

Research Data Management in Switzerland: National Efforts to Guarantee the Sustainability of Research Outputs

Abstract

In this article, the authors report on an on-going Data Life-Cycle Management (DLCM) National project realized in Switzerland, with a major focus on long-term preservation. Based on a extensive document analysis as well as semi-structured interviews, the project aims at providing national services to respond to the most relevant researchers' DLCM needs, which includes: guidelines for establishing a data management plan, active data management solutions, long-term preservation storage options, training, and a single point of access and contact to get support. In addition to presenting the different working axes of the project, the authors describe a strategic management and lean startup template for developing new business models, which is key for building viable services.

Keywords: Research data management services, preservation of digital data, data life-cycle management, value proposition canvas, business model

Introduction

The ongoing technological developments in digital content give rise to new ways to collect, capture, store, manipulate and transmit large volumes of data and stimulate communication and collaboration between researchers. Research data, one important class of digital content, is defined by the Organisation for Economic Co-operation and Development (OECD, 2007) as “factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, and that are commonly accepted in the scientific community as necessary to validate research findings”.

The information contained in research data thus generally constitutes valuable assets to researchers. The US National Science Board (NSB, 2005) distinguishes three types of data: observational, computational, or experimental. This distinction is crucial in the choices made for archiving and preserving information: In the cases where observations will not repeat (for instance in astronomy, earth science, etc.) data are consequently unique and priceless. Conversely, for computational data, and if comprehensive information about the model is available (typically contained in the metadata), preservation in a long-term repository may not be necessary because the data can be reproduced. In between, data from experiments that can be accurately reproduced need not be stored indefinitely; yet in practice it may not be possible to reproduce precisely all of the experimental conditions, and/or the costs of reproducing the experiment are prohibitive. Apart from these kinds of data, other can be regrouped under the term of “record” (Borgman, 2015) covering general documentation (of government, business, etc.), to which we might add electronic laboratory notebooks and scanned artefacts in the research context.

Commonly, researchers don't pay attention in particular to the importance of the lifecycle of their data, a concept that encompasses many facets (as exemplified in CEOS, 2011), some of which, as illustrated previously, might depend on the relevant discipline. The data lifecycle begins with the acquisition of raw data. Subsequently these data are analysed to test hypotheses and validate models. The results of these studies usually lead to publications, an important milestone in the life of those processed data, as this can correspond to the end of the project and/or its funding. At this stage, data are often neglected, typically abandoned on the researcher's self-bought storage, improvised servers and/or institutional servers (in the best case), to be quickly forgotten by failing to promote them in other contexts. Disciplines with huge data amounts, i.e. Life Sciences, depend on their respective institutional or external proposed infrastructure. Also, data management Know-How is usually transferred among research teams and does not come from an outside service with data expertise. Even if backups at a minimum are advised and used, loss of data could be a serious issue independent of research discipline.

Preservation of digital information is nowadays rather well mastered (Schumacher et al., 2014), even if it remains challenging to keep bit streams unaltered on long periods of time (Rosenthal, 2010). Indeed, information is subject to the second law of thermodynamics, that is, the disorder (or entropy) tends to increase permanently, with the result that any data necessarily becomes corrupted with time. This implies that to preserve information this natural disorder must be prevented, for instance through the ability of matter to perform calculations to restore a corrupted information. This mechanism has limitations, however, by the required energy, and therefore a selection of information becomes a necessity. It is established that the exponential growth of production data, the multiplication of computer uses, the obsolescence of objects, etc. will increase demand for energy (Bihouix, 2014).

While preservation is key to save research data, it is by itself useless if the different stages of the data lifecycle are omitted. Researchers should indeed make explicit at an early stage of the project what they intend to do with their data during the project and afterwards, whether they plan to share them (or keep them confidential), how long they need to preserve them, etc. They should also document their datasets so as to be able in the future to understand them, for themselves but also for other researchers mainly within the same discipline, but not only (Goodman et al., 2014; Wallis et al., 2013; Wallis, 2014). Such Data Management Plans (DMP) are increasingly asked by funding agencies and research institutions.

