentive to further development. The review is conceived as a dialogue with the digital resource, both of them working towards improvements.

Submission and review process

The envisioned process goes as follows: A scholar or a group of scholars submits a data paper and an OAI-PMH access to the corresponding metadata. This allows to gather the version of the data which will be reviewed. At that point, it is up to the editorial committee to decide whether technical and content review should be separated, whether this should be double-blind, single-blind, not blind at all or open and in which time frame they want to operate.
The publication can integrate a link to the review, which can be done in the form of a certification, but since there are scholarly contexts in which certifications can be a risky modus operandi, a simple link to the review seems at this point the most viable system.

The review can raise points that could be improved, and the resource's team could be offered to re-submit data when these points have been taken into account. It would then be possible to show clearly the progress achieved along time. Such an organ needs two driving forces:

- a motivated editorial board willing to define a review model and to gather a critical mass of reviewers
- a solid web interface

What DARIAH wants to offer is the technical background, so that the workflow is backed by a solid structure and team. We hope that scholarly communities will find this offer appealing enough to take advantage of the structure we are currently developing. The data journal sandbox is now opened, metadata have been imported from Ortolang and Nakala, the Deutsches Textarchiv and others trusted repositories should follow soon.

Data Journals on the episciences platform

- Episciences platform: episciences.org
- Our sandbox: datajournal.episciences.org
- Data journals: episciences.org/page/journals
- Other examples: https://www.cms.hu-berlin.de/de/dl/dataman/teilen/dokumentation/datajournal

The episciences platform has not been developed for data journals, it is an overlay journal platform, on top of a preprint archive or repository. An overlay journal is an open access electronic journal based on and composed of research articles that are submitted after being deposited in an open archive. The implementation has clearly been made easier by the French centralized repository structure HAL for the Arts and Humanities. An overlay structure requires submissions to be written and formatted properly before being submitted. It spares time in copyediting and formatting for the editorial team, but it requires that the authors take responsibility for their texts much more strongly than it is the case in traditional arts & humanities journals. Usually papers are submitted with linguistic problems, typos, but it doesn't matter because an editorial assistant will do it for you, but when you submit to a repository, it is your way of working that is becoming visible to the scientific community.

Why use the episciences platform for data journals?

In the context of the Journal of the Text Encoding Initiative, I have been working with Open Journal System (OJS), one of the major open access editorial workflow system. In comparison the editorial interface of episciences is incredibly flexible. It can be adapted for practically every editorial need, with a lot of functionalities. For example, as an editor, you have to send reminders to authors and reviewers. In episciences, you can completely automatise the whole process. In OJS, you have to do that by hand, OJS sends only one reminder and you can't change it.
The episciences platform has the advantage and the inconvenient that it relies on the quality of data repositories and requires a clear vision of the amount and type of relationships with the repositories that are envisioned. One of the very great advantages is

that it allows to certify or evaluate any kind of data: not only a research paper, it can also be video, software, a data set, etc.
The episciences platform is designed to harvest metadata via OAI-PMH, which is useful and necessary to gather the information needed for a data journal. Each scholarly community has to identify the resources or repositories relevant to their field, but some technical elements such as the OAI-PMH interface are necessary to exchange information on a reliable basis. This also means that the repository you will be working with has to have clear versioning strategies to allow to re-review data.
On episciences, nothing is kept on the platform itself, everything is harvestable and can be "called" via the metadata and the OAI-PMH interface, at any time as long as the repository offers such an interface. It is a great advantage compared to having data as "supplementary files" or to gather the data for the review as is currently most often practiced. It also allows to avoid proprietary archiving strategies of repositories. Episciences is built on top of open access repositories.
On the other hand, the layout question is left unsolved in the hands of the authors. It is only a minor issue when the scholarly communities are used to work with LaTeX, but arts and humanities scholars are used to editors taking care of the layout. And this is important because what makes a journal is also to have something nice to read in the end and not just an ugly word document in times new roman.

The dashboard offers different options depending on the role you have, but the general review process is:

- Submission
- Attribution of reviewers
- Reviewing process
- Final acceptation
- (Publication)

Hands-on session

Let's build a data journal in digital humanities within our episciences sandbox http://datajournal.episciences.org/

- Group 1: create a rating grid
- Group 2: define the form of peer review
- Group 3: write a rationale for the journal
- Group 4: find potential resources

Elements of guideline
=> Group 1

- Create your own rating grids by defining your evaluation criteria
- Examples (DH Commons):
    - http://dhcommons.org/journal/2016/women%E2%80%99s-print-history-project-1750-1836
    - http://dhcommons.org/journal/issue-1/collaborative-text-annotation-meets-machine-learning-heurecl%C3%A9-digital-heuristic
    - http://dhcommons.org/journal/review-guidelines

- Think about the level of “visibility” of each criterion. Why don’t we have access to one specific criterion? Or why do we have access to another? For instance: level of visibility of the review report? If a review is closed, what might be the consequences on the reviewer’s work?
- Quality of manuscript: writing, clarity, organization, adherence to template (of the journal)
- Criteria for assessing the effectiveness of the data paper content as a mean for accessing the data set(s)
- Data quality, criteria for assessing the methodologies leading to the production of the data set(s)
- Data reusability, criteria for assessing the actual reusability of the data set(s)
- Utility and contribution of data, criteria for assessing the potential of the data set(s) for the community

=> Group 2: Define the action scope of the different roles + type of peer review

=> Group 3: Write the rationale for the data journal (what is the journal about? what do we want to evaluate, with what aim?)

=> Group 4: Look for potential resources and defining which metadata fields are of use for the evaluation to work. Analyse the “quality” of the repositories in terms of metadata. Do the datasets comply the data paper criterias

- Make a list of data repositories. Ensure that data set(s) are usable for a data journal dedicated to Digital Humanities.
- Have a look at the conditions of use
- Do you easily find the information that you need? What do you have to do to obtain them? (email, form…)
- Metadata fields
    - Title
    - Authors
    - Abstract
    - Key words
    - References
- Potential resources - Repositories
    - Nakala
    - Ortolang
    - DTA
    - Registries of repositories: re3data and Open access directory (OAD)
- OAI-PMH
- Availability to provide data set access attributes => DOI or URL
- Competing interests: fundings declaration of any factor that might influence the data set (personal, financial)
- Coverage to provide data set “extent” attributes, including spatial and temporal coverage
- Format (format, encoding, language)
- Licence

- Microattribution: all the creators who contribute to the datasets
- Project: goal and funding
- Provenance: methodology and tools leading to the production of the dataset
- Quality (including data set limitations and anomalies)
- Reuse: information on the potential uses of the data set(s)

Conclusion

What is the benefit of opening a data journal for a scholarly community?
First of all, an editorial board wanting to engage in such an endeavour would benefit from the technical infrastructure and the ongoing reflections on workflow and assessment procedures. Then, a data journal by definition recognizes the value of data, something often still difficult to cope with in arts and humanities scholarly communities. This will contribute to the change in mentality this initiative wants to induce or at least contribute to. Beyond the certification, which might be considered as a first level of readability (for example for our colleagues and students that are too often unaware of quality criteria for digital resources), the second level is the reconciliation of the research process and the evaluation process. One part of the research process that will gain great recognition from this, namely data modelling. This is certainly one big mentality change but re-evaluating data modelling within the frame of data journals is something that could, in the end, also help people to understand better what DH are doing.

Contact

Anne Baillot was a trainee civil servant at the École Normale Supérieure in Paris between 1995 and 1999. She completed her PhD in 2002 in Paris. Since then, she has been living in Berlin where she worked as a post-doctoral researcher at various institutions. Between June 2010 and January 2016, she was junior research group leader at the Institute of German Literature of Humboldt University, funded by the DFG (German Research Foundation). As a junior group leader, Anne Baillot is the editor of Letters and Texts: Intellectual Berlin around 1800. Since 2013, Anne has been a member of the editorial board of fr.hypotheses and en.hypotheses, and since 2015, Anne has been a board member of the German DH association (DHd) and of the European Society for Textual Scholarship. She blogs about her research in English on http://digitalintellectuals.hypotheses.org/ and tweets as @AnneBaillot. Since February 2016, she has joined Laurent Romary's team and is working at the interface between research, infrastructure and cultural heritage institutions. She is Managing Editor for the Journal of the Text Encoding Initiative and is working towards developing new models for journals in the scholarly ecosystem. Her next book (to appear 2017) is dedicated to the relationships between writers and publishers between the late 18th and early 20th century in Germany.

Marie Puren is junior researcher in Digital Humanities at the French Institute for Research in Computer Science and Automation (INRIA) in Paris, members of the Alpage laboratory (INRIA – Paris Diderot University). As collaborators to the PARTHENOS H2020 project, she focuses her research on the development of standards for data management and research tools in Arts and Humanities, and she currently works on the creation of a Data Management Plan for this project. Marie Puren also contributes to the IPERION H2020 project, especially by upgrading its Data Management Plan. After being a lecturer and a responsible for continuing education projects at the Ecole nationale des chartes, Marie Puren has been a visiting lecturer in Digital Humanities at the Paris Sciences et Lettres (PSL) Research University. Her main publications belong to fields including intellectual history of the XXth century, French studies and digital humanities. Marie Puren has been awarded a Ph.D. in History at the Ecole nationale des chartes – Sorbonne University. She holds Master’s degrees in History and Political Science from the Institut d’Etudes Politiques de Paris, and in Digital Humanities from the Ecole nationale des chartes.

