Data Life-Cycle Management Project: SUC P2 2015-2018

Eliane Blumer
Eliane.Blumer@unige.ch
https://orcid.org/0000-0002-0972-5396
Division Systèmes et technologies de l'information et de la communication (STIC), Université de Genève

Pierre-Yves Burgi
Pierre-Yves.Burgi@unige.ch
https://orcid.org/0000-0002-4956-9279
Division Systèmes et technologies de l'information et de la communication (STIC), Université de Genève

Abstract

This text is about the Swiss University Conference (SUC) program 2 (P2) project entitled "data life-cycle management" (DLCM), which is partnered with library-, IT-, research-, and/or Life Sciences and Digital Humanities-departments from eight Swiss Higher Education Institutions (EPFL, HEG / HES-SO, UNIL, UNIBAS, UNIZH, ETHZ, UNIGE and SWITCH). The project, accepted in August 2015, aims at offering services to the Swiss research community to help them manage their research data more conveniently. Herein, after introducing the research data management topic, we present the methodology and organization of such a comprehensive project, starting by collecting researchers' opinions, expressing them into needs, and ending with a detailed description of what the already existing and future outcomes of the project will be, in 2018.

1. Introduction

With “datafication” of the research sector, we are living in a data deluge era and thus have to address data management before it gets completely out of control. Internationally, the trend to optimize research outcomes by providing the scientific community with access to the original data and knowledge required to reproduce published results strongly contributes to find new solutions to data management. However, caring for scientific data, ensuring their reuse in the future for yet-unknown applications, and ensuring their accessibility and understandability over time requires deep-rooted strategies and established practices and tools that appeal to the digital curation field [1], of which the key elements will be addressed in this manuscript.

Many funding agencies and publishers now impose accessibility to research data in a form that allows reproducibility, but also permits reuse of the generated knowledge. Funding bodies in Europe and USA have therefore started large programs aiming at providing researchers with institutional tools and services to manage, store and give access to their data. For instance, the Economic and Social Research Council (ESRC) in UK has a research data policy since 1995, where issues about data management are cited. In its current version, a data management plan (DMP)(1), a document describing the research project and the handling of data from collection to preservation, stipulates(2): “Planning how you will manage your research data should begin early on in the research process and Economic and Social Research Council applicants who plan to generate data from their research must submit a data management plan as part of their Joint Electronic Submissions application.” Following the same path, in 2003 the National Institutes of Health in USA has mandated data sharing for most grants receiving over $500’000 a year, as specified in its own data sharing policy(3). Moreover, the Organization for Economic Co-Operation and Development (OECD) published in April 2007 its Principles and Guidelines for Access to Research Data from Public Funding(4), a document containing directives for facilitating cost-effective access to digital research data from public funding. Over the past five years, the number of federal agencies with DMP guidelines for grant proposals has been increasing rapidly. For instance:

- 2010: the German Research Foundation mentioned the importance of accurate management of research data(5);
- 2011: the National Science Foundation in USA set up a DMP requirement for all grant proposals(6);
- 2013: the White House Office of Science and Technology Policy demanded that all agencies receiving $100 million or more in research and development funds submit policies for data sharing in order to increase the availability of the products of federally funded research(7);
- 2013: the European Commission published its Guidelines on Data Management in Horizon2020(8) and the League of European Research Universities (LERU) produced a Roadmap for Research Data, which provides a guide for research-intensive universities engaged in data-driven research(9).
- 2014: the Department of Energy in USA requires DMPs in order to apply for a grant(10).

In Switzerland, the *Swiss* National Science Foundation (SNSF) has not yet asked for such DMPs, but recently, its CEO (Martin Vetterli) mentioned during an Open Science Day in

October 2014 at the Swiss Federal Institute of Technology – Lausanne (EPFL) that the SNFS is following the issue very closely(11).

To catch up with these still foreign but approaching trends in dealing with data management, we have set up a national project involving eight institutions (EPFL, HEG / HES-SO, UNIL, UNIBAS, UNIZH, ETHZ, UNIGE and SWITCH). After an extensive period of maturation (almost two years), on September 1st 2015 the kick-off meeting of the “Data Life-Cycle Management” project took place in Bern. This ten-million Swiss Francs and three-year long project aims first and foremost at offering new services to researchers to help them handle their research data. Given that each partner institution has major issues in data management to resolve, the SUC P2 national program provided us with the just-in-time opportunity to gather our respective forces and to grapple with this important and complex topic so as to serve the whole Swiss research community.

Before describing in more details the project, we start with some definitions of what we mean by “data life-cycle management” and what are the links with the digital curation field. We then address the researchers’ needs and discuss how the DLCM project intends to respond to these needs.

1.1. Data Life-Cycle Management in research

In the studied context, “data life-cycle management" (DLCM) refers to the various steps of a specific research process, which can be summarized as: *planning, collecting, analyzing, preserving, publishing and re-using scientific data*(12).In the sense that we suppose data are digital (which is not necessarily always the case), such a process is akin to digital curation [1], which deals with a set of techniques that address data management on electronic platforms with an emphasis on their maintenance and some added value to them for current and future use. In any case, data management practices cover the entire life cycle, from planning the investigation to conducting it and from backing up data as it is created and used, to long-term preservation of data after the research investigation has concluded. Well thought-out data management is therefore increasingly important due to the evolution of research practices, advances in research equipment, tools and data formats as well as new technologies, which permit sharing and collaborating on international scales(13). Recently developed concepts such as the DCC Curation Life-Cycle Model(14), help us to get a high-level overview of the topic with all the necessary stages involved in (digital) data curation [1]. DLCM is of course reminiscent of the three ages theory [2], in which data pass from the active status (data analysis and processing) to semi-active (e.g., data deposited on a repository linked to a publication), to finally end in a long-term preservation system (archived data, mainly with a passive status). The challenge resides in being able to delimit these stages in a scientific context where such archivistic notions are mainly lacking.