One important part of the DMP concerns data formats and their eventual obsolescence. In principle before ingesting data into a repository, researchers should comply with recommended formats so that migration cycles can be ensured on the long term. Such a compliance is however not always possible for various reasons, which include performance and volumetry, and also simply researchers' willingness of doing it. In such cases, data description should be further developed so as to become self-described, a complex and time consuming process, which has also its limits (Lee, 2011). Researchers will rarely comply to do it, precisely because they prefer using their time for tasks they deem more pertinent for advancing their research work. To support preservation of any data regardless of format, Klindt and Amrhein (2015) define two levels of preservation: passive and active. At the passive level, data are preserved at the bit-level only, providing researchers with the possibility of storing data at minimal cost while complying with the editors' and/or funders' constraints. Obviously, this is not what information professionals are seeking, but this corresponds to a field demand. Conversely, at the active level all the necessary

preservation mechanisms are applied to ensure data remain interpretable throughout the migration cycles, yet this necessitates a more intensive preparatory work from the data producers. In both cases, preserved data should remain as accessible as possible for further uses, and not kept as dark archives. Access should consequently be facilitated, opened as much as possible following standard formats, with the aim of serving other researchers in the near and far future. Periodic value assessments must be performed so that the costs for preserving bitstreams match the intrinsic value of the preserved datasets, given that this value can lessen with time.

How many researchers have in mind all those management issues when acquiring their (increasingly) large volumes of data? Preservation of digital information is clearly a complex and costly process but cannot anymore be circumvented. Researchers need to be at the least rendered aware of these issues through training and assistance dispensed by information professionals and/or data curators to maximize digital curation of their data. By digital curation it is to be understood in simple terms and following the Digital Curation Center's definition (DCC, 2007) "the active management and appraisal of data over the lifecycle of scholarly and scientific interest" (see also Kim et al., 2013, for more extensive definitions). The Swiss project that we describe in this paper has such an ambition to cover all main issues contained in Data Life-Cycle Management (DLCM), taking for granted that preservation of digital information over the long term is representing the core element of this ambitious enterprise.

In the following sections we expose the main methodological considerations, the need analysis to DLCM guidelines and tools in an academic environment, and present the Swiss DLCM project along with its related dimensions and focus on long-term preservation.

Methodological Considerations

This section presents the general methodological approach, the main data collection and analysis techniques pertaining to the DLCM field and an overview of the major expected results by the end of the project. Given the exploratory nature of this work, the realization of these objectives is based on a qualitative approach. Regarding the gathering of data and their analysis, three steps were performed:

1. A large document analysis aiming mainly at performing an exhaustive academic and professional literature review with which the DLCM project can build a set of best national and international practices in terms of DMP and policies.
2. Semi-structured interviews to confirm the researcher's needs; more than 50 such interviews were performed with researchers from different departments in a variety of disciplines (see below), providing a deep knowledge of the diverse research data practices in the Swiss universities.
3. The analysis of the data, collected both from the literature review and semi-structured interviews, to lead to an overview of the main considerations that should be taken into account to offer a general framework/model for rational research DLCM with related guidelines, tools, competences and technologies to allow its effective operationalization in the academic environment.

Expected Results

The following outcomes are expected by the end of the project:

- A DMP adapted to the Swiss research communities and compliant with the main funders (e.g., Swiss National Science Foundation, Horizon 2020, etc.);
- Guidelines for Researchers and Information Professionals on research data management (including guidelines for data and metadata dissemination) to encourage the use of best practices in data management;
- A policy template for the higher education institutions (HEI), which can be used to establish sound policies to manage research data with all related issues (such as data privacy, intellectual property, storage costs, etc.);
- A National Portal on research data management with recommended tools and practices for researchers and information professionals;
- A toolbox for building SIP, AIP, and DIP OAIS packages from subsets of research data (including graphic user interfaces adapted to different tools);
- A prototype of a scalable OAIS-compliant infrastructure;
- Business models for the delivery of viable long-term preservation services;
- An inventory of existing data management training modules, including expert networks in collaboration with other Swiss and international projects;
- Specific data management training modules and teaching modules for integration into Library Information Systems (LIS) courses;
- Publications on data management in proceedings and journals.

Research Validity

To ensure the quality of our studies, the DLCM project team investigated scientific validity by

- Establishing the state of the art regarding the work performed in data management with local, national and international field experts;
- Being aware of the main theories, standards, projects in link with research data governance;
- Sharing and publishing intermediate results at peer-reviewed international conferences and journals.

To check practical relevance, the DLCM team met several professionals in public institutions and private companies with the intent to collecting their feedback experience in the field of long-term data preservation.