Anne Baillot & Marie Puren

Twitter: @AnneBaillot & @puren1406

Email: anne.baillot@gmail.com & marie.puren@inria.fr

8-Economy of Open Access & Open Data Publication

Open access business models for humanities

Pierre Mounier, OpenEdition & EHESS, France

The goal of this presentation is to give some landscape about the development of open access in the SSH, particularly under the perspective of business models. It is maybe puzzling to speak about "*business model for open access*", there is like contradiction between the two terms and my objective is to demonstrate that there is no. In fact, there are not only one but many developing business models for the SSH publication in open access.

STEM disciplines: the rise of APC model

STEM are Science, Technique, Engineering and Medicine. It is the leading force driving the whole ecosystem of publication and scholarly communication towards open access. The firsts initiatives were developed in STEM disciplines and where the most important business model emerged. I will then show that those models doesn't fit well with the Humanities. With open access there is a distinction between the two roads to open access:

- *the green open access model*: development of open archive and self archiving from authors of publication into institutional or disciplinary repositories
- *the gold open access model*: it used to designate the fact of publication in open access; an open access journal or book series by itself whatever the business model behind. But things have changed over time so that now for many people it only became the APC model. APC stands for Article processing charge. It means that journal publishing open access charges the author to publish an article open access. This is the dominant business model for open access in STEM disciplines.

Finch Report

In Europe, UK and the Netherlands have developed this model but the real start at political level was given in UK with the Finch report: it was commission chaired by Janet Finch, a sociologist specialised in sociology of science, where specialists were asked to evaluate the cost of a major shift for UK research publication from subscription based journals to open access APC based journals. The report has been really important because it was the first time that there was a global evaluation of the financial cost to his shift. It evaluated an important additional cost during the transition period from subscription based scholarly publication system to an open access publication system and gave an important impetus for this shift. Now the entire research in UK is shifting towards this new model. Research funders are really helping in this process. This is a “shift from a reader-pays to an author-pays system, which in turn requires a shift in publications processes and business models”.

Jisc Report

Jisc has been monitoring the evolution of the APC cost paid by 14 UK institutions to allow their researchers to publish open access with Article Processing Charges in different journals. During the three last years, there was a huge growth of the number of paid APCs and a huge growth of expenditures paid through APC, but it doesn't mean yet that the cost is growing up because this is just the number of APCs paid. This indicates that the shift is on the way. But this report also demonstrated that the average cost of each APC is also evolving and appear to be growing. It is the case with the so-called full open access journal, it means that all the articles are open access, the business model is completely based on APC, but it is also true with so-called *hybrid journals*. An hybrid journal is a traditional journal subscription based that allow researchers to publish individual articles open access provided that they pay the APC. If as an author you are mandated by your funder to publish your article in a particular traditional journal, you can do it, but then you have to pay the APC. The hybrid model is in fact really common and it is the most criticised as well because this it is based on two sources of revenues: from the authors and from the libraries subscribing to the journal, it is sometimes called "*double dipping model*".

PLOS

A good example of massive model based on APC, not form a commercial publisher but from a non-profit pro open access initiative from the academy community is the Public Library Of Science (PLOS). It is a *massive business model* because it is based on massive publishing, massive collecting of APC, so the average APC of each article to be published in PLOS is lower than with commercial publishers but the business model can run because it is based on massive publishing. You can find all information on their website: last year they published 31 000 articles, this is really huge. It generated around 42 million dollars revenues. And it works, they are balanced - they are non-profit so they have to break even their budget not to make profit.

This trend is powerful, at least in the STEM disciplines and the APC model is supporting this trend.

Max Planck digital library

A very interesting initiative in Europe comes from the Max Planck digital library. They published a white paper to propose to the scientific community to gather funding to allow a major shift towards APC model and open access model. They proposed the generalisation of what the physicist are doing for several years now with the project SCOAP3. The idea is simple, the consortium gathers money from subscriptions that are paid by the libraries of this institution in order to shift the usage of this cost to APC cost. This gives a major incentive to publishers to shift all their journals towards open access APC based. During the last open access conference in Berlin, they presented this white paper called "Disrupting the subscription journal's business model for the necessary large-scale transformation to open access". They took into account all the articles published during one year referenced and indexed into the Web of Science. Then they filtered to how many articles are open access. They multiplied that with the average cost of an APC to have the cost of shifting all the articles to open access and they compared it with the current subscriptions paid by the institutions. They showed that in three countries, Germany, United Kingdom and France, if you compare the current costs covered by subscriptions to the potential cost of shifting everything towards APC, then the comparison stands. They explain that it would not cost more and in some case it can cost less, or at least the same.

APC pre-exists to open access and open access exists without APC

We must remember that APC pre-exist to open access. Open access didn't invented the APC model, even in subscription based journal, in many cases the author also had to pay Article Processing Charges. And the APC model previously w**as a Page Processing Charge**, the charge was paid page by page with additional costs for figures, tables and additional data. So when a researcher wanted to make available more data in his article, he had to pay more because it entailed more cost. Of course, open access exists without APC model. There is no need to make the equivalence between gold open access publishing and APC model. If you look at the Directory of Open Access Journal (DOAJ), the article Open access Article Processing charges: DOAJ survey May 2014 has been published two years ago and it showed that around 60% of the journals registered in the DOAJ are non APC journal, they do not charge the author. So in fact, the majority of open access journals today are non-APC.
But how do they live? How do they support their operation?

OA business models in the humanities and social sciences

The invisible rise of institution supported journals

Financial support from the institutions should be highlighted. They support with work force and infrastructures most of the open access journal today. Examples:

- Revues.org: almost none of the journals are APC based because they are in SSH and it is not a meaningful business model for SSH. Most of the journals are supported by their institution.
- SCIELO: It is much bigger than revues.org, it covers all Latin America, particularly Brazil. You have more than 1200 journals published open access and most of them are also supported by their institution to operate.
- Redalyc: It is based in Mexico and covers journals from different Latin American countries.

One major supporting tool for the development of those institution supported open access journal is OJS from the Public Knowledge Project (PKP consortium). It is an initiative from the Simon Fraser University in Canada who developed a lot of tools to allow small publishers or institutions to set up websites or platforms to publish open access their own journal; one of these tool is Open Journal System, but they also recently developed Open Monograph Press (OMP), the equivalent of OJS for books. In 2015, there was more than 10 000 installation of OJS everywhere in the world. It means several thousands of journals run by this kind of tool which are supported by their institution. The library of a university can set up a publication platform for journals of the university to be published open access.

Another report in French

Publier: à quel prix ? Etude sur la structuration des coûts de publication pour les françaises en SHS. It was a study made by a French consortium to evaluate the cost of publication of an article. They used an online survey and made interviews with university presses and different journals in France to evaluate the cost of edition and publication of an article, open access or not. The result showed that the average cost of article publication was around 1 300€, taking into account proofreading, typesetting, peer-reviewing, etc.

Weakness of the institutional support model

This model is interesting because it ensures revenues for the journal to operate, backed up by its own institution. But it has some weakness: when the institution decides not to subsidise anymore because the policy has changed for example.

- Terrain is an important anthropological journal in France which was subsidised so far by the Ministry of Culture. One day, for many reasons, the Ministry of Culture decided to stop to subsidise this journal. So it was the end of the journal. Their last issue was on the topic of *Nostalgia* to say goodbye. But, a few month ago another institution, the CNRS, decided to step in, to take back the journal and to allocate one full-time editor to run the journal. The journal will soon reborn and it will be the topic of the next issue! Even if it is good news, it was disrupted during several months.
- HAU journal (also from anthropology) => If there is a weakness because a journal is supported by only one institution, then it can interesting to have support from different institutions, like this it lowers the risk that institution disengage from the journal. It is a journal of ethnographic theory founded a few years ago by scholars from different countries. Very rapidly, they made the development of their journal supported not on one institution but many. Now they have set up a network of ethnographic theory: HAU-N.E.T. It is supported by a huge number of different institutions from various countries. So each year, some institutions can step out but it is not a problem anymore because there are other institutions stepping in. In fact, it regulates the budget of the journal and ensures some sustainability.
- Open Library of Humanities: The business model of this platform of journals is not based on institutions directly but on the libraries of the institutions. You can see on their website how it works: libraries pay a small amount each year and if they have a sufficient number of libraries, it can cover the cost of operation of the platform and of the journals hosted; if more libraries participate, it lowers the cost by article published for each library. And this support gives the right to participate to the board, no other privilege. So the **libraries can participate to the governance of the project**.

Freemium model for journals. The example of OpenEdition

This model is used by several players in the field amongst which you can find OpenEdition. When we, at OpenEdition, decided to use the freemium model to propose to our journals to use this model, it was not primarily for financial or economic reasons. We have chosen it because we made a survey around 2009 in the libraries about the visibility of our open access content in the libraries and we found out that this visibility was less important in those libraries than the visibility of the paid content in the same libraries. We have commercial competitors which are not open access and we saw that the journals disseminated or publishing by this competing platform were more visible in the library

catalogs and more used by library. It was counterintuitive because when you publish open access you assume that those resources would be more visible than the "gated content".
=> *When you are in a library, you have limited human resources, so you have to make a choice and of course allocate human resources to maximise usage on what you pay for. It is just good management.*
So we had this paradox that every librarian we talked to was supporting open access, but on their ordinary daily business they were maximising their time on what they pay for. So we thought that if we want our open access resources to be visible and used inside the libraries, then we have to make them pay for something. And that was the beginning of the logic of using this freemium model where the resources are still open access because it continues to increase visibility but not inside the libraries specifically. It increases discoverability of the resources on the open web where most of the users are but we developed specific services to be paid by librarians to support this open access resources and to justify the fact that they give more visibility to these resources inside the library.