This is why effective data management asks for a variety of competencies provided by information specialists, information technology specialists and researchers. Furthermore, research departments, funding bodies, publishers and other partners may play an important role in data management as well [3,4,5]. Table 1 summarizes for each step the possible attributions of different roles as well as associated tasks. The main message of this table is that data life-cycle management is a teamwork effort, which is not necessarily perceived as such by researchers. That’s why, the implementation of data management asks for a cultural change as well as some flexibility and mutualization in the underlying tasks. This is not to say that research data will be curated following the archivist’s model for administrative documents.

Indeed, for instance who would select the content to archive and decide their conservation duration, if not the researchers themselves? However, digital curation principles should be applied to improve the quality of the information to be preserved whenever these principles are deemed reasonable within a scientific context, with the hope that the boundaries between scientific and administrative contexts will blend in the future as researchers become more acquainted with the curation techniques. This could happen in instances where a virtual research environment (for instances in Humanities) evolves towards a digital curation center. Certainly not all scientific disciplines apply for such a transformation, but for those suitable, then data life cycle could become a common language.

Table 1: The Data Life-Cycle steps and corresponding roles [6]. Note that the tasks and responsibilities might vary from one institution to another.

Step	Role	Possible actions
Plan	Research Support	Keeps up-to-date with latest changes in research politics, new synergies, new obligations
	Funding Bodies	Impose Data Policies
	Researchers	Know the latest changes in research politics, fill in proposals
	Libraries	Assist in filling in data management plans/proposals
Collect	Research Support	Provides best practice guidelines
	Researchers	Collect data according to their project
	Computing Services	Offer infrastructure (i.e. storage) for data collection
	Industrial partners	Offer LIMS/ELN solutions connected to repositories
Analyze	Researchers	Analyze their data, describe them according to pre-established DMP
	Computing Services	Offer infrastructure for data analysis, collaborate with preserving/publishing infrastructure
Preserve	Research Support	Points to guidelines
	Researchers	Hand data over to preserving infrastructures
	Library	Offers expertise for submitting datasets
	Computing Services	Offer technical infrastructure for long-term preservation

N° DOI : 10.55790/journals/ressi.2015.e1602

Publish	Library	Provides visibility to published datasets, along with citation mechanisms
	Publishers	Link data to associated publications
	Computing Services	Offer technical infrastructures for publishing
	Researchers	Decide on the basis of established policies the publication modalities (open vs. restricted, time period, etc.)
Reuse	Researchers	Predispose for data re-use and sharing
	Library	Facilitates data re-use through publication mechanisms

2. The researchers' needs

In preparation to the DLCM project, we wanted to identify the researchers' primary needs at each partner institution, along with the existing solutions in place in those participating institutions. For that, all DLCM partners conducted semi-structured interviews during two months (September and October 2014). The structure of these interviews contained four major parts, namely: (1) initial data and workflow, (2) analysis and data exploration, (3) publication, archiving and long-term data management, and (4) research data in the future: challenges, risks, perspectives. Table 2 presents the compilation of all interviewed disciplines. The results of these interviews have been profusely used to orient the project's deliverables.

Table 2: List of the interviewed disciplines during the preparation of the DLCM project.

Institution	Number of interviews	Disciplines
UNIGE	8	Theology, Informatics, Linguistics, German, Cognitive Neuroscience, Educational Sciences, Geomatics, Archeology, Vulnerability, Political Sciences, Medicine (Child Cancer Research), Political Sciences
ETHZ	8	Biosystems Science and Engineering, Seismic Networks, Sociology, Consumer Behavior, Quantum Optics Group, Scientific Computing/Photon Science, Physics
UNIL	15	Social Medicine, Social Sciences, Digital Humanities, Genomics, System Biology, Bio-informatics, Public Health, Imaging and Media Lab, Cancer Research

EPFL	5	Transport and Mobility, Quantum Optoelectronics, Supramolecular Nanomaterials and Interfaces, Audiovisual Communication Laboratory, Virology and Genetics.
UNIBAS	7	Biology research (Biozentrum, 2), Biology (Core facilities, 2), Molecular Psychology, Public Health (STPH), Digital Humanities
UNIZH	5	Law Science, Biology/Microscopy, Biology/Proteomics, University Hospital, Geosciences
Total	**49**	30 different disciplines

2.1. Results of the interviews

Every interview was entered into a summary table, organized by discipline and dispatched in the four main interview parts with a finer classification based on similarities in the answers when applicable. On this basis, an analysis of the main trends was performed, and citations deemed representative of the researcher's community were further extracted. This analysis shows that the outcome of the interviews are diverse, sometimes even contradictory, but clearly depend on the institutional strategies and/or research habits in the specific considered disciplines.