The interdisciplinary character of this work (data curation, information sciences, computer engineering, researchers' practices, etc.) brings a variety of competences that have an important impact on the quality of the outcomes. To ensure however a fluent collaboration between those several partners and researchers from different disciplines, a terminological glossary was defined and shared in regular meetings, workshops, etc.

Need analysis

As mentioned before, an exploratory need analysis was conducted in order to get a deeper knowledge about the researchers' needs and the solutions in place. For this, every partner institution 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 1 presents the compilation of all interviewed disciplines.

Institution	Number of interviews	Disciplines
University of Geneva (UNIGE)	8	Theology, Informatics, Linguistics, German, Cognitive Neuroscience, Educational Sciences, Geomatics, Archaeology, Vulnerability, Political Sciences, Medicine (Child Cancer Research)
ETH Zurich (ETHZ)	8	Biosystems Science and Engineering, Seismic Networks, Sociology, Consumer Behavior, Quantum Optics Group, Scientific Computing/Photon Science, Physics
University of Lausanne (UNIL)	15	Social Medicine, Social Sciences, Digital Humanities, Genomics, System Biology, Bio-informatics, Public Health, Imaging and Media Lab, Cancer Research
EPF Lausanne (EPFL)	5	Transport and Mobility, Quantum Optoelectronics, Supramolecular Nanomaterials and Interfaces, Audiovisual Communication Laboratory, Virology and Genetics.
University of Basel (UNIBAS)	7	Biology research (Biozentrum, 2), Biology (Core facilities, 2), Molecular Psychology, Public Health (STPH), Digital Humanities
University of Zurich (UNIZH)	5	Law Science, Biology/Microscopy, Biology/Proteomics, University Hospital, Geosciences
Total	49	

Table 1 Summary of the Need Analysis Interviews

As Table 1 shows, interviewed disciplines are various. In a second step, 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 (see Appendix). The main outcomes of those four interview parts are the followings:

Initial data and workflow

First and foremost, generally no formal DMP are being used, unless the funding instances require it at the time of the project application. As a consequence, data loss and description difficulties are often mentioned as a main issue.

Another challenge, concerning data description and storing, is that there are no common guidelines 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 view of data preservation.

Even if common description standards exist in some internationally well-organized disciplines, such as in Geography, 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. 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 the exclusive task of the institution, but also of the individual.

Analysis and data exploration

The biggest challenge in this part seems to be the freedom within data organization. Every research project, department, sometimes even researcher has her/his own habits and it is difficult to harmonize any of them. As for sharing and preserving, at the moment, Dropbox¹ remains a useful solution, even if it is subject to the American Law for privacy and copyright.

Publication, archiving and long-term data management

The notion of long-term preservation is generally absent and to the question of how long should data be kept their answers are either elusive or indicate that 10 years might be representative of such a long-term time period. Further questions without answers include what should be the best strategy for long-term preservation and whether there are any existing guidelines or rules.

Another difficulty is related to copyright, as there is no clear view on who the owner of specific data is or what has to be done in order to publish data in respect with the law in force.

Research data in the future: challenges, risks, perspectives

Even if the opinions of the importance for long-term preservation differ, one point is mentioned regularly: Researchers indicate 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 the successful publication of the scientific results. Often, data vanish in the wilderness of offices or on more or less well-cared servers. While the survival of these data is of concern, their proper management is cited as a more difficult matter.

¹ Dropbox (www.dropbox.com) is a file hosting service, operated by the American company Dropbox, Inc., which offers services like cloud storage and file synchronization.

Interviewed researchers repeat the importance of an incentive for data sharing. Such an incentive might be e.g. data citation or new ways of peer-review (as mentioned for instance in Tenopir et al., 2015). In order to do so, data sets have to be permanently identified by a DOI. Having a supplementary person who manages their data would be appreciated as well as the possibility to get more cited in the literature through data citation.

Finally, disciplines with huge amount of data see a risk as well in the rising costs of storage place with the related question of who will pay for it.

Main needs

An analysis of the main trends shows that data management clearly depends on the institutional strategies and/or research habits in the specific considered disciplines. Most of the interviewed disciplines are confronted with various challenges during different stages of their research projects. As a matter of fact, today's research cycle depends on ad-hoc solutions as well as specific habits within research departments. Also, based on the above-mentioned results, the following researchers' needs have been identified as development axes to be further proposed as deliverables:

- Guidelines and support for helping researchers properly manage their data,
- Ad hoc data storage, computing and analysis solutions,
- Solutions for active data management with storage of research progress based on periodical snapshots of defined datasets,
- Development and/or maintenance of online research data long-term repositories.

