nccr on the move

Data Management at UniNe

This document outlines some aspects that underscore the benefits for the University of Neuchâtel to develop an expertise point on data management. These lines are drawn from the experience of the nccr – on the move, which has implemented such a position since late 2014.

Archive and networked archive

With the use of electronic data, researchers face new challenges when it comes to accessing existing data or bringing new one into the world, and making sure it stays there. Fortunately, many solutions have emerged from solving painful misadventures, they reach from adapting the organization of global and local infrastructures to the development of best practices at the researcher's level.

Long-term preservation of digital documents

In contrast to paper, digital files do not age well; storage media get erased, file formats become obsolete, programs to handle them are no longer available and even devices used to read the records become scarce. Whenever such events happen, the cost of recovery is very high. These costs can be avoided when users acquire specific work methods that are oriented towards preservation of essential data. This means being aware of the limitations of the devices used in daily work, having access to adequate storage infrastructures and choosing the right formats and strategies. A specific expertise and guidance is needed in order for these digital files to become suitable for archival.

Inter-operability of archive catalogs

Institutional repositories are one of the ways to insure the long-term availability of digital data. On one hand the documents they ingest must follow specific rules defined by the repository, on the other hand, the repository needs to follow specific archiving standards[1] so that the catalog remain "harvestable" from other repositories and partners. This remains true weather the files are statistical datasets, publication references or any collection of information. Although the standards are widely published, they are more commonly circulating among archiving community than the providers of IT solutions, and the University of Neuchâtel would benefit from a better awareness of these standards.

Inclusion of documentation standards

Adoption of standards eases the institutional archival of data, it also facilitates the exchange of research output among researchers and helps structuring the advancement of scientific projects. When scientists organize their survey data along the Data Documentation Initiative, the implementation of surveys is simplified, dataset become readable across technological platforms. With the adequate support, researchers can document their deliverables in open formats that are easily derived to produce high quality codebooks. This knowledge is not trivial and requires a tight collaboration with the research teams.

[1] https://www.openarchives.org/pmh/

Digitalization of research tools

Effects on research ethics

The relationship between the researchers and the population they study is affected by the digitalization of data collection and dissemination, mostly by the ease in which potentially confidential information circulates. Anonymisation techniques become crucial, as well as early collection of informed consent. Working on social media data, documents or databases also requires special care with intellectual property. Researchers are looking for guidance in this field as well. Multiple aspects of their practices are also affected by the data sharing requirements, with issues such as disappearing of embargo periods or mandatory use of open access publishers.

Move towards digital humanities

Social scientists work increasingly with complex data such as Twitter feeds, relational databases or image banks. When working with traditional literal media, they apply standardized protocols for text mining and tagging or create augmented editions. This movement crosses over disciplines: network analysis concerns econometrics as well as philosophy, interactive visualizations are useful both for anthropologists and legal scholars. Some tools might be area specific (geography, text transcriptions), but more than often can be applied in many different settings. Again, there is here an area of expertise that needs to be made available to the researchers at the University of Neuchâtel.

Extension of the syllabus

The shift towards digitalization requires more from the research community than being fluent with office productivity suites. Students routinely learn specific tools of their discipline (Geographic information systems, Computer-assisted qualitative analysis software, Statistical software), but they now need to be trained in the production of data visualization, data documentation, sustainable file formats, interaction with repositories, the fellows need also guidance on data ownership and publication strategy. Since output is no longer limited to article or monographies, scientists must also be able to produce convincing posters and slide productions used in scientific communication.

Promotion

Institutional visibility

The set of communication tools used by the University of Neuchâtel works well within the institution, the websites and intranet serve adequately the local needs. On a broader level, the institution is networked with external actors at the national or international levels that today are not able to ingest the specific local information. The Swiss National Scientific Fund needs to access a publication catalog structured in a specific way in order to define if publications follow its open access guideline, Datasets produced in Neuchâtel need to be formatted in specific ways so that web harvesters around the world can see them, this will also serve actors outside the academic world, such as data journalists who are also increasingly using production issuing from scientific research.