Generally, within the interviewed disciplines, researchers organize themselves for all matters concerned with description, sharing, storing, and publishing and archiving of research data, usually following their own appreciation and methods, as expressed by one interviewee(15): "*Yes! There is a high need to centralize data management. Everyone just does it as he/she likes."* Also, the degree of organization varies from, for instance the Physics departments, which are already well organized on international, European and/or disciplinary level, emphasizing that their current system is working and doesn't need to be changed, compared to other disciplines, such as Educational Sciences or Humanities, which are less organized and struggle with most of the DLCM issues. For instance, there can be a lack of storage possibilities, with the usage of own-bought servers stored on the floor or in inadequate rooms.

2.1.1. Initial data and workflow

Data loss is often mentioned as a main issue, as for instance an interviewee stating that *"It happens a lot that we lose data, when a PhD candidate leaves."* Also, generally no formal DMP are being used, which could help to better structure the research process and its outputs, unless the funding instances require it at the time of the project application. Yet researchers become increasingly aware of them, which is a very encouraging observation.

Concerning data description and storing, unsurprisingly, there are no common guidelines shared between disciplines and thus data exist in a plurality of formats (vector, video, audio, image, text, graph, raw bit streams, and so on), proprietary or/and open, depending on the software application. Those formats are tailored to the needs of the research projects (and team) and rarely in the optics of data preservation, a fact nicely summarized by the following answer of an interviewee to the question of what happens to research data after the end of a

project: *"That's the big problem. Nothing! The results are sleeping in the cupboards. We need to encourage the reflection about long-term preservation.*

Yet, common description standards exist in some internationally well-organized disciplines, such as in Geography, where large volume of satellite data are processed and stored by partnering up with CERN, Swisstopo, and various other international institutions. On the contrary, in Humanities and Educational Sciences, no standards are used, with sometimes even the question of what exactly represents a "datum", which surprisingly can in some cases remain difficult to answer.

As for data storing, in most of the cases, self-bought improvised servers are used, as institutional IT departments are often slower in providing solutions than the rate of data produced by researchers. Ad hoc infrastructures (institutional or externalized such as with Genbanks) are especially used in disciplines having to cope with huge amounts of data (like in Life Sciences). Independent of research discipline, researchers are aware of the need to back up their data, as loss of data is a recognized worrying issue. However, the organization of back-ups is not always seen as a task of the institution, but also of the individual.

The majority of the interviewed disciplines highlighted the issue of the surge of data volumes with all its underlying issues (storage cost, curation needs, etc.). This fast increase in data volumes is often cited as due to higher data quality (e.g., higher resolutions or higher sampling rates) and/or frequent versioning of those datasets (to keep trace of their processing steps).

2.1.2. Analysis and data exploration

According to the survey, both analysis and data exploration clearly depend on specific habits and software akin to the research disciplines. A very common answer is that data analysis "*is work of the PhD candidate*". How they organize themselves are their responsibilities, with sometimes some "good practices" communicated, such as "*don't lose your data, back-it up*". But there are no common guidelines.

Sharing of data also depends on the disciplines' habits, with a general tendency to agree to do it as long as it is within the same field and not against the researchers' own interests, such as bearing prejudice on future publications and academic reputation.

Dropbox is cited as the most used solution for sharing data, but an institutional infrastructure would be preferred if available. More difficult however seems sharing of raw or processed data for a variety of reasons:

- as soon as it concerns other disciplines than their own, or targets the "public domain";
- due to copyright issues;
- return on investment: too much efforts to do it without clear benefit, in return.

Interviewed researchers repeated the importance of incentives for data sharing. Such incentives might be, e.g. data citation or new ways of dataset peer-review, as it is increasingly appearing with the so-called "data journals"(16). A documented example on the resistance against data publication is FORS Lausanne(17). FORS aims at archiving all Social Sciences data coming from a plurality of disciplines by offering the needed infrastructure within their community. Nonetheless, so far data is not deposited deliberately; the main reason being,

according to our interviewed researchers, that it takes too much time to do it relatively to the benefits one can get from it.

2.1.3. Publication, archiving and long-term data management

The end of a research project, concomitantly to the acceptation of publications describing the main project results, often represents an important milestone for data, transforming the project status from active to passive. Given that the publication of datasets is not the rule (even if some journals ask for it, i.e. in Genetics discipline), the end of a project often means the definitive loss of its research data from an institutional point of view (a more or less properly named file might still reside on the researcher's computer). Also, the cited temporary solutions for preserving data, such as for example local storage or those based on Dropbox-like solutions, are obviously not a satisfactory answer. Yet, researchers often express the wish to keep data over longer periods, from 5 years (i.e. Computer Sciences) to forever (especially in the Humanities disciplines), with ten years representing a tangible time period for many disciplines.

In apparent contradiction with those cited periods of time, the notion of long-term preservation is generally absent in the answers. In one case an interviewee said: "*However, the real problem is : will we be able to read the PDF in 30 years? I am not even able to read my own documents from 20 years ago.*" Those cases, in which former research data cannot be used because of improper description, missing context and/or technical format changes, illustrate well the problematic behind long-term preservation.

In the cases where the ingestion of data into a repository for long-term preservation is mentioned as a good solution, then comes the open question whether the repository should be at national, institutional or disciplinary levels.

2.1.4. Research data in the future: challenges, risks, perspectives

Interviewed disciplines point to an awareness of upcoming changes in research, but underline that current habits will only change if the new ones make sense and offer an academic interest. They also suggest, as expressed by one interviewee, that *"funding agencies need to give money for storage as well".*

Furthermore, having additional expert staff for managing their data would be appreciated, as well as the possibility to have their research work more referenced through data citation.