Freemium: a business model to address 3 challenges

- To increase sustainability for Open Access publications
- To set up a business model adapted to specific needs of humanities and social sciences scholarly communication
- To increase impact of Open Access publications in libraries

Freemium is a business model by which a product or service is provided free of charge, but money (premium) is charged for *advanced features, functionality, or virtual goods*. The word "freemium" is a portmanteau neologism combining the two aspects of the business model: "free" and "premium". But in fact, freemium is a common business model in the digital world, in press but not only: Lemonde, Wired, The New York Times, OpenBook Publishers or OECD. Some articles are free and some others are not, or additional services can be paid for. If you have a smartphone, most of your apps are freemium, you can download them and use them freely but you can buy additional services inside the app. But there are many ways of implementing freemium. For OpenEdition, the freemium model is from libraries for publishers:

- Open Access on the Web (html) – free (you can read online, copy/paste, print, save, etc.)
- Pdf and epub download and other services licensed to libraries – premium (no DRM, no download quotas)
- 66% of income for books and journals publishers
- Partnership with libraries consortia: Couperin, Crepuq, CIFNAL

There is an incentive in the freemium model to add constantly new premium services, that is why we developed added values services for the libraries. It is not only the ability for their public to download pdf and epub, it is also the ability for them to have Counter statistics on their campus, it is a standard to evaluate the usage electronic resources in libraries. They also have hotline, long-term preservation of the content, all metadata that can directly fill their catalog, etc. It gives the library incentives to maximise usage of this open access resources but it also gives us incentives to answer the specific needs of the libraries. For example, libraries used to say us for years that it would be nice to have a MARK record of

metadata to be able to be ingested in our catalog. So the incentive was to make this development for the libraries.
=> We are in constant dialogue with the libraries that ask us new services and we have to answer to this.

The results of the freemium program at Openedition are pretty good, we have around 150 journals from our platform, participating to this program and getting revenues. We have 50 books publishers and more than 110 libraries subscribing to the program. It took a lot of time to convince the American libraries because we are far from them but eventually we did it. So year after year the revenues are growing. The results also are good on a usage level - the first aim was to maximise usage in the library. We can calculate retrospectively the usage before we implemented the premium program, make a comparison and deduce some impact of the program. For example, Year after year, when Cybergeo entered the program, there was a growth of the usage inside the library. Is is only a hint, not a global survey.

Freemium can be complementary to another stream of revenue which can be the support of one or several institutions. Freemium is not here to replace other business models, but it can be an additional model.

How to cite an article in freemium? => Html is available. We developed a feature numbering each paragraph in the html, so your citation practice will be better than citing a page because it is more precise and meaningful as you cite a meaning part of the content and not a formal part which is a page, which depends on the layout.

The case of books

Some say that open access for books is infancy, somehow it is right, but it is just the beginning. There are many initiatives trying to develop open access book publishing in SSH.
Some fundamentals:
Oapen reports, a book publishing platform from the Netherlands with a European scale, that produces a lot of reports and surveys, studies to assess the business models of open access book publishing, the usages, to identify challenges to overcome, etc. For example in the report "A project exploring Open Access monograph publishing in the Netherlands. Final Report", they calculated the average cost of publishing a book in the Netherlands, as the French did, so the total average cost for one book publishing, open access or not, is around 12 000€. Other studies confirmed this evaluation. They also divided the different parts of the cost and separated:

- the first copy preparation cost, without printing, distribution and commercialisation costs, the "open access cost". It is the cost publishing a book, in open access or not, the cost of preparing the first digital copy of a book.
- Then you have the cost of selling, printing, distributing the book.

They found that the open access cost was around 50% of the total cost. So the cost of publishing an open access book on a platform, a pdf on a platform is around 6 000€, not so

much in fact. Their message to funder is that if you want to fund open access book, then it is not so much to add to a research project budget.
Important output of this study:

- Visibility and discoverability of open access books are higher than traditional books. In the methodology of the study, they took 50 books open access and 50 other books, non open access. Scientifically, they tried to take the same type of books, authors have the same prestige, publisher the same, same subject, etc. In term of usage, OA books are higher because they are more discoverable.
- They also showed that there was almost no impact on the sales of the print copy of OA books. It means that when you publish a book OA on the Internet, if you have a print copy of the book that you distribute by other means, you can also sell your printed copy of the book at the same level as if the book was not open access.
=> There is no negative impact of distributing OA book on the Internet on the sales.

London Economics Report "Economic analysis of business models for OA monographs January 2015". This report helps us a lot to categorise the business model we can find in OA book publishing:

- *Traditional publisher*: Oxford University Press, for now, OA is very marginal is this business model
- *New university press* (NUP): UCL Press, the library is leading this completely digital and OA initiative
- *Mission-oriented OA*: Language Science Press - Unsatisfied scholars about publication set up their own publishing organisation
- *Freemium OA*: OpenBook Publishers
- *Aggregator/distributor*: OAPEN (works with Knowledge Unlatched) or KU Books - they are the middleman between the publisher and the libraries.
- *Author payment model* (Book Processing Charges, BPC): Ubiquity Press, they separate the different functionalities that an author want to subscribe to order to publish his book; the cost can be different if for example the author does by himself the typesetting or proofreading, he will not have to pay for it; he can also have additional services to maximise the visibility for example. It is not a package, you can choose the services you want.

Freemium model for books: the example of Open Book Publishers

Books are available in different format and you can buy the paperback, the print version of the book. You can read everything OA in html. The book is OA on the platform but if you want to download additional formats, then you have to pay for that. First, they disseminate their books on several platforms, for example on OpenEdition or GoogleBooks, so it increases the visibility and the usage as it reaches out different part of the audience through different channels. Regarding the revenues, they have a business model mainly based on the sales revenue but they also have other streams of revenue: they have grants, for example, the authors receiving grants can give it to the publisher to support the business model of the publisher; they also have a specific library membership program so their users can download freely the pdf or epub files or have a rebate on the print version.

Freemium for data: the example of OECD publishing

Business model for open data publication: I have found only one very good example of business model on data with OECD freemium with the editorialisation of data. The OECD has several research projects and they publish on a daily basis journals, books and statistical series with a freemium model. It means that most of their content, data and publication, is available OA on their platform. They also propose to institutions to subscribe to additional services based on the editorialisation of their content, publication and data. So as a user, unregistered, you can download data but if you want, for example, to access to data in different format or if you want to aggregate some data and download this aggregation you made yourself on the platform, or if you want specific representation of data to copy/paste in your own publication, or if you want to do some advanced search in data, then you have to be affiliated to an institution that subscribed to OECD iLibrary. I think that what is interesting here is that you have free access to raw data, but there is lot of work of editorialisation to support these premium services that to be paid. They are also constantly improving their business model: some previously premium services become free because they invent new services that are premium. In fact, one major point made by the London Economics report is that **with the freemium model, there is a strong incentive for innovation** because you cannot maintain forever the same service as premium. Usage is evolving, demand from users is evolving too.

Conclusion

“It is a numbers game, so bust out your Excel spreadsheet. It's all about finding things in the margins - lots of little things rather than one key thing”.
From the inventor of Dropbox, D. Houston, in "Case studies in Freemium: Pandora, Dropbox, Evernote, Automattic and MailChimp", Gigaom, March 2010.
It means that you cannot elaborate your budget on one stream of revenue, you have to elaborate it on many different streams of revenue which are complementary. This is a way to make a sustainable business model.
=> Diversification of the publishing business model! It is not anymore a matter of selling books in bookshops and libraries but also:

- Funding (gold)
- Print (on demand) sales
- Premium services income
- In-kind institutional support
- Crowdfunding

Contact

Pierre Mounier is deputy director of OpenEdition, a comprehensive infrastructure based in France for open access publication and communication in the humanities and social sciences. OpenEdition offers several platforms for journals, scientific announcements, academic blogs, and, finally, books, in different languages and from different countries. Pierre teaches digital humanities at the EHESS in Paris. He has published several books about the social and political impact of ICT, digital publishing and digital humanities.

Associate Director for international development OpenEdition

Coordinator of OPERAS: http://operas.hypotheses.org
ORCID: http://orcid.org/0000-0003-0691-6063
Twitter: @piotrr70
Email: pierre.mounier@openedition.org

Sdvig Press

Patrick Flack, Sdvig Press, Swiss

Patrick Flack is managing editor of Sdvig Press, an open access non-profit academic publishing house. He is a researcher who came into this role of publisher because of the demand and the structure of his research project. The publishing house itself is not a traditional one but more a hybrid between a publishing house, a digital infrastructure and a research project. Its mission is to respond to a specific research challenge. The point is not OA per se, but OA is an integral part of the project, a mean to make it function and a manifesto for the research project that gave birth to the publishing house.