The DLCM Project

This section begins with an overview of the five major axes that compose the project and then focuses on the preservation topic and the viability of the proposed services.

The DLCM Project Organisation

The DLCM project started in September 2015 and is planned for lasting 3 years. It is organized into five tracks (or subprojects), headed by a specific partner institution, and which focus on : 1) Guidelines and policies; 2) Active research data; 3) Long-term preservation; 4) Consulting, training and teaching; 5) and Dissemination. The following sections briefly present the objectives of each of these axes.

Guidelines and policies

This part aims at defining guidelines based on an exhaustive academic and professional literature reviews with which the DLCM project can build a set of best national and international practices in terms of DMP and policies applicable in the HEI. Rather than dictating closed policies, a research data management policy template has been elaborated and presented at the steering board of one of the highest HEI political instances, i.e., the Swissuniversities research delegation (whose members are Rectors/Presidents from 9 HEI, and the directors of the Swiss National Science Foundation and of the Commission for Technology and Innovation). The aim of such a

top-down approach is to promote/recommend the development of policies based on a common national framework, while being flexible enough to allow adaptations for local specificities.

Active Research Data

Active research data refers to the stage of data collection, processing, and analysing early in the lifecycle. Based on the researchers' interviews, three main distinct scenarios could be identified: (1) "Single Endpoint", consisting either in raw data, or a number of data processing steps before archiving datasets; (2) "Open-Ended Work", which is representative of active data management that has no definite end and whose data can continuously evolve, and/or even refer to other data contained in the cloud (linked data) with obvious implication for properly archiving them; and 3) "Times Series Data", data continuously collected, possibly pre-processed and which have to be archived before being further processed. These three scenarios give a hint of how complex can be the articulation between active data management and preservation, as in most case it is not a linear process but rather one with a large number of cycles and subcycles. Thus this part aims at providing to a broad spectrum of researchers concrete technological solutions and best practices to properly manage their active data according to the three identified scenarios, with particular focus on collecting, processing and analysing those data. Along this line, three main domains are being considered: (1) Electronic Laboratory Notebooks (ELN) and Laboratory Information Management Systems (LIMS) solutions and support; (2). Virtual Research Environment (VRE) for Digital Humanities and (3) a range of working solutions for scientific facilities and software storage solutions for active data in variety of disciplines (with however a focus on life sciences).

Long-Term Preservation

This axis aims at establishing a bridge between active data and long-term preservation and publication solutions. For doing so, we consider well-established concepts, such as the Curation Lifecycle (Ball 2010; Pouchard 2015) and the OAIS Model (ISO 14721:2012). While this axis is described in more details below, the main parts composing it are: (1) the establishment of a gap-analysis of the repositories currently in use in the project partners' institutions; (2) the development of the required toolbox for building the OAIS information packages (SIP-AIP-DIP); (3) the design and tests of a prototype of a scalable OAIS-compliant infrastructure; (4) the elaboration of a business model with associated costs for preserving (large amounts of) data on the long term.

Consulting, Training and Teaching

The main targets for this axis are: (1) the establishment at national level of consulting, training and teaching within the DLCM field with both general and specific needs; (2) the creation of a consulting service at each partner institution, coordinated by a central desk; and (3) in parallel, the research data management knowledge will be integrated in Bachelor and Master courses in the field of Information Science, so that the freshly trained librarian can ensure the sustainability of this knowledge for the future generations.

Dissemination

This axis aims at promoting the results of the DLCM project and establish relevant contacts and collaborations with other institutions and projects not directly involved or linked to this one. For this, several sensitizing campaigns are planned during the whole project, including the organization of one-day annual workshops targeting researchers and information experts.

Participating to international conference and workshops, as well as publishing in proceedings and journals constitute other important activities of this track.

Preservation

Preservation of research data on the long term is one of the core objectives of the project by the fact that it serves at establishing a bridge between active data management and the mean to reliably store data (on the long term) so as to give access to the dataset that resulted into a publication, or simply allow to reuse it for furthering research.