Promotion of individuals and networked projects

Local projects are more than often part of collective initiatives, whether or not funded as such. The SNSF's nccr projects or European framework programs all produce deliverables that all need to be identified,

findable, archived and ready for dissemination. This is much easier when catalogues of data, publication and other results follow common standards and data collections follow consistent guidelines.

At an individual level, researchers can (and should) have web pages where the consistency of their research is made visible. When the information follows bibliometric and other documentation standards, the individual curation of personal collections is easily moved between contents, from institutional to personal websites and becomes harvestable by external indexing tools.

Promotion of individual researchers

Recent efforts in scientometry have led to a considerable quantification of individual research activity. Researchers can (and should) have an active role in their self-promotion, favoring open platforms for their publications, supporting their findings with well documented datasets, producing results that are ready for knowledge transfer. Being present in research inventories, data repositories, using Digital Objects Identifiers are also strategies that promote the individual researchers, independently from their trajectory between academic institutions.

Requirements of funders and editors

Data management plans

Most sponsors of research expect academics to be able to describe what kind of data they intend to collect, the ethical, legal and technical challenges that their collection imply, and also how they insure that the investment made in these data collections is preserved. Data management plans (DMP's), when collected and structured along common lines help both the institution and the scientists conducting research to avoid breach of confidentiality or loss of data. Again, this is not a trivial endeavor, and although disciplines tend to produce guidelines to help answering such requirements, researchers need to be trained and offered an adequate expertise in order to meet those demands.

Data availability

In addition to the administrative request of sponsors to make the data produced with their funding available for further research, comes now the demand of scientific journals to access data prior to article publication. Accessing data serves both purposes to replicate results and offer further secondary analysis. Defining which data is to be made available varies from discipline to discipline. Econometric and survey data are more straightforward than ethnographic data, the latter being best discussed with the researchers. Even partial data contributes to the research, as a deposit in an institutional repository usually comes with metadata describing the research and contributes to promoting the results.

Availability of publications

As for their sustaining data, scientific publications themselves need to be available. This request is now generalized from the Swiss National Fund and will only broaden in the future. This implies an increasingly complex set of rules combining the status of the publications (Author Accepted Manuscript, etc..), the economic model of the journals (On subscription or Open Access), as well as individual agreements between academic institutions and publishers. Where to publish and under which conditions are no longer the decisions of the sole researchers; they need additional specific information in order to accomplish this basic scientific activity.

Examples from the nccr – on the move experience

Since the first phase of the project, Data Management at the nccr – on the move has contributed to the production of science. Considering this relatively recent field of activity, the deliverables have very well matched the organizational and individual expectations.

Data management strategy

Every single of the 16 sub-project of Phase II has elaborated a detailed data management plan. The fully structured global document spans over 74 pages, validated by the SNF, publicly available and regularly updated. This is a result of a global strategy within the nccr network office, intense communication and individual support for all disciplines involved. The DMP's respect both the formal requirements of our funding agency and the diversity of the activity of researchers based across Switzerland. Although the information is managed in a common relational database, DMP's circulate as PDF's to insure their long-term preservation.

Durability of deliverables

In connection with the DMP's, the project's deliverable data is almost entirely documented along the open source Data Documentation Initiative standard, both as xml and pdf files (codebooks between five and 460 pages). Forty-three datasets are available at the time of this writing. Most datasets are available on the Zenodo.org (CERN)[2] repository, some on the OSF, EUI, Dataverse and Fors, along the wishes of individual researchers. We keep an electronic library (Zotero)[3] listing our more than 500 publications, enriched with specific tags of our project. A general linked inventory[4] (in PDF form) acts as a hub allowing access to any dataset, publication or DMP produced during the project.