One point is regularly mentioned: Researchers understand that there is no adequate answer to the question of what could and should happen to research data after the end of a project and/or after the successful publication of the scientific results. They acknowledge datasets disappear in the "office wilderness" or on unmaintained servers.

As long as researchers do not see any incentive for managing their data, they will not document neither annotate them, which render the preservation of datasets even more challenging. Yet, one interviewee acknowledges that *"Standardization is needed: often, same data is reproduced, because no one knew that the other one was doing it."*

Of course, there is the possibility of enforcing research data policies at institutional and national levels, but if enforced top-down, researchers might refrain deliberately from applying such policies.

2.2. Expressed needs

Based on the aforementioned results, we have reformulated them into needs, which we intend to address with high priority in the DLCM-project. Those needs are summarized as follows:

(a) Publication of guidelines and providing support to help researchers manage their data

The interviewed disciplines, aware of upcoming changes (as several citations illustrate), would like to have access to guidelines and templates to help them elaborate DMPs, as well as tools for analysis, publication and preservation of their data. Research teams would further appreciate expert staff in data management. One interviewee nicely expressed this fact: *"It would be like paradise if someone helps me organize my files, names them, etc."*

(b) Developing ad hoc short- and long-term storage, computing and data analysis solutions

These three interviewees' citations:

"It would be convenient to have some kind of collective data storage."

"Interesting would be an offer where data independent from the storage place could be managed 'all in one'."

"SNF does not fund long term storage. This is a big issue."

are only representative out of several, but they clearly mention the need for comprehensive solutions, addressing the whole cycle of research, but also the difficulty of finding appropriate financing solutions for these services. Raw data typically go through multiple processing steps before being interpreted. The input parameters, provenance, and data workflows must therefore all be documented if (parts of) the processing workflow has to be rerun on (selected) data. Under other conditions, researchers ask for the possibility to capture snapshots of the processed data along with contextual information. These snapshots have to be automated as much as possible so as to make their capture intuitive. In all cases, processing results must be stored in a database in an adequate format so that researchers can query them through easy-to-use interfaces, and further used them for not-as-yet devised purposes. The solution must ensure that the identification and extraction of data subsets respects the donor's consent conditions and that could potentially restrict the usage to specific field(s) of research. Furthermore, as data volumes augment, the issue of the cost of storage will have to be given special attention.

(c) Developing electronic laboratory notebooks

As the survey has shown, data analysis and its documentation in the laboratory is a major issue. Consequently, proper description is of utmost importance for research reproducibility and/or for affirming intellectual property on the research findings, yet is very difficult to put into practice. (For intellectual property reasons, "Dropbox-like tools", in which privacy in not guaranteed, would have to be proscribed.) In the future, there is an avowedly need for Electronic Laboratory Notebooks (ELN) that will allow the storing, linking and annotating of all digital data generated during the research process, and which will also serve as a preparatory step to further submit datasets onto long-term storage infrastructures. All research fields are concerned, including those in the Humanities.

N° DOI : 10.55790/journals/ressi.2015.e1602

(d) Development and/or maintenance of online research data repositories

The question of what happens to data after the end of a project or the final publication of the results has not been answered, but was raised as an important issue. Data are stored wherever it seems convenient, often in an unstructured manner and left on their own with the belief that they will be available for 5, 10 and even more years. To permit effective long-term preservation, there is thus a need for OAIS-compliant data repositories(18). Ideally, each dataset should get a permanent identifier, such as a DOI (Digital Object Identifier(19)), so that it can be referenced and linked to corresponding publications in a sustainable way, which should answer the interviewee's expressed need: "*Databases take too much time and are not user-friendly to handle. We need to start to link data with publications.*"

(e) Incentives for data management

Researchers will not do data management deliberately as long as they doubt the benefit in spending time and money to perform it. One interviewee nicely expressed this: *"But, the only way to bring us to describe our data is to make it mandatory. Furthermore, there must be a value, as well. We are collecting and analyzing data for publications, so, if the data is not useful for this, there will be no sense to publish it."*

Incentives have to be provided, and could be, for instance, the adoption of a reference mechanism to give more visibility to data, granting money for data management into the research project proposal, improve the impact factor of a publication, etc.

3. Organization of the DLCM Project

We based the project content on the above-mentioned needs. As for the choice of the considered scientific disciplines, that is Digital Humanities and Life Sciences, it came naturally from the expertise of the project partners; and by taking into account two "distant" disciplines, we encourage the setting up of generic solutions.

The project has been divided into six complementary tracks, which are summed up in Figure 1. With the exception of track 0, which concerns the project management, we now describe each of these tracks in more details.