Gap-analysis

The project partnership is made of seven institutions, some of them with solutions for long-term preservation, but mainly used for archiving publications. Consequently, we initiated the project with a gap analysis to assess the different stages of maturity and compliance of these solutions with the Open Archival Information System (OAIS) standard. Also, a large panel of technologies was encountered among the participating institutions: the University of Zurich (UNIZH) is operating a repository based on the Eprints software². The Swiss Institutes of Technology are using Rosetta from Ex Libris³ (in Zurich - ETHZ), and a repository based on Invenio, with the intention of linking it to Zenodo⁴ in the future (in Lausanne - EPFL), while the Universities of Geneva and Lausanne are both using Fedora Commons⁵ for their repository. To perform this assessment with most objectivity, we applied the evaluation tools conceived of by the Digital POWRR team⁶ (Preserving digital Objects with Restricted Resources – Schumacher et al., 2014). The methodology 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. This results into an evaluation grid (Figure 1), which represents the intersection of the Digital Curation Centre's digital curation lifecycle⁷ steps and the OAIS Reference Model (ISO 14721:2012) specifications. From this grid it is apparent that no institution completely fulfills the required compliance, though some (e.g. ETHZ) are close to the target. Moreover, most repositories do not comply with the “Reliable, Long-Term Bit Preservation” feature, which is a requisite for long-term preservation, yet a challenging one. Indeed, it is a known fact that disk storage based for instance only on RAID⁸ systems does not detect all kind of errors and is subject to “silent data corruption” (Rosenthal, 2010; Bairavasundaram et al., 2008), which thus requires other higher-level checksum mechanisms to guarantee errors are promptly detected.

² <http://www.eprints.org>

³ <http://www.exlibrisgroup.com/category/RosettaOverview>

⁴ <http://zenodo.org>

⁵ <http://fedorarepository.org>

⁶ <http://digitalpowrr.niu.edu>

⁷ <http://www.dcc.ac.uk/resources/curation-lifecycle-model>

⁸ Redundant Array of Independent Disks


	Copy	Fixity Check	Virus Scan	File Dedupe	Auto Unique ID	Auto Metadata Creation	Auto Metadata Harvest	Manual Metadata	Rights Management	Package Metadata	Auto SIP Creation	Public Interface	Auto DIP Creation	Auto AIP Creation	Reliable, Long-Term Bit Preservation	Redundancy	Geographically Dispersed Data Storage Model	Exit Strategy	Migration	Monitoring	Auto Recovery	Open Source	Clear Documentation	Cost
digitalpowrr.niu.edu																								
ETH Zürich	(X)	X	X		X	X	X	X	X	X	X	(X)	X	X	X	X	(X)	X	X	X			X	(X)
UNIGE Archive/Fedora 3	X				X	X	X	X	X	X	X	X	X	X		X		X				X	X	Free
UNIL Serval/Fedora3	X	(x)	X	X	X	X	X	X	X	X	X	X	X	X		X	(x)	X	X	(x)		X	X	Free
UZH/ZORA Eprints	(X)		X		X		X	X	X	X		X			(X)	X	X					X	X	Free
Zenodo (EPFL)	(X)	X	X	(X)	X			X	X	X	(X)	X	(X)	(X)	X	X		X		X	X	X	X	Free

Figure 1 Evaluation Grid. “(X)” indicates partial or similar function.

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

At some stage in their research work, researchers will want to select a specific dataset from the active data storage area for ingesting it into a longer-term storage repository. 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, researchers should have the possibility to push their selected data from the active storage area into a data repository (semi-active status) and/or a long-term preservation system (passive status) “by a click”. Following the OAIS Reference Model requires the preparation of a Submission Information Package (SIP). In the process, several micro services are usually activated and include checksum calculations, format detection, virus check, etc. with at the end the assembly of an Archival Information Package (AIP). Tools exist to provide such micro services and for automatizing AIPs (for instance Archivematica, Curator’s Workbench, iRods, etc.). From the consumer side, specific user interfaces must be designed to allow users to 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.

Scalable OAIS-compliant infrastructure

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. SAFE PLN is an ongoing project for building such an OAIS-compliant based on the LOCKSS infrastructure⁹ (Maniatis et al., 2005) and regroups several partners in Europe and Canada¹⁰.