Research challenge

My research itself is focused on the history of structuralism, not only French structuralism, but structuralism in Central and Eastern Europe. The idea behind this project is to counter the usual vision we have of structuralism with Saussure, the great Genevan linguist, with schools in between Prague, Copenhagen and Geneva; and Claude Levi-Strauss from whose work the structuralist movement as we usually know it (Barthes, Lacan, etc.) evolved. In fact, the origins of structuralism are far more complex and we want to show that the network of the history of structuralism is not only French focused, but involves a lot of scholars and thinkers from different disciplines in Prague, in Russia, etc.
The research challenge is important as some authors are little known and have written in different languages (German, Russian, Polish, etc.). Moreover, the corpus necessary to represent the history of structuralism and counter narratives that are strongly established since the French movement from the 60's is enormous. Worst of all, most sources have not been curated and edited in a proper way: many books have not been republished and critically edited. This is why it is now necessary to make as many texts as possible accessible otherwise no researchers will be able to take an interest and conduct their own research. The corpus has to be curated and made accessible in a multilingual way. This task has to be done sustainably, internationally and cooperatively. Publishing and curating the corpus is an integral part of the research, otherwise the established narratives will not move.

Publishing house

It has been developed as a solution to this research challenge. The main problem is not really a specific research problem because the corpus is really wide (and allows for competing interpretations), but it is to get the project financed for a long period of time. The publishing house basically started with the idea of print on demand publishing and OA, through a presentation of Pierre Mounier. But moving beyond the OA publication of such digital material, we then wanted to have a virtual environment where we could work on and structure the whole corpus. We took some inspiration from OpenEdition, for example hypotheses.org the blog platform and turned this into a digital project and a curated data and text repository with overlay services. We have a few journals but most importantly, we have platform oriented services: Acta Structuralica, Phenomenological Reviews, Open commons of Phenomenology, Structuralica, PACEM. Organizing the publishing activity around a community was a decisive moment for this project.

Structure

The most mature project, the Open Commons of Phenomenology is discussed here as an example. The most important part is in the repository, with the major authors of phenomenology. You can select authors, have a short bibliographic introduction, then a complete bibliographies. This is not as in a library catalog or Worldcat where they have basically everything that has been inputted; we went to the Husserl archive and we found a complete edited bibliography and inputted everything: http://ophen.org/pers-100275
The granularity is at the chapter and article level. We do not only input a book, we also include all chapters and all articles. Ex: Maurice Merleau-Ponty, La structure du comportement.
We respect copyright, it is always indicated on the page of the publication, so the pdf is made available if the copyright allows it. If not, for example Merleau-Ponty is free of right in Canada, so there is a copy on the Canadian website but we didn't repost it on the website, we linked it externally.
All the data is strictly and carefully curated and structured. We have a function like in Worldcat, if you estimate that something is duplicated or missing; or you can find all the translations, or other editions, etc. It is a very efficient way if you are doing the history of publication of a philosopher: these information are structured and presented in a way that is useful to users.
=> The platform offers not only contents or data itself, it structures it, it presents it in ways that are immediately useful to users.

Finally, it is exhaustive, up to a point of course; you can always find new publications of an author, but if you go to the Merleau-Ponty's page on the open commons, you will have everything from him. If it is the Heidegger page which is under copyright, you will have links to all the text. Every single text is quickly and efficiently accessible in pdf. We carried out OCR (Optical Character Recognition) and a real digital edition. This is the backbone of the platform, in a way, the part we intent to sell.

Another way the data is organized is by journal and author. It can be very hard to carry it out in library catalog, especially if it is an old journal that has a banal name, like "People and School" because it will always give a hundred results to find the review. The database is still quite *small*, still we have 25 000 entries and 2000 full texts.

When you are logged in, you can submit references with a two-stage process and declare metadata (title, subtitle, editor, language, DOI, rights, etc.). You need to link to translation or original edition, that's how we can move so efficiently behind. When you save your submission, it goes to the moderator, it will create an input from you and it will be reviewed and corrected if needed. This brings quality to data. It really makes sense to do curate carefully and check the data is correct and structured, linking to the author, to other publications.

The other aspect of the project which was inspired by hypotheses.org is to have blogs for research projects or society pages. For example, North American Society of Early Phenomenology has a blog here where they post call for papers and so on. Its feed is replicated on the main page. Scuola di Milano, a group of researchers who want to input a lot of bibliographical information and present them also has a blog here. It is really at the frontier of being a blog, a project, a page and almost a journal. We haven't given it a ISSN, individual posts don't have a DOI, but we can decide to do that. They have a research group that could become the editorial board, it could change into a journal.

An interesting feature that explains why we didn't use hypotheses.org for this blog is that each philosopher has a page and it is then linked directly to the main database, it gets the bibliographical data and publishes it on the blog page. This gives the possibility to present results or data which they have imputed in a common database (accessible from different platforms). This idea is to have 20 or 30 research projects which we select, then they input information about authors with linguistic capacity. We aim at integrating visualisation tools (timelines and maps).

We also have templates to input biographical data with *events* in the structured database, so all will be connected as well with an author page. This allows to know all the courses an author gave and to link it to who were the students attending as well. This will be combined with the bibliographical data. It obviously gives a lot of possibilities of visualisation that can be integrated in the research project or institutes' webpages.

With a structured database where you have all phenomenology and structuralism and 10 to 20 platforms which would cover the whole field of relevant thinkers, I can do my work - and many other researchers can carry out their own. It is the institutional name that is an incentive for quality; a quality dimension, so in that sense it is not completely open structure: it is an academic scholarly publication or communication form.

How to finance the project

We have two important parts. The thing is not about making money, but how to finance the project and its labour cost.

- Digital library can use as an infrastructure and contribute with qualified working hours, which cost us nothing except answering some emails, setting up the blogs, etc.
- Freemium model: all the content is open access and we offer extra features: full bibliography in a full list, visualisation with timelines, maps, biographical data, etc.

Editorial and Governance board

Our strategy is to become *incontournable* and make scholar recommend the platform to libraries. It is a project for the community, community based, like the commons, scholars just need librarian to support by subscribing. The organisation is non-profit, it means that all the money goes directly back to the project, to develop the technical side, to do translations, reedition and develop further projects. And the big dream is that the commons doesn't stay digital, it is also real world community, so we would love to make conferences and make an institute of open commons, to have real world places. My colleague also owns a second hand bookstore in Lausanne, so we have this completely digital global open access project but the physical place where people would communicate and talk about research can be part of the project as well. It is a revolution, as Stiegler says, with the Internet compared to invention of the printing press, it changes a lot of things, not only about publishing books, it changes the way you cooperate.

Contact

Patrick Flack is the managing director of sdvig press, an open access, non-profit academic publishing house. He is also associate member of the Central-European Institute for Philosophy (Czech Academy of Sciences, Prague). Since completing his PhD in 2011 (Comparative Literature, Charles University in Prague), he has worked in Helsinki, Leuven and Berlin as a post-doctoral researcher funded by the Swiss National Scientific Foundation. His research focuses on structuralism and a trans-cultural, interdisciplinary approach to its historiography.
With sdvig press, he is currently developing a number of open access thematic platforms – such as the Open Commons of Phenomenogy – designed to function as sustainable infrastructural and communication hubs for their respective scientific communities. The development of these platforms is linked directly with international institutions (Husserl Archives, Czech National Library, etc.), embedding their research projects, archival holdings and editorial outputs.

Sdvig Press: http://sdvigpress.org/
Twitter: @panflack
Email: flack@sdvigpress.org

9-Infrastructure & Platform

Contrasting platforms and infrastructures as configurations for data sharing

Jean-Christophe Plantin, LSE, UK

I am the co-director for 2016-2017 of the Master program Data and Society at the LSE, where I work with students on topics such as the governance and public policies in sectors that are increasingly *data driven*. As it is in media and communications, we talk a lot about journalism and social media, but a lot of things we see apply also to the world of library, data sharing and infrastructure. The larger umbrella of my research is the *platformization of social life*, which designates services provided typically by infrastructures that are increasingly provided by digital platforms. For example, I work a lot on maps and cartography, such as the IGN in France. Since the arrival of the World Wide Web, we have seen platform-based mode of cartography that are increasingly reaching the scale, and the nature of essential service of infrastructure. This is the configuration I am working on.
In the world of archiving, we see the same tension, between, on one side, infrastructure and data archive that have been existing for a long time, and more recent web-based platforms on the other side, that present their activities as doing the same services sometimes *in addition* to infrastructure, sometimes to *replace* infrastructures.
The purpose of this talk is to compare these two entities, to "map" their relationship and to see what are the risks and benefits when it comes to data sharing for scholarship.

Data as scholarly output

Incentives:

- Rise of big data and data across disciplines
- New data sharing requirements (incentives from funding bodies)
- Diversification of materials considered as scholarly outputs: greater interest from researchers, librarian, etc. to extend the artefact of scholarly communication beyond journals (datasets, simulation, softwares, etc.)

The decentralisation of scholarly infrastructure

- Rise of the World Wide Web and the possibilities to decentralise scholarly infrastructure
- Web as technology and culture that challenged the traditional vertical and central system of scholarly infrastructures: publisher, library, archive
- Reduction of the publication cost with electronic media, scholarly productivity measure (hyperlinks as alternative to citation count), e-print movement and repository effort of the early 2000's (ArXiv.org)

Figshare

Figshare is very much a product of these two tendencies, using technical environment of the Web, adopting values and characteristics of platform, as well as positioning itself towards these new needs and incentives around data. Figshare describes itself as a "*platform where researchers can store, share and get credit for all of their research*" but the broader objective is "*improving and opening up the dissemination and discovery of scientific research*". It invites individual researchers to self-archive their outputs (datasets, graphics, presentation slides, almost anything) through personal profiles you can create, such as on Academia or Facebook. It was created in 2011 as a "pet project" from Dr Mark Hahnel, a stem cell graduate, before being a company hosted since 2012 by digital science based in London.