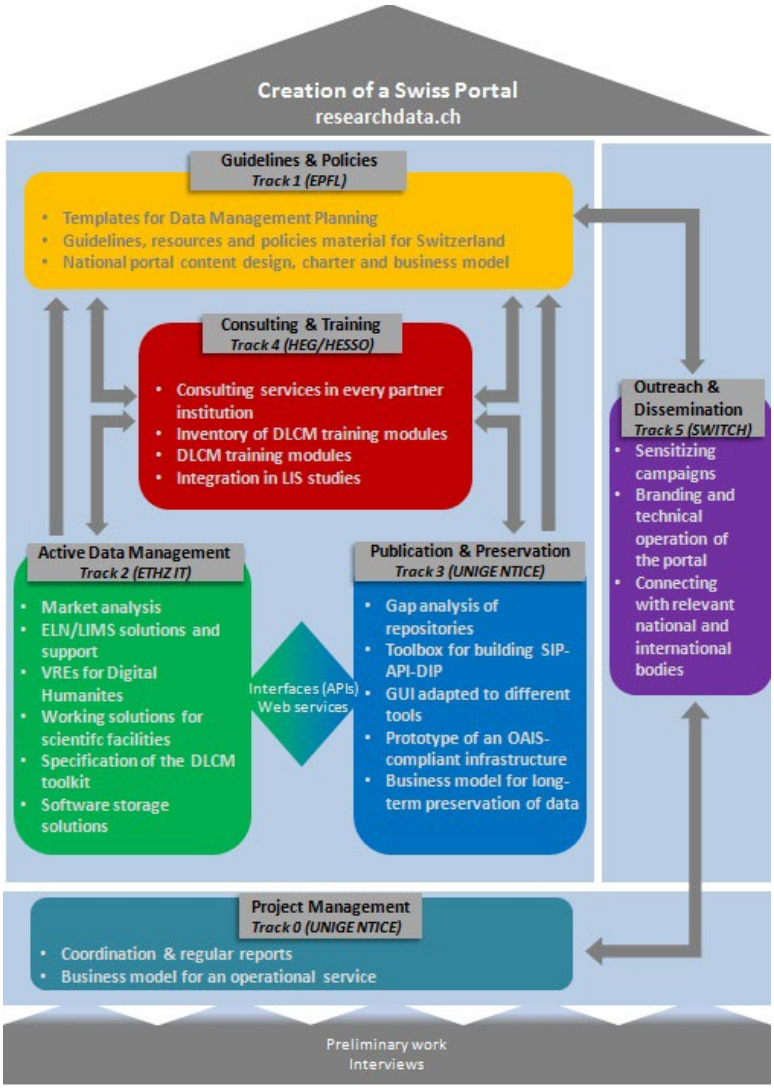

Figure 1: Project organization in 6 tracks

Track 1: Guidelines and Policies

Managed by the Swiss Federal Institutes of Technology in Lausanne (EPFL), this track shown in Figure 1 at the top of the building blocks was the first the DLCM team started to work on. Indeed, strongly connected to all other tracks, this track serves as an entry point to all future services through the setting up of a National Portal. Following the examples of national websites focusing on research data management available in the UK with the Digital Curation Center (see footnote 14), in France with the donneesdelarecherce.fr(20), or with DANS in the Netherlands(21), the proposed Swiss National Portal will provide access to customized resources, policies and guidelines regarding DLCM targeting scientific communities and information professionals based in Switzerland. The portal will also be of great importance for the whole project as it will make widely available the outputs of the different tracks of the project and a selection of a variety of external resources (e.g., training modules). At the time of this writing a first and basic version of this portal has been launched(22), which contains up-to-date information about the DLCM project and ongoing events concerning research data in Switzerland.

N° DOI : 10.55790/journals/ressi.2015.e1602

Besides the setting up of a National Portal, track 1 aims mainly at performing an exhaustive international literature review with which the DLCM project can build a set of best practices in terms of Data Management Plan (DMP) and policies. (The EPFL team has already worked on a similar project(23) and has therefore an up-to-date knowledge in this field.)

One of the first deliverables of this track is to provide at the Swiss level researchers and information professionals (librarians, IT engineers, etc.) with customized DMP templates regarding ongoing research projects. As previously mentioned, researchers are increasingly asked to provide DMPs when applying for research funding. As a direct consequence, the DLCM-project members are, in some cases, already facing researchers' questions concerning such documents. In addition to providing ready-to-use templates to the research community, relevant international links to existing resources (i.e., DMPonline(24)) will be proposed on the portal so as to get a wider offer on this topic.

Focusing on the political and strategic levels, this track will allow the Swiss institutions to set out the expectations for the management and sharing of research data. It should thus help define data policies for the provision of research data management services to the community and clarify the roles and responsibilities at the institutional level. For instance, the University of Edinburgh and the Humboldt University in Berlin have both implemented a research data management policy, respectively in 2011(25) and 2014(26). However, there is yet no such example in Switzerland. Consequently, based on pre-existing policies, intended not as prescriptive but as a basic starting point, track 1 aims at proposing such policies to the Swiss institutions which could then be adapted to specific disciplinary or institutional contexts. In this context, the DLCM project is partnering up with the European Project LEARN, which intends to offer policy templates at the international level (27).

Track 2: Active Data Management

The main objective of this track, managed by the Swiss Federal Institutes of Technology in Zurich (ETHZ), is to provide to a broad spectrum of researchers concrete technologic solutions and best practices to manage their active data, with particular focus on collecting, processing and analyzing those data. Four main axes will be followed in this track:

a) Electronic laboratory notebooks (ELN) and laboratory information management systems (LIMS) solutions and support

b) Virtual Research Environment (VRE) for Digital Humanities

c) Working solutions for scientific facilities

d) Software storage solutions for active data

Concerning ELN/LIMS (part a), nowadays in Life Sciences, the majority of data produced in a scientific laboratory is stored in an electronic format, but the description of the experimental procedures used is often still recorded in conventional (paper) laboratory notebooks. In Humanities, it is very similar: the whole research process is still on paper and collection, analysis as well as temporary data storage is generally dissociated.