⁹ <http://www.lockss.org>

¹⁰ <http://www.safepln.org>

LOCKSS is based on the Byzantine fault tolerance concept, originally known as the Byzantine Generals' Problem (Lamport et al., 1982). To be able to defend against Byzantine failures, in which components of a system fail with symptoms that prevent some components of the system from reaching agreement among themselves, redundancy is required to form a voting poll. For example, for tolerating two node failures, at least 7 nodes are necessary. The current implemented version offers a limited storing capacity (less than 1 TB) over 7 geographically distant nodes. Also, and even if in the future SAFE PLN could be extended to tens of terabytes (with many issues yet to be solved, such as bitstream transfers between nodes), to manage even much larger volumes, typically of the order of several petabytes, we have to consider more scalable architectures. One particularly promising solution is based on Archivematica for the ingest part, iRods for automatizing the archival storage (mainly for managing two replications, one being geographically distant), and Fedora Commons for exposing the DIP through a dedicated user interface. The Zuse Institute Berlin (ZIB) has developed such an infrastructure for managing more than five petabytes of data on disk (Klindt and Amrhein, 2015). Other solutions relying on object-based storage (such as Ceph¹¹) are currently being investigated. Among the still many open questions yet to solve to render the infrastructure scalable, one is technical: how to check and replicate petabytes of data possibly distributed over hundreds of millions of files? Another issue, dealt with in the following section, is related to the financing of such large-scale infrastructures on the long term..

Viability

Given the DLCM project is limited in time and resources (as all projects are), a viability methodology based on strategic management and lean startup templates for guaranteeing the sustainability of the services on the long term has to be considered. For this, the Business Model Canvas (Osterwalder and Pigneur, 2010) and Value Proposition Design methodology (Osterwalder et al., 2015) are being applied. The methodology provides tools to simulate how an institution can potentially make, supply and earn value, while the BMC provides the essential elements for developing the accompanying business models. One important component of the Value Proposition is to place the customers (i.e., researchers) at the centre, and develop services in relation to their real needs. This should in principle avoid developing services that will not be used. Yet, this does not solve the sustainability issues, which are related to the business model. Currently, the project members familiarize themselves with such a methodology (which per se is not an intuitive approach) and start to institutionalize it, with the help of a business analyst (who is also a member of the project team).

Value Proposition and Business Model Canvases

Within the Value Proposition Canvas (see Figure 2) one typically describes the considered customer' practices, including their every day tasks (job), potential gains and incumbent pains. In a further step, one attempts to find gain creators, pain relievers and transform them into products and services. This way, every created service answers an existing and preliminarily identified need and provides a viable value in reducing pains. One example of Value Proposition corresponding to three service developments for Long-Term Preservation (LTP) is illustrated in Figure 2.

¹¹ <http://ceph.com/ceph-storage/object-storage/>

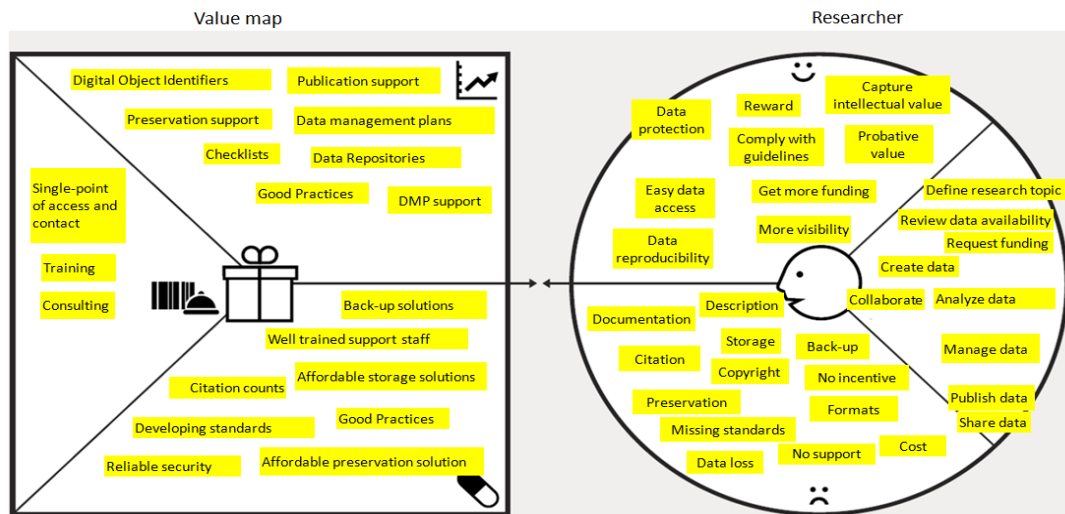


Figure 2: A Value Proposition Canvas corresponding to three LTP services (value map at left) with on the right (circle) the Gains, Pains, and Customer Jobs indicated.