This is figshare.com, but Figshare also has a technological side, which is at the basis of their second target: Figshare can be deployed as a middleware service marketed as *Figshare for Institution* (e.g. with Monash University) or *Figshare for publishers* (e.g. with PLOS). It links together institution-based and publisher web portals with a custom-made data deposit and publication platform. If you are a university, a research lab, a publisher, a library, you can contract with Figshare and get custom interface, storage services, search capabilities, etc.

Figshare as a case study is a good example of a platform based-technology. We know that both platform and infrastructure rely on *principles*, they have *technical characteristics* that are different, so what happen when they are both conflated? What does that mean for scholarship? For data accessibility?

Infrastructure and platform properties

What defines an infrastructure and what defines a platform according to a series of criteria:

- Architecture
- Relation between components
- Market structure
- Focal interest
- Standardisation
- Temporality
- Scale
- Funding
- Agency of user

You can find further details in the article "Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook".

Relation between components

- Infrastructure: Interoperability through standards
- Platform: Programmability within affordance, APIs

Market structure

- Infrastructure: Administratively regulated in public interest; sometimes private or public monopoly

- Platform: Private, competitive, sometimes regulated via antitrust and intellectual property

Focal interest

- Infra: Public value; essential services
- Platform: Private profit, user benefits

Standardisation

- Infrastructure: Negotiated or de facto
- Platform: Unilaterally imposed by platform

Temporality

- Infrastructure: long term sustainability, reliability
- Platform: Frequent updating for competitive environment.

=> This is a typology of these two separated entities, but it allows to see what is in between, when these two properties conflate and sometimes conflict.

ICPSR

The Inter-university Consortium for Political and Social Research (ICPSR), university of Michigan, is an infrastructure, a data archive created in 1962. They are specialized in archiving and sharing social science data, especially large-scale survey data produced and published by research institutes. They are a membership-based institution: everybody can deposit datasets on their website, but if you want to access it, you have to be part of its network of more than 800 member-institutions. They obtain data either from researchers who directly deposit datasets on the website, or they proactively acquire some datasets that their community wants to get.

Care of data through manual processing

ICPSR works as a library when it comes to data circulation, data acquisition, data processing and sharing. Researchers deposit their datasets, if it fits the appraisal criteria of the institution, all the datasets are processed the same way to fit the standard of the ICPSR. Data processors are the people who make sure deposited datasets are “clean enough” to fit the criteria of the institution. The reconstitution of the "pipeline" for data processing includes the following steps:

- Deposit of dataset
- Dispatch
- Repair
- Contact with the PI (principal investigator)
- Prepare
- Verify
- Publish

Every submitted dataset go through this pipeline. The two specific actions that I want to stress here are what I call “repair” and “prepare” because it is the action of data processors that really reflects this care of data. *The institution consider by default that datasets are not* *perfect, are "broken" or contain mistakes and problems, so they hire dedicated people who putting their expertise and work time to take care of it before publication*.

- "Repair": data processors rely on different scripts and softwares to go through huge SPSS files, and to flag some mistakes, inconsistent code, some missing documents, etc.
- "Prepare": data processors make sure data is presented in a way that fits the ICPSR standards, to make sure every document is presented in a specific way, using templates, make that metadata fit with the catalog of the institution.

Figshare: Automatic data provision

Of course, it contrasts with Figshare, which has a very different model. On Figshare.com, the website where you have a profile, you can just drop your dataset. There is a minimal curation, there is no added value, no labour. Data circulation is also different:

- No processing, self-deposit
- Centrality of the API (Application Programming Interface)
  - To connect web-based actors with the scholarly world (ex. reference platform like Zotero and Mendeley, repositories like Github, academic or institution libraries, cloud storage, online scholarly journals). All is mediated through their technology, their API.
  - To connect with institutions and publishers (PLOS and Monash University), this is how they design custom search capacity or custom portals.

To understand the way APIs mediate between different actors, you can read "Code as a research object". Figshare, Mozilla Science Lab and Github got together in this project and designed a Firefox browser extension that generates a DOI for datasets, code are deposited on Github and can be released on Figshare. Different platforms connect together because they talk the same language. Here, the API is central, and constitute a brokerage point between Github and Figshare, using their APIs and developing systems so that any Github repository can be processed as a package. We see here how technically a *web based data model circulation* is applied to the data sharing and data reference.

Consequences on scholarship

Relations between these two entities for scholarship:

- Infrastructures: for example, ICPSR developed an expertise over more than 50 years, network of members, reputation, recognized standards and quality, labor intensive, slow but considered as maintenance work, so there is an high turnover, so you need to train again new people, so it costs time and money for an institution.
  - Manual processing for specific type of data, but what about big heterogeneous data?
  - *Path dependence* and *reverse salient* (Hughes 1983, historian of technical systems who studied electrification in the US): when you look at how the ICPSR works and how it developed an expertise, they are very good at what they are doing but there is a path dependency towards one specific type of datasets. They became extremely good at it, they are highly specialized

mostly on survey data but as a result, they have tremendously narrowed down the amount of datasets they can accommodate. If you want to work with other kind of data, you have to redesign the whole pipeline.

- Platforms:
    - Figshare presents itself as having a strong commitment to open data and open science. Mark Hahnel is involved in lots of scientific open data events. He is very active in the community. Figshare released a report on the state of open data. They also have a clever way to contact with libraries, they integrate themselves with the existing standards, digital preservation network, citation survey such as data science. They are doing it in a great way if we have open access as goal. But there is no mean, technically or philosophically, to make sure this commitment is going to stay for the long run. We have a lot of examples of platforms who changed tremendously their data access with regards to a change in their business model, cf #DeleteAcademiaEdu because Academia.edu added a new feature going towards freemium model but we don't know what is going to happen afterwards.
    - *Splintering infrastructures* (Graham, Marvin, 2001). They use this term for urbanism in a huge study on cities, showing that with the increase of what they characterize as neoliberalism, cities that usually had a provision of essential services with infrastructure are being replaced by what they call "*networked premium spaces*". It is the idea that instead of having essential services for everybody, we have pockets of priorities where people are going to have customized access to the services. So the infrastructures are “splintered.”

Conclusion

=> More heterogeneous data
=> More incentives to deposit
=> Two candidates to accommodate these data: infrastructures and platforms

- Infrastructures developed a specific expertise following high standardisation but making it hard to accommodate a wide range of data.
- Platforms jump right in in that situation, using the flexibility and plasticity of its structure to organise data sharing, direct deposit, API. But with different configuration come different forms of care through data processing (absence of care through automatic provision).

Platform can commit to openness but there is no way to make sure there is commitment on the long run. Platforms are part of a larger decentralization of traditional scholarly infrastructure, a risk that can emphasize the tendency of the splintering of infrastructure by showing how institutions can do more with less.

Contact

Jean-Christophe Plantin is Assistant Professor at the London School of Economics and Political Science, department of Media & Communications. He investigates the civic use of mapping platforms, the collaborative challenges in big data science, and the evolution of knowledge infrastructures. His research was funded by the Alfred P. Sloan Foundation, the Gordon and Betty Moore Foundation, the European Regional Development Fund, and the University of Michigan MCubed Program. His work was published in New Media & Society, Media, Culture & Society, and the International Journal of Communication.

Twitter: @JCPlantin
Email: j.plantin1@lse.ac.uk

Huma-Num infrastructure

Nicolas Larrousse, Huma-Num, France

We prepared this presentation with Jean-Christophe Plantin with the idea to show a (relatively) conservative infrastructure compared to those new platforms which are really dynamics and different in their goals. I will focus on our own sort of Figshare in Huma-Num which is Nakala in order to show the differences of approach.

During the last decade, we saw, as Jean-Christophe Plantin showed, that Humanities are really in a *digital turn* and regular researchers deal now with digital data. It means that there is a need for tools, softwares, mediation between raw data and researchers. They also need to appropriate data, they need to shape data to do something useful for their research and this is a long process, from the beginning to the end. Between, there is a huge cycle. And now that we have more and more data, it is getting a problem. It means that researchers can't work only with their personal computers anymore. They need to have some tools able to deal with the amount of data produced. At Huma-Num, we also see that there is a need for more sophisticated tools, the classical Filemaker is no longer sufficient. For instance, they use Geographical Information System and you need to be aware to use it. They also need to preserve data which is, in my opinion, a great problem for the future of research.

But what is an infrastructure today? Is it a data center?
In the definition of the European Community, it is a whole set of things. Of course, you need to have some computers and a data center somewhere but you also need to develop expertise and a network of people.
=> An infrastructure is no longer computers somewhere. Above all, you deal with data. So you need to show it, to share it, to disseminate it and to preserve it.

Huma-Num is a French infrastructure for Humanities. In order to address this kind of question, we support groups of people with expertise (we call them consortia) but we also provide virtual machines, softwares, hosting, preservation, storage, etc. We organized it as a sort of onion with at the center the "users" (researchers, network of researchers, projects). We also create consortia, that means that we fund group of people who share the same interest about scientific objects, not necessarily coming from the same discipline, and so they can work together. We expect from them to build expertise, good practices, tools, standards etc. that we can share with other communities. With this process, we try here to build a *virtuous circle* which is really the center of our project. And beside that, we provide some core services, machines, virtual machines, softwares which are generic services, but we also provide specific services centered on data: that our way to fit to the technical needs of research projects. At Huma-Num we want to offer original services to the community for data processing.
The last part of the onion is that we need to have exchange with other people doing the same thing, especially in Europe. European Infrastructures for Humanities are on the way (e.g. DARIAH) and Huma-Num intends to be a sort of hub to Europe for the French community in Humanities. The main idea is to valorise what the French community is doing

(tools, expertise, data). In return, we can get information about the way communities are organized abroad and what they are doing, so we can situate our actions and reinforce networks of expertise.