Systems that combine ELN and LIMS, as well as VREs for Humanities (part b), all offer solutions to this practical problem by storing and describing the experimental data in a unique place and offering the possibility to easily find data according to relevant search criteria. These tools thus will allow scientists to keep track of their experiments and to select data for future publication. Furthermore, the target is to apply a market analysis to compare existing and already implemented ELN, LIMS (i.e., sLIMS@EPFL(28)) as well as VRE (i.e., SALSAH@UNIBAS(29)) solutions, so that the selected tools fit as much as possible the

researchers' requirements and ultimately can be adapted to the Swiss research community in other disciplines than life science and digital humanities.

The main objective of part (c) is to use partners' assets and know-how to propose and extend generic building blocks for selected use cases, particularly in medicine. Indeed, current research projects often require a contribution of multiple specialized partners and the generation of large primary datasets is often delegated to specialized platforms, so-called "core facilities". Data provided by these core facilities are of increasing size and complexity. Moreover, they are the result of complex processing workflows consisting of multiple steps involving the integration of various files and associated metadata. Most facilities do not provide management means for the data they distribute. This delegation of data management to the researcher is possible for very simple data sets but fails to scale to modern research applications. Furthermore, facilities have neither the mandate nor the funding to manage their customers data. Also, the ETHZ, with its solid experience both with core facilities and in managing of data for distributed research consortia, has developed building blocks that address parts of the problem and has gained experience in using them. Similarly, at the University of Lausanne (UNIL), Vital-IT and IT RC(30) teams have developed know-how covering specific needs: Vital-IT has successfully managed large distributed genomic datasets for partnerships involving both academic and private companies (e.g., Imidia(31)), while IT RC implemented an Information Technology platform that will be used for the management of all UNIL cohorts(32). All these cited experiences form a strong asset for progressing on these demanding aspects of the DLCM project.

Finally, part (d) aims at developing a DLCM system that can robustly handle distributed datasets, while keeping track of data origin. As a starting point, the ETHZ openBIS system(33) will be used. This system typically associates a dataset with a single storage resource. In a further step, concepts such as distributed datasets managed by various DLCM systems, active cloud storage for enabling cloud-based analysis, actively managed caching copies and ingestions of results from cloud storage back into DLCM systems will be included into the system solution.

Track 3: Preservation and Publication

The University of Geneva (UNIGE) is heading this track, which aims at establishing a bridge between active data and long-term preservation and publication solutions. For doing so, we will base ourselves on well-established concepts, such as the Curation Life Cycle and the OAIS Model. The track is organized in five main parts:

a) Gap-analysis of repositories

b) Toolbox for building OAIS information packages (SIP-AIP-DIP)

c) GUI adapted to different tools

d) Prototype of an OAIS-compliant infrastructure

e) Business Model for Long-Term Preservation of data

The gap analysis (part a) will focus on each partner institution's repository in use. Currently, they are at different stage of maturity and based on a large panel of technologies. For instance, the University of Zurich (UNIZH) is using ZORA(34), which is an example of a repository operated with the software Eprints(35). Eprints is a widespread open source repository software also used by the Universities of Bern and Basel. ETHZ is using Rosetta from Ex Libris(36), UNIGE, UNIL and SWITCH are using Fedora Commons(37), while EPFL's repository is based on Invenio(38) with the intention of linking it to Zenodo(39). Consequently, it

N° DOI : 10.55790/journals/ressi.2015.e1602

seems a logic step to assess the current status of these various systems and evaluate what is missing to comply with the Open Archival Information System (OAIS) standard, a strong requirement for guaranteeing Long-Term Preservation (LTP) of digital information.

For the methodology behind the gap analysis, we intend to exploit the evaluation tools developed by the Digital POWRR team (Preserving digital Objects With Restricted Resources)(40). They base their approach first on the National Digital Stewardship Alliance's, which defines five functional areas (storage and geographic location, file fixity and data integrity, information security, metadata, and file formats) and four levels of digital preservation (protect, know, monitor, and repair of data) akin to digital preservation systems. Second, they developed an evaluative grid resulting from the intersection of the Digital Curation Centre's digital curation life cycle and the OAIS Reference Model. They applied their approach on a selection of tools and services, inventoried on the Community Owned digital Preservation Tool registry(41). For the DLCM project, this methodology will be applied to the set of tools and services already used at the different partner institutions.

Based on the gap analysis, a toolbox will be proposed (part b) to allow researchers to deposit subsets of their data into a repository and/or a sustained long-term storage system. Typically, the OAIS Reference Model requires the preparation of a Submission Information Package (SIP); this process should be as transparent as possible to researchers. Tools exist to provide such microservices, as for instance Archivematica, Curator's Workbench, etc., and will further have to be interfaced to the various outputs of track 2. In the ingest process, the SIP will then be verified, before creating an Archival Information Package (AIP) from it, to be transferred into the archival storage for long-term preservation. This approach follows the logic of the DLCM, where at some stage those researchers who are working with an active storage for processing and analyzing their scientific data might need to select a subset of these data so as to ingest them into a longer-term storage system and/or for accompanying a publication (see above the three ages theory). Motivations for researchers to accomplish this step (passing from the active to semi-active or passive status) are various and mainly include: publishers asking for sustainable access to the data used to get the results in the published paper, need of a Digital Object Identifier (DOI) (or any other permanent identifier) for openly sharing the dataset, or simply archiving data at the end of a research project. To make this step as flexible and transparent as possible, the researchers using openBIS, SLims (LIMS@EPFL) or other ELN/LIMS tools should have the possibility to push their selected data from the active storage area into a data repository (semi-active status) and/or an LTP-system (passive status) "by a click". The aim for the output from the Active Data Management tool is to be a generic web-based service flexible enough to be adapted to different systems. Specific user interfaces (part c) then will allow researchers to deposit and retrieve information from the repository and/or from the archived AIP. In this latter case, this will be accomplished by generating a Dissemination Information Package (DIP) delivered to the user who has requested the information.