In a further step, the BMC is completed by first specifying the “Customer Segment” part, by integrating the Value Proposition Canvas into the “Value Proposition” field, and then by filling the other sections (customer relationship, key activities, key resources, key partnerships, revenue streams, and cost structure) in order to describe and finalize the targeted services. An example of BMC corresponding to a National Portal service is illustrated in Figure 3. In this case several customer segments have been identified (specified in the outer right column of the canvas).

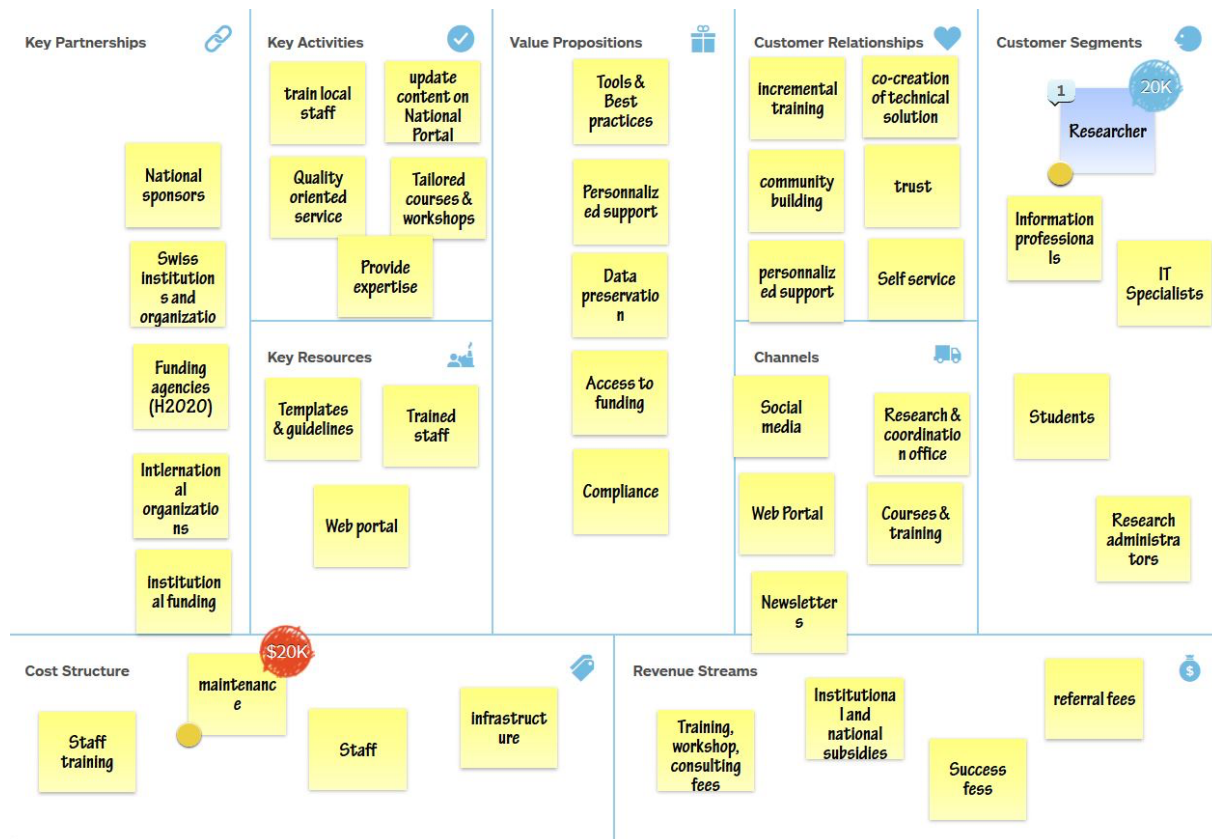


Figure 3: Business Model Canvas for a National Portal service. A population of 20'000 Researchers have been evaluated as potential customers (blue circle), and maintenance costs of the infrastructure was estimated to 20'000 US\$ (red circle).