Nowadays, Huma-Num funds about 10 consortia in very different disciplinary fields (Musica: music encoding initiative, etc.). The idea is to have people form different structures to work together, which is not always easy in France. For example, we have the 3D consortium, which is not disciplinary, it is a group of people interested on 3D in general for their work, it can be archeologists or people coming from geography. There is another consortium, Archipolis, on Political Science which concerne is long term preservation of surveys. They invent their activity in fact, for example new set of metadata to describe a common object (3D or surveys), because standards need to be adapted in order to address new needs.

The second pillar of Huma-Num is technological services: along big storage, you have softwares and hosting websites, virtual machines. The infrastructure is hosted in an already existing data center, we didn't want to build a new one for Humanities in France, it is a total non-sense for us. So we pay to be in a huge data facility specialised in physics and so we can provide a lot of storage, garanty power and network availability, etc.

The originality of Huma-Num is about data services - it could be an issue in Humanities because researchers generally store data on their computer and it can get lost easily (e.g. computer crash, end of fundings, retirement, etc.). There are a lot of *nice* ways to lose data. For instance, if you have a project funded for three years and you have a beautiful website done by a private contractor, a postdoc or an engineer and then the project stops to be funded. As your data is in the website, it is lost because nobody is going to maintain it. Besides the technology gets obsolete rapidly nowadays. So, after a while, there is no way to access data and to cite it.
So we try to provide a set of tools in order to preserve data and the preservation for us is to be able to share data, to disseminate data, to inform people that data exist. We also provide long-term preservation service, which is more specialized: for this, we rely on an existing specialized data center. Our added value for this data center is to provide them with new type of data and metadata, it helps them to improve their way of dealing with the long term preservation service.

NAKALA & ISIDORE

Since many scientific data producers do not have the digital infrastructure to provide persistent and interoperable access to their data, Huma-Num has implemented a tool to expose and share research data called “NAKALA”.
NAKALA provides mainly three types of services:

- A PID (Persistent IDentifier) to data and metadata
- Permanent data access
- An exposition of metadata through a Triple Store and OAI-PMH

NAKALA is a simple repository for sharing resources:

- The main API to access data is the Triple Store
- You can cite your Data and your MetaData

- Data and MetaData are immediately available

But if you wish to show your data, you need another application, not provided by NAKALA:

- A search Engine
- Tools for visualization

We decided to develop NAKALA in order to address the issue of data being lost in a website. NAKALA is a repository with the idea of separating the place where data is and the place where you show it. As Jean-Christophe Plantin mentioned, there is nowadays a competitive war between technologies. When you choose a technology, after a few years it will no longer be trendy or available. This is why data should be somewhere else to be able to share and show it on the mid-term.

So NAKALA is this repository and then you will have to build something upon it to show it. In Huma-Num we use Omeka to do so. It is not really flashy but it works very well. Then we connect NAKALA and OMEKA. If one day you want to get rid of OMEKA, it is not a problem, your data is still untouched in NAKALA. You can also use hypotheses.org for example to show data from NAKALA.

ISIDORE is the place where you harvest a lot of repositories, including NAKALA but not only. ISIDORE harvest about 4000 sources and show 4 Millions records. Then there is a chain of treatment to enrich, classify and link all these metadata. Isidore only deals with metadata, never with data, it attributes a handle to be able to cite your data and there is a huge work, a sort of *automatic curation of metadata* to enrich, to classify and to put it in Linked Open Data.
=> Every word or term used in Isidore is linked to the LOD by using Semantic Web Technologies.

So, you put your data in NAKALA, it is safe, Huma-Num takes care of that. Then you can advise people your data are here by connecting Isidore. Isidore will harvest metadata, classify them and disseminate them. ISIDORE can be viewed as a specialised search engine, but Google is also really fan of Isidore semantic classification. We also have a Triple Store In ISIDORE as well as in NAKALA. NAKALA is totally built on Semantic Web technologies, there is no relational database.

Then, when your dataset is complete or finished, you can preserve it on very long term, for example now we deal with huge set of digitized manuscripts. We work with the French National Computing Center for Higher Education (CINES) which has the official mandate of long term preservation for scientific data - at the beginning it was dedicated to thesis in digital format). So, in France, data produced by researchers from public money are supposed to be, one day, on the National Archive. Long-term preservation is a specific process, based on the archive field. Basically we put it on special devices in the CINES and they take care of it, they take the responsibility of data by making copies and they try to re-read data every month; if the format is obsolete, they will convert it and this is a huge responsibility. So they make sure that in 20 years we can re-read and understand this data by adding metadata and context information. There is a technological part, but also an archivist one.

An example of good tool for today's interoperability: semantic web technologies
Even if it not easy to use for people, it is perfect for machines as it works greatly to exchange data and to link data to other repositories too. We provide tools for hosting Triple Store content, the basis of semantic web. The idea of NAKALA and ISIDORE is that if you are a researcher you can put your data in NAKALA, or to make your repository harvested by ISIDORE, then you are automatically in the graph of semantic web, even if you don't know anything about semantic web.

We have a lot of links with French repositories like data.bnf.fr, which the National Library and it is remarkable to see that the National Library publishes this kind of data in RDF since 3 years whereas they are working with these technologies for about 10 years now. This is because in fact, you need a lot of curation work to permit people to publish and to maintain this kind open data repository.
There is also the DBpedia project, which is supported by the Ministry of Culture for the French part of Wikipedia. We also use geonames, lexvo and other repositories.
Nakala and Isidore provide a SPARQL EndPoint: for NAKALA it is considered as the API to access the metadata associated to NAKALA's TripleStore.

NAKALA is in fact associating different bricks, which you can replace. For example if you want to get rid of this TripeStore which is today very trendy, you can replace it; if you want to change storage, it’s possible; if you don't want to use an other OAI-PMH software, you can change it etc.
NAKALA is organized around a TripleStore, we don't have any database, so we will be able to switch to another technology in the future.

So we deal with data, any kind of data: it can be code, archeological data in a zip file, voice recording. Still NAKALA provides with some tools to suggest good format, for example when you add a picture, you have a tools that checks formats. To decide what is *right or wrong*, we rely on the work done by the CINES, which maintains a list a recommended formats. You can add Filemaker projects, but we suggest that it could be better to prefer another more reliable on long-term format. The only requirement is to give at least four elements of metadata which are title, author, date, type. Deposited data are securely stored and a PID is attributed to it. Figshare uses DOIs, we give handle. It is the same old technology. For metadata, you can access it through the TripleStore. Everything in NAKALA is organised around TripleStore. So as soon as you add your data to NAKALA, it is accessible.
The curation is made when you ask for an access to NAKALA: Huma-Num does an evaluation of scientific goals of the projects as well as the future of data (e.g. openness, licence etc.).

Data deposit in NAKALA

There is a web interface, not as sexy as Figshare, still it is easy for researchers to add for example 20 or even 50 videos and to share it. If you have more expertise and more data, for example an archeological project in Egypt uploaded 120 000 pictures from the walls of Karnak, you can use batch processing, add metadata in XML, make a packet and process it

into NAKALA. PIDs are attributed to data and metadata: the “handle” technology for PID (used also for DOI) gives two possibilities to access data: using a NAKALA URL or a regular handle URL.
For the access, right now, we don't provide any specific API because the TripleStore is supposed to be the API, but it might evolve in the future.
To get back to the main topic of the talk, we can say that NAKALA is not really a platform, but more an infrastructure. For instance, it doesn't provide any visualization tool, the idea is to have a simple repository. Huma-Num provides on the long term the maintenance of the handles system because the citation is really useful for data and metadata. And we also hope to provide with usage statistics: a good measure of the infrastructure use.

To display your data, you will need something else built above NAKALA, for example you can use Wordpress provided by hypotheses.org. Huma-Num provides a CMS OMEKA to display easily your set of data plugged with NAKALA that we called NAKALONA. For example, a French research center in Kenya made a 40 years press archive and added it to NAKALA. We had several exchanges with the persons in charge of this work and we decided to use a batch processing because there are more than 10 0000 pages. They want to share it at large, to show it, so we used NAKALONA and within minutes it was possible to show it or to search with metadata (see https://ifrapressarch.nakalona.fr).

Others can develop their own interface, etc. And Isidore can be considered as an interface. There is no limitation in term of size of data. We use a distributed technology called Active Circle that allow to aggregate parts and make an abstraction of the physical storage, so the size of the storage is not a problem. Anyway, we now think about providing more curation, more advices about good practices, helping people even to prepare data to go to NAKALA, but there is an important human cost. The issue is not to have only good tools or infrastructure, it is also to make people use them.