Expertise and Training

During the course of the eight last years, the data manager has been involved with virtually every single project of the nccr – on the move. Demands for support have spanned from suggesting software to solve specific problems to the redaction of data management plans and ethical controversies. He has also been active in disseminating the knowledge of data management by offering multiple workshops addressing issues such as scientific posters, data documentation, open access, sponsor requirements, cartography or bibliography. In the last part of Phase II, five post-doc or Phd fellows have been trained to act as data management delegates with the task to produce standardized dataset documentation for their local research groups.

Visualization

Thanks to specific competences of the person holding the position, data management at the nccr – on the move also involves data visualization. We have created more than forty interactive visualizations[5], either part of our indicator project or as an individual support to specific projects. Many are results of international collaborations and cover the needs from geographers to legal scholars. They have been consulted thousands of times and many are associated with scientific publications and disseminated datasets. These visualizations are embedded in many webpages of the nccr – on the move website.

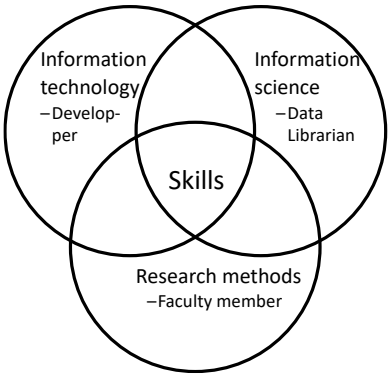
[2] https://zenodo.org/communities/nccr-onthemove/search?page=1&size=20&q=
[3] https://www.zotero.org/groups/393573/nccr-onthemove/tags/SNSF/library
[4] https://nccr-onthemove.ch/DataManagement/Inventory.pdf
[5] https://public.tableau.com/app/profile/nccr.on.the.move

Skills, technology and tools

The expertise point combines skills, technologies and tools that originate in distinct areas, as well as traditionally distinct organizational in the University of Neuchâtel. Although the position is strongly oriented towards computing professions, it has an overwhelming academic role to play and builds on expertise that generates in the world of librarians.

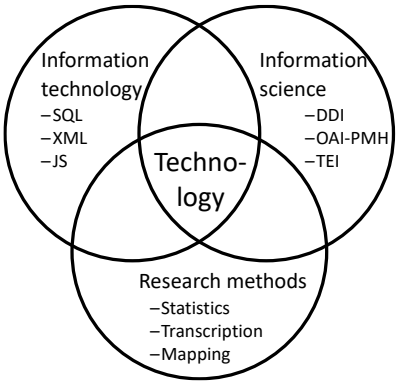
The Data Manager takes after three professions; the IT developer in Information Technology, the librarian from the Information Science, and the faculty member in the research methods.

Each domain brings specific knowledge, each domain has a set practices that is complementary.



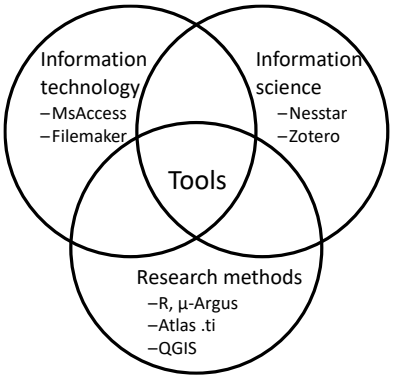
Data Management is rich in jargon. Inherited from their respective domains, the technologies bring either methods of standardized collaboration (Information Science), programming tools (Information Technology) or methodologies.

Methods define how research is done, standardization how the knowledge is shard and the tools provide the means to achieve this.



Each domain also carries emblematic tools used in Digital Humanities, some dedicated to knowledge sharing such as documentation or bibliographic software, some are essentially specific to disciplines, such as µ-Argus (anonymization), GIS or transcription tools.

Many generic tools such as relational databases are also handled by the Data Manager, for instance to generate the integrated inventory.



For more information, please contact Andreas Perret, Data Manager (andreas.perret@nccr-onthemove.ch)

apt / 02.03.2022