Another important constraint of the OAIS Reference Model concerns physical storage. Storage must be highly redundant, self-correcting, resilient, and must consist of multi-copies geographically distributed, while maintaining integrity and traceability of the stored information. This will involve robust "low-level" layers for data management. There is thus a clear need of developing a novel concept for a nationally distributed storage infrastructure, OAIS compliant (part d). While each partner institution will have to decide about what architecture they will implement on the basis of the gap analysis, UNIGE (responsible for this track) will propose an implementation relying on Fedora Commons 4, coupled to a community-own network such as LOCKSS for guaranteeing (dark-archived) copies in different geographic locations, with fixity checks and mechanisms for repairing corrupted data.

Finally, also key to this work package is how to make the LTP-systems economically sustainable (part e). It is indeed intended to propose LTP-services against a fee, which must be as much as possible aligned with what users can find on the market. Typically, today a hard disk installed on a personal computer, not secured, but which can easily be configured in mirror (RAID 1) costs about 100 CHF (and thus 200 CHF in RAID 1). The proposed services will be difficult to offer at such a price level for the corresponding storage capacity and consequently will have to offer attractive value propositions to researchers; otherwise there will be few incentives for researchers to place their data onto a central storage.

Track 4: Consulting and Training

Track 4, under the management of the Haute Ecole de Gestion (HEG) and the direction of the HES-SO, is addressing on one hand the training services, found to be missing in the other tracks, and on the other hand the creation of consulting services for data management where necessary. This track is due to start once the three previous ones have produced enough know-how to build pertinent training modules.

By delivering an up-to-date list of already existing training modules and highlighting them on the National Portal, this part aims at ensuring adequate DLCM knowledge transfer through these modules. If there are missing modules, they will have to be created.

In order to facilitate the creation, centralization and exchange of internal DLCM know-how and in order to answer as best as possible every type of researchers' questions, it is also intended to create DLCM consulting services at each partner institution, which will be coordinated by a central desk where necessary.

As a last important outcome, all acquired knowledge will be adapted to Bachelor and Master Courses in Information Science, where the freshly trained Library and Information Science students will practice and therefore ensure the sustainability of this knowledge for the future generations.

Track 5: Communication and Dissemination

Having closed the research data life cycle, there is still one important point missing. As all above mentioned topics are essential to the whole research community and should not be limited to the "biggest" players in the field, and in order to achieve the maximum possible impact of the results, a communication and dissemination track (track 5), headed by SWITCH, aims to reach institutions and other projects that are not directly involved or linked to the project.

For this, several sensitizing campaigns are planned during the whole project. Furthermore, establishing connections and links with relevant national and international bodies (i.e., RDA(42), DANS(43) etc.) will be pursued as well. One of these stakeholders is the Swiss National Science Foundation (SNSF(44)), the primary funding agency in Switzerland. Especially when it comes to DMPs, the direct involvement of the SNSF is a precondition for promoting their use.

As already mentioned, the Swiss National Portal for DLCM should become a reference website not only for researchers, but also for librarians, administrative personnel and IT staff. In order to keep the portal content as open as possible while ensuring copyright protection of its contributors, the appropriate CC (Creative Commons) license will be applied whenever feasible.

4. Conclusions

The DLCM project intends to respond to the primary researchers' needs related to data management. It can be seen as a direct consequence of recent international initiatives concerned with scientific data management, such as Horizon2020 or the LERU Roadmap for Research Data. Researchers' needs are various, and clearly discipline-dependent and so far no common ground between the Swiss institutions exists to provide satisfactory solutions and services. Also, after having conducted a consequent number of semi-structured interviews with researchers in different disciplines (covering about 30 different fields), five main needs could be identified, namely, (1) guidelines, (2) ad-hoc and long-term storage solutions, (3) electronic laboratory notebooks, (4) repositories for active data, and (5) development of incentives for data management. These needs set the basis for the project structure – five tracks that cover the whole data life cycle of a research project, and include training modules, and a National Portal to disseminate in a structured way as much as possible best practices in this somehow intricate field.

At the term of the project we reckon that researchers will have the appropriate tools to manage their data in a way reminiscent of digital curation methods. To be as transparent as possible, these tools will have to rely on microservices packaged into easy-to-use human-computer interaction interfaces as much as possible agnostic to the considered scientific disciplines. In any case, the challenge of closing the gap between the ways research data are managed by scientists and those in vigor in more administrative contexts managed by archivists will certainly guide us in the choices of the proposed solutions.