=> The problem is that there is no reward, no incentive for properly sharing data (with sufficient metadata). The only incentive is to be cited so today it is like a loss of time for a career. For Figshare, it is a question they have. For the self-deposit website, the main difficulty is to get people adding metadata in addition to just dropping files with no description. Sometimes, if they see that a dataset has an important download rate, they contact the researcher depositor and ask him to improve it, but it is not systematic, it is a kind of *download driven metadata*.
For publication, in ISIDORE, we can try to find if there is a PID related to a dataset and add a link to metadata. Like this, we try to build links between datasets and publications in the sense of RDF and semantic web graph.

Incentives for open data
In France, when you apply in humanities through the National Research Agency (ANR), you can declare that you are in touch with an infrastructure, but it is not yet mandatory. However with DMPs in H2020 projects, things are changing a little bit. Funders are aware that data cost a lot of money, so they need to preserve it and to share it. In France, it is just the beginning.

Contact

Nicolas Larrousse is head of the long-term archiving department at Huma-Num, a French infrastructure which aims to provide services to researchers in social sciences and humanities. He is particularly focused on interoperability and is involved in European infrastructures and projects. Huma-Num is promoting collaboration and providing services to manage, enrich and expose research data through a wide network of partners and consortia. Huma-Num is the National Coordinating institution of DARIAH European infrastructure for France and is involved in H2020 European projects.

Huma-Num: http://www.huma-num.fr/

Twitter: @Huma_Num

Email: Nicolas.Larrousse@huma-num.fr

10-Social impact

Gregory Crane is an Alexander von Humboldt Professor of Digital Humanities at Leipzig University.

In the 1980's, we thought that Greek and Latin texts needed to be online, in an open way, so people could use them with the goal, as being humanist, to transcend transnational cosmopolitanism. Since then, interesting projects have emerged and improved our knowledge. There is for example the coordination project OCR-D, which is aimed at the development of methods of Optical Character Recognition (OCR) for printed historical material. We can today try to extend these tools to quotation detection, networks and mapping, or the analysis of social network data.

Open Greek and Latin, a subset of a global philology project

Greek and Latin remain the central topic of the great national trends of the cultural heritage of Europe. I would like to emphasize the importance of Latin as a transnational aspect of European identity, as an ideal of unified language, which all Europeans can share and which is, on the same time, nobody's language. Scholars and scientists can help transcend short-term political issues. The rise of the great national vernacular languages is very interesting as it shows the choice between using a national language and having a limited audience or the adoption of a language with a broader impact. Open Greek and Latin is in a sense a large-scale open data project as we have reached the first million OCRed books downloaded from Internetarchive.org and this allows to trace translations over time.

Philology

Philology is the cognitio, the mental understanding process of the universal past, historical and philosophical. But it can also be understood in a narrower way as being the preparation of editions, the analyze variances, etc. I see it very differently: anything you can do with a textual record that allow to reconstruct anything that happened in human mind or in the world around us. By definition, it is expensive, never satisfied and has no inherent methods. Methods can evolve depending upon the question. And statistical methods can help our understanding of the past. So, philology is also data driven, every statement is directly backed with the primary evidence which upon the statement is based. Within philology, mechanism for citation is essential as you have to be able to reference any word, symbol, element of any surviving text or object (see Homer Multitext Project documentation)

Humanities and visualisation

1. Difference in the proof and discovery in the Humanities opposed to other scientific analytic fields
2. Evaluation for humanities questions that may have no ground truth?
3. DH and text visualisation scenario

We today have a lot of interesting emerging projects and tools that can address text analysis for data sets or social media, in a multilingual way, but the question is about the usage of this tremendous potential. What kind of questions can you ask given this level of heterogeneous and vast data?
You now have huge databases, such as Internet Archive.org, Europeana, Gallica from the French National Library, etc., but the question then is the connection with national databases on a larger scale. Germany has digitised and carefully scanned about 400 000 books printed in Germany or in German outside of Germany from 1500 to 1800. When it is OCRed, the implications would transform the way we have to think about the history of thought, as we would have a bigger scale. We now have tools that allow the visualization of automatically detected text reuse for a document in millions of books, which is absolutely nice for the humanities. This is really something that should happen to all our documents.
=> Digital Humanities are the space of creative destruction where students of the humanities are forced to rethink their larger goals in light of the challenges and possibilities of the digital world, beyond the fixed stream of print on the page.

One opportunity with big data is the use of geotags and place-names as it allows to generate a map from the content of a document, and its relations, it allows to see things you couldn't see before.
=> This is why it is interesting to get access to the source data, but, the problem with big data is the multiplicity of languages and the volume, which cannot be done by hand anymore.

You can now produce visualisations of cluster of words that statistically co-occur. These clusters often correspond to a topic, but the ground problem then is to know what you can do with that. You can also use bi-lingual text display tool that helps understand the structure of a text with grammatical and syntactic relation of every word; you can soon figure out how to read a text in another language. This is an environment where you can do something with a text that 30 years ago was completely inaccessible. This is beyond translation because you can see all the functions of words. This is one way to understand data and open the barrier of language for inspection. And it can also be applied to music as a textual object (with score), or to mathematics.
How far can you get? How do you generate data that you need to understand? How do you do syntactic analysis at scale? How do you make parallel text alignment automatically?
We have examples of students and citizens who are voluntarily producing a data driven manuals and this illustrates a new form of intellectual production with distributed work and decentralizing power. It is in fact a democratic ideal because there is no academic reward or financial credit in this involvement.

Global Philology Planning Seminar Report

We have just created a Bachelor of science in DH in Leipzig University that aims the integration of computer science with humanities work and we are also organising an open conference. It tends to gather disciplinary needs and specificity to determine the structures and organisations we need.

Europeana, in a sense, seems to be stuck at the metadata level for further integration. Full integration will only happen when data circulates, otherwise it is just metadata. My goal here is to understand better how we can collaborate better than we do and to share sustainable development to connect resources and components. We are trying to have a list of core services and every historical document should have these services applied to it, to allow good things to happen.
For example, we could see all the places that were quoted in a document. All the names, every word should be analyzed. You should have syntactic analysis for all the words, you should be able to align all different versions that you just digitized and to compare them.
It would be nice too to have text alignment across languages, ideally you could discover if there were translations of this work and align them. We could also produce on the fly lexicon for any word, use automated text mining, sentiment detection and lexicography to see patterns emerge.

Resources

- Global Philology Open Conference, Feb 2017: http://www.dh.uni-leipzig.de/wo/events/global-philology-open-conference/
- Open Greek and Latin as open data:
  - http://www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project/
  - https://github.com/OpenGreekAndLatin
  - http://www.culingtec.uni-leipzig.de/ESU_C_T/node/379
- Patrologia Graeca:
  - https://mimno.infosci.cornell.edu/patgrec/
  - https://mimno.infosci.cornell.edu/
- Homer Multitext Project: http://www.homermultitext.org/
- Perseus Project: http://www.perseus.tufts.edu/hopper/
- CITE/CTS architecture:
  - http://www.homermultitext.org/hmt-doc/cite/index.html
  - http://cts.dh.uni-leipzig.de/
- Alpheios Reading tools: http://alpheios.net/
- Epidoc: https://sourceforge.net/p/epidoc/wiki/Home/
- Book alignement
  - http://books.cs.umass.edu/mellon/alignment.html
  - http://books.cs.umass.edu/beta-sprint/Demonstration/Demonstration.html
- Transkribus Project: https://transkribus.eu/Transkribus/

Contact

Gregory Crane is an Alexander von Humboldt Professor of Digital Humanities at Leipzig University. He is a specialist in classical philology and computer science.
He completed a doctorate in classical philology at Harvard University and worked as an assistant professor. He has the reputation of being a pioneer of digital humanities due to his development of the Perseus Digital Library, a freely accessible online library for antique source material. He was associate professor at TUFTS University and is now Winnick Family Chair of Technology and Entrepreneurship. He has received, among other awards, the Google Digital Humanities Award 2010.

Institutional website: http://www.dh.uni-leipzig.de/wo/gregory-crane/
Email: crane@informatik.uni-leipzig.de

Bibliography

Altman, Micah, and Mercè Crosas. “The Evolution of Data Citation: From Principles to Implementation.” *IASSIST Quarterly*, 2013. http://www.iassistdata.org/iq/evolution-data-citation-principles-implementation.

Austin, Claire C, Bloom, Theodora, Dallmeier-Tiessen, Sunje, Khodiyar, Varsha, Murphy, Fiona, Nurnberger, Amy, Raymond, Lisa, et al. “Key Components of Data Publishing: Using Current Best Practices to Develop a Reference Model for Data Publishing,” 2015. doi:10.5281/zenodo.34542.

Bargheer, Margo. “Practical Experiences at Göttingen University in Providing Open Access Services to Researchers, Scholars and Faculties.” May 20, 2015. http://zenodo.org/record/46060.

———. “The Dissertation Almost Done, but How to Publish Now?” February 15, 2016. http://zenodo.org/record/46059.

Birnbaum, David. “What Is XML and Why Should Humanists Care? An Even Gentler Introduction to XML.” Accessed October 19, 2016. http://dh.obdurodon.org/what-is-xml.xhtml.

Bordier, Julien. “Open Peer Review : From an Experiment to a Model,” February 2016. https://hal.archives-ouvertes.fr/hal-01302597.

Borgman, Christine. “Data Attribution and Citation Practices and Standards.” October 27, 2011. http://sites.nationalacademies.org/cs/groups/pgasite/documents/webpage/pga_066705.pptx.

———. “HarvardDataCitationSymposiumBorgman20120516.pptx - harvarddatacitationsymposiumborgman20120516.pdf.” Accessed April 25, 2016. http://projects.iq.harvard.edu/files/attribution_workshop/files/harvarddatacitationsymposiumborgman20120516.pdf.

Borgman, Christine L. “The Conundrum of Sharing Research Data.” *Journal of the American Society for Information Science and Technology* 63, no. 6 (June 2012): 1059–78. doi:10.1002/asi.22634.

———. “Why Are the Attribution and Citation of Scientific Data Important? In: Uhlir, Paul and Cohen, Daniel (eds.). Report from Developing Data Attribution and Citation Practices and Standards: An International Symposium and Workshop.” *National Academy of Sciences’ Board on Research Data and Information. National Academies Press: Washington DC. Http://www.nap.edu*, 2012. https://works.bepress.com/borgman/265/.

Buddenbohm, Stefan, Nathanael Cretin, Elly Dijk, Bertrand Gaiffe, Maaike De Jong, Jean-Luc Minel, and Blandine Nouvel. “State of the Art Report on Open Access Publishing of Research Data in the Humanities.” Report. DARIAH, August 12, 2016. https://halshs.archives-ouvertes.fr/halshs-01357208/document.

“Canonical Text Services Protocol Specification.” Accessed September 9, 2016. https://cite-architecture.github.io/cts_spec/.

Christine, Berthaud, Capelli Laurent, Gustedt Jens, Kirchner Claude, Loiseau Kevin, Magron Agnès, Medves Maud, Monteil Alain, Riverieux Gaëlle, and Romary Laurent. “EPISCIENCES - an Overlay Publication Platform.” *Stand Alone*, 2014, 78–87. doi:10.3233/978-1-61499-409-1-78.

“CLARIN - Persistent and Unique Identifiers.” Accessed September 2, 2016. http://www-sk.let.uu.nl/u/D2R-2b.pdf.

Cliggett, Lisa. “The Qualitative Report 2013 Volume 18, How To Article 1 , 1 - 11 http://www.nova.edu/ssss/QR/QR18/ cliggett1.pdf.” Accessed April 18, 2016. http://www.nova.edu/ssss/QR/QR18/cliggett1.pdf.

Corti, Louise. “Progress on Open Data Publishing in the Social Sciences,” October 30, 2014. http://dspace.ut.ee/handle/10062/44055.

Dávidházi, Péter, ed. *New Publication Cultures in the Humanities: Exploring the Paradigm Shift*. Amsterdam: Amsterdam University Press, 2014.

djenzar. *Robot*. Photo, October 25, 2015. https://www.flickr.com/photos/djenzar/26431038934/.

“Electronic Sources and Locator Information - Electronic-Sources.pdf.” Accessed April 19, 2016. http://www.apastyle.org/manual/related/electronic-sources.pdf.

“Four URLs.” Accessed September 9, 2016. http://folio.furman.edu/projects/cite/four_urls.html.

“H2020 Programme Guidelines on FAIR Data Management in Horizon 2020.” Accessed September 2, 2016. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

Joachim Schöpfel, Stéphane Chaudiron, Bernard Jacquemin, Hélène Prost, Marta Severo, and Florence Thiault. “Open Access to Research Data in Electronic Theses and Dissertations: An Overview.” *Library Hi Tech* 32, no. 4 (November 11, 2014): 612–27. doi:10.1108/LHT-06-2014-0058.

“Joint Declaration of Data Citation Principles - FINAL.” *FORCE11*, October 30, 2013. https://www.force11.org/group/joint-declaration-data-citation-principles-final.

King, Gary. “Restructuring the Social Sciences: Reflections from Harvard&#039;s Institute for Quantitative Social Science.” *PS: Political Science and Politics* 47, no. 1 (2014): 165–72.

Konkiel, Stacy. “Tracking Citations and Altmetrics for Research Data: Challenges and Opportunities.” *Bulletin of the American Society for Information Science and Technology* 39, no. 6 (August 2013): 27–32. doi:10.1002/bult.2013.1720390610.

Kratz, John, and Carly Strasser. “Data Publication Consensus and Controversies.” *F1000Research*, April 23, 2014. doi:10.12688/f1000research.3979.1.

Margo, Bargheer, and Schmidt Birgit. “Göttingen University Press: Publishing Services in an Open Access Environment.” *Information Services and Use*, no. 2 (2008): 133–39. doi:10.3233/ISU-2008-0569.

Maxwell, John W. “Beyond Open Access to Open Publication and Open Scholarship.” *Scholarly and Research Communication* 6, no. 3 (October 22, 2015). http://src-online.ca/index.php/src/article/view/202.

Mišutka, Jozef, Amir Kamran, Ondřej Košarko, Michal Josífko, Loganathan Ramasamy, Pavel Straňák, and Jan Hajič. “Linguistic Digital Repository Based on DSpace 5.2.” *Https://github.com/ufal/lindat-Dspace*, 2015. https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-1481.

Mooney, Hailey, and Mark Newton. “The Anatomy of a Data Citation: Discovery, Reuse, and Credit.” *Journal of Librarianship and Scholarly Communication* 1, no. 1 (2012). doi:10.7710/2162-3309.1035.

Moore, Samuel, ed. *Issues in Open Research Data*. Ubiquity Press, 2014. http://www.ubiquitypress.com/site/books/detail/12/issues-in-open-research-data/.

Mounier, Pierre, Laurence Allard, Luxembourg) Équipe du Centre virtuel de la connaissance sur l’Europe (CVCE, Pierre Grosdemouge, Marion Lamé, and Fred Pailler, trans. *Read/Write Book 2 : Une introduction aux humanités numériques*. Read/Write Book. Marseille: OpenEdition Press, 2012. http://books.openedition.org/oep/226.

“OECD Principles and Guidelines for Access to Research Data from Public Funding - OECD.” Accessed September 2, 2016. https://www.oecd.org/sti/sci-tech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm.

“Open Access for East and West.” Accessed September 15, 2016. https://sdvigpress.org/projet.

“OpenAIRE’s Experiments in Open Peer Review / Report.” *Zenodo*. Accessed September 28, 2016. doi:10.5281/zenodo.154647.

Pienta, Amy M., George C. Alter, and Jared A. Lyle. “The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data.” In *Organisation, Economics and Policy of Scientific Research” Workshop, Torino, Italy, in April*, 2010. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/78307/pienta_alter_lyle_100331.pd?sequence=1.

Preparing the Workforce for Digital Curation. Washington, D.C.: National Academies Press, 2015. http://www.nap.edu/catalog/18590.

Prost, Hélène, Cécile Malleret, and Joachim Schöpfel. “Hidden Treasures: Opening Data in PhD Dissertations in Social Sciences and Humanities.” *Journal of Librarianship and Scholarly Communication* 3, no. 2 (2015). doi:10.7710/2162-3309.1230.

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.

Research Data Netherlands. *Persistent Identifiers and Data Citation Explained*, 2014. https://www.youtube.com/watch?v=PgqtiY7oZ6k&feature=youtu.be.

Romary, Laurent, Mike Mertens, and Anne Baillot. “Data Fluidity in DARIAH – Pushing the Agenda Forward.” *BIBLIOTHEK Forschung Und Praxis* 39, no. 3 (2016): 350–57. https://hal.inria.fr/hal-01285917/document.

Romary, Laurent, and Marie Puren. “Datasets of IPERION CH,” 2016. https://hal.inria.fr/hal-01289058/document.

Roorda, Dirk, and Charles van den Heuvel. “Annotation as a New Paradigm in Research Archiving.” *ResearchGate*, October 7, 2014. https://www.researchgate.net/publication/269711156_Annotation_as_a_New_Paradigm_in_Research_Archiving.

Schöpfel, Joachim, Južnič Primož, Hélène Prost, Cécile Malleret, Ana Češarek, and Teja Koler-Povh. “Dissertations and Data.” In *GL17 International Conference on Grey Literature*, 2015. http://hal.univ-lille3.fr/hal-01285304.

———. “Dissertations and Data,” 2015. http://hal.univ-lille3.fr/hal-01285304/document.

Schöpfel, Joachim, and Hélène Prost. “Research Data Management in Social Sciences and Humanities: A Survey at the University of Lille (France).” Accessed August 24, 2016. http://edoc.hu-berlin.de/libreas/29/schoepfel-joachim-71/PDF/schoepfel.pdf.

Schöpfel, Joachim, Hélène Prost, and Cécile Malleret. “Making Data in PhD Dissertations Reusable for Research,” 2015. http://hal.univ-lille3.fr/hal-01248979/document.

Schöpfel, Joachim, Hélène Prost, and Violaine Rebouillat. “Research Data in Current Research Information Systems.” euroCRIS, 2016. http://dspacecris.eurocris.org/handle/11366/501.

Silvello, Gianmaria. “A Methodology for Citing Linked Open Data Subsets.” *D-Lib Magazine* 21, no. 1/2 (January 2015). doi:10.1045/january2015-silvello.

Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, et al. “Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications.” *PeerJ Computer Science* 1 (May 27, 2015): e1. doi:10.7717/peerj-cs.1.

“The Canonical Text Service (CTS) · The CITE Architecture.” Accessed September 9, 2016. https://cite-architecture.github.io/cts/.

“The OHCO2 Model of Text · The CITE Architecture.” Accessed September 9, 2016. https://cite-architecture.github.io/ohco2/.

Uhlir, P. F., National Research Council (U.S.), eds. *For Attribution--: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop*. Washington, D.C: The National Academies Press, 2012.

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.