For any further information, please contact: pierre-yves.burgi@unige.ch or eliane.blumer@unige.ch

NOTES

(1) Examples of DMPs can be found at www.dmptool.com or dmponline.dcc.ac.uk/

(2) Economic and Social Research Council (ESRC) « Research Data Policy », March 2015: http://www.esrc.ac.uk/funding/guidance-for-grant-holders/research-data-policy/

(3) National Institutes of Health Data Sharing Policy (April 17th 2007): http://grants.nih.gov/grants/policy/data_sharing/

(4) OECD Principles and Guidelines for Access to Research Data from Public Funding, April 2007: http://www.oecd.org/sti/scitech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm

(5) Pampel, Heinz. « DFG erwartet Aussagen zum Umgang mit Forschungsdaten ». wisspub.net, May 25th 2010: http://wisspub.net/2010/05/25/680/

(6) National Science Foudnation Data Management Plan Requirements, January 18th 2011 : http://www.nsf.gov/bfa/dias/policy/dmp.jsp

(7) The Interuniversity Consortium for Political and Social Research Guidelines for the Office of Science and Technology Policy Data Access Plan, February 2013: http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/ostp.html

(8) European Commission, « What is Horizon 2020? », 2014: http://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020

(9) LERU Research Data Management Group. LERU Roadmap, 2013 : http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

(10) US Department of Energy, Office of Science. Statement on Digital Data Management, July 28th 2014: http://science.energy.gov/funding-opportunities/digital-data-management/

(11) EPFL Library. « Open research data: the future of science », October 28th 2014: http://library2.epfl.ch/conf/opendata

(12) Based on UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

(13) UK Data Archive « Research Data Life-Cycle », 2015: http://www.data-archive.ac.uk/create-manage/life-cycle

(14) Digital Curation Centre, « Digital Curation Life-Cycle Model», 2015 : http://www.dcc.ac.uk/

(15) All citations have been translated into English by the authors of this manuscript and are kept anonymous.

(16) One of such a data journal is for example: http://www.nature.com/sdata/about

(17) http://forscenter.ch/fr/

(18) http://www.iso.org/iso/catalogue_detail.htm?csnumber=57284

(19) http://www.doi.org/

(20) http://www.donneesdelarecherche.fr/

(21) http://dans.knaw.nl/nl

(22) http://www.dlcm.ch

(23) http://research-office.epfl.ch/funding/international/horizon-2020/open-research-data-pilot

(24) https://dmponline.dcc.ac.uk/

(25) University of Edinburgh, « Research Data Management Policy », May 16th 2011: http://www.ed.ac.uk/information-services/about/policies-and-regulations/research-data-policy

(26) Simukovic, Elena, « Forschungsdaten-Policy, Forschungsdatenmanagement, HU Berlin ». September 17th 2014 : https://www.cms.hu-berlin.de/de/ueberblick/projekte/dataman/policy

(27) http://cordis.europa.eu/project/rcn/194936_de.html

(28) http://www.genohm.com/slims/

(30) http://www.salsah.org/

(30) http://www.vital-it.ch/

(31) http://www.imi.europa.eu/content/imidia

(32)For example, Biobanque Lausanne (BIL), Swiss Kidney Project on Genes in Hypertension (SKIPOGH) http://www.skipogh.ch/index.php/Welcome_to_the_SKIPOGH_study! and the Cohorte Lausannois Study (http://www.colaus.ch).

(33) http://www.cisd.ethz.ch/software/openBIS

(34) https://www.zora.uzh.ch/

(35) http://www.eprints.org/uk/

(36) http://www.exlibrisgroup.com/de/category/Rosetta; see also http://dx.doi.org/10.3929/ethz-a-007362259

(37) http://fedorarepository.org/

(38) invenio-software.org

(39) zenodo.org, launched in 2013, was created by OpenAIRE (www.openaire.eu) and CERN to provide a place for researchers to deposit datasets

(40) http://powrr-wiki.lib.niu.edu

(41) http://coptr.digipres.org

(42) https://rd-alliance.org/

(43) http://www.dans.knaw.nl/en

(44) http://www.snf.ch/fr/Pages/default.aspx

BIBLIOGRAPHIE

[1] H. Ross (2010) “Digital Curation: A How-To-Do-It Manual”, Neal-Schuman Publishers, Inc.

[2] G. Kern, S. Holgado, M. Cottin (2015) “Cinquante nuances de cycle de vie. Quelles évolutions possibles ? Les Cahier du Numérique, 11(2) pp.37-76. doi:10.3166/lcn.11.2.37-76

[3] S. Büttner, H.-C. Hobohm, L. Müller (2011) “Handbuch Forschungsdatenmanagement”, Bock + Herchen Verlag, Bad Honnef 2011 (p.20)

[4] H. Pampel, R. Bertelmann, H.-C. Hobohm (2010) “Data Librarianship“, Rollen, Aufgaben, Kompetenzen. In: C. Schmiedeknecht & U. Hohoff (p.166)

[5] S. Corrall, S. (2008) “Research Data Management: Professional Education and Training Perspectives”, Vortrag. 2nd DCC / RIN Research Data Management Forum. Roles and Responsibilities for Data Curation. Manchester, UK, 26-27 Nov. 2008 (p. 6)

[6] M. Eckard, C. Rodriguez (2013) "Thinking Long-Term: The Research Data Life Cycle Beyond Data Collection, Analysis and Publishing", GVSU "Big Data" Conference 2013. Grand Valley State University, Apr. 2013