Cost Models

One key component of the business model canvas is Cost Structure (see Figure 3), which should match the Revenue Streams. The issue of the LTP costs, has been grappled through a number of international studies, most notably the KRDS (Keeping Research Data Safe) project (Beagrie, 2008; Beagrie, 2010). This project focused on UK case studies to draw conclusions on the main cost drivers, and on this basis a costing framework containing two major elements, economic and service adjustments, was proposed. Economic adjustments consist of inflation (e.g., salary), deflation (e.g., storage media), depreciation of the assets (historical/purchase costs across its useful life) and cost of return for financing and investment. Service adjustments consist of costs related to (1) staff, (2) acquisition, disposal, and ingest phases, (3) archive storage, preservation planning, and data management. Interestingly, backup and long-term file storage represent only a tiny portion of the total costs. According to this study, acquisition and ingest count for up to 42%, access for 35%, and storage for 23%. Even more extreme is a typology of the activities that considers data creation, which counts for 73%, with curation and storage accounting respectively for 24% and 3%.

Apart from the KRDS study, another notable work in this domain is the comparative study of several cost models realized in the context of the 4C European project¹². However, among all

¹²See D3.1—Evaluation of Cost Models and Needs & Gaps Analysis, available on <http://4cproject.eu>

these works, one model in particular retained our attention, the Total Cost of Preservation TCP from the UC Curation Center (Abrams et al., 2012), which encompasses the full economic costs associated with long-term preservation of digital assets. This model takes into account 11 preservation activities (System, Services, Servers, Staff, Producers, Workflows, Content types, Storage, Monitoring, Interventions, Oversight), and considers two price models: “Pay-as-you-go” and “Paid-up”. If at first it seems considering such a comprehensive cost model is beneficial to determining financing load, in practice it brings more questions than it answers them. Indeed, to be applicable, first one has to know in advance the number of “Producers”. And as the number of Producers increases, the associated costs diminish through economy of scale, which is to say that a Producer has better to come late to benefit from cheaper prices. Second, evaluating all 11 preservation activities is far from being trivial and/or necessarily reliable. Consequently, the DLCM project team developed a simplified TCP cost model independent of the number of Producers, based on the “pay-up price model” without assuming any investment return, but which takes into account a global cost for maintaining the infrastructure, including staff’s stipends and an annual percentage rate of price decrease for the hardware. It is acknowledged that other important costs related to services such as creation and curation of datasets exist, but to evaluate basic costs associated to safe storage of information, it was assumed that other services could be billed separately in function of the level of services customers ask for. And for researchers, having reasonable costs for storing their data on the long term is clearly an important incentive for avoiding letting data end their life on inappropriate storage infrastructure (such as USB keys or private computer).

Main Realizations, Future Steps and Challenges

This Swiss DLCM project started in September 2014 with a prestudy, followed by a first implementation phase initiated in September 2015. The concrete outcomes after one year are the followings:

Guidelines, DMP, and Policies

A Web site with relevant resources, tools and guidelines for researchers has been set up (www.dclm.ch). A Data Management Policy Template for guiding institutions in establishing Research Data Management policies has been written and presented to the Swissuniversities Research Delegation, which is one of the highest Swiss assemblies for HEI. This top-down process is still ongoing with the intent of homogenizing the Swiss policies in this domain in coordination with the Swiss National Science Foundation, whose mission is to finance research at National level through competitive grants. Finally, a Data Management Plan Checklist has been elaborated based on previous experience of two partner institutions (EPFL and ETHZ), with the idea of collaborating with DMPOnline¹³ developed by the UK Digital Curation Center to produce an online tool adapted to the Swiss needs.

Active Research Data

This part of the project is particularly complex, as it must deal with the working environment of researchers from various disciplines. Also, one of the main axes concentrates on LIMS and ELN,

¹³ <https://dmponline.dcc.ac.uk/>

a topic of concern for researchers willing to document their research processes and data. A relevant market and gap analysis concerning Swiss LIMS software (SLims¹⁴, OpenBIS¹⁵, ViKM¹⁶), and other tools mainly used in life science (e.g., Labkey¹⁷), has been accomplished along with video tutorials to facilitate their use. In a further step how these tools can be applied to other fields than life science will be assessed. For Digital Humanity (DH), a virtual research environment (Salsah/Knora) is currently being evaluated on several DH projects¹⁸.

Long-term Preservation

During the first year of the project a gap analysis using the methodology (Schumacher et al., 2014) has been conducted in order to identify the relevant gaps in institutional repositories for becoming OAIS compliant (see chapter Preservation). The main outcome of this work is an inventory of the tools and repositories expected to constitute the future ecosystem of the LTP National service (see Figure 4). Next steps will include specifications for interoperability between institutional repositories and research tools, and of the SIP, DIP and AIP (in progress), which will lead to the development of APIs for interfacing the National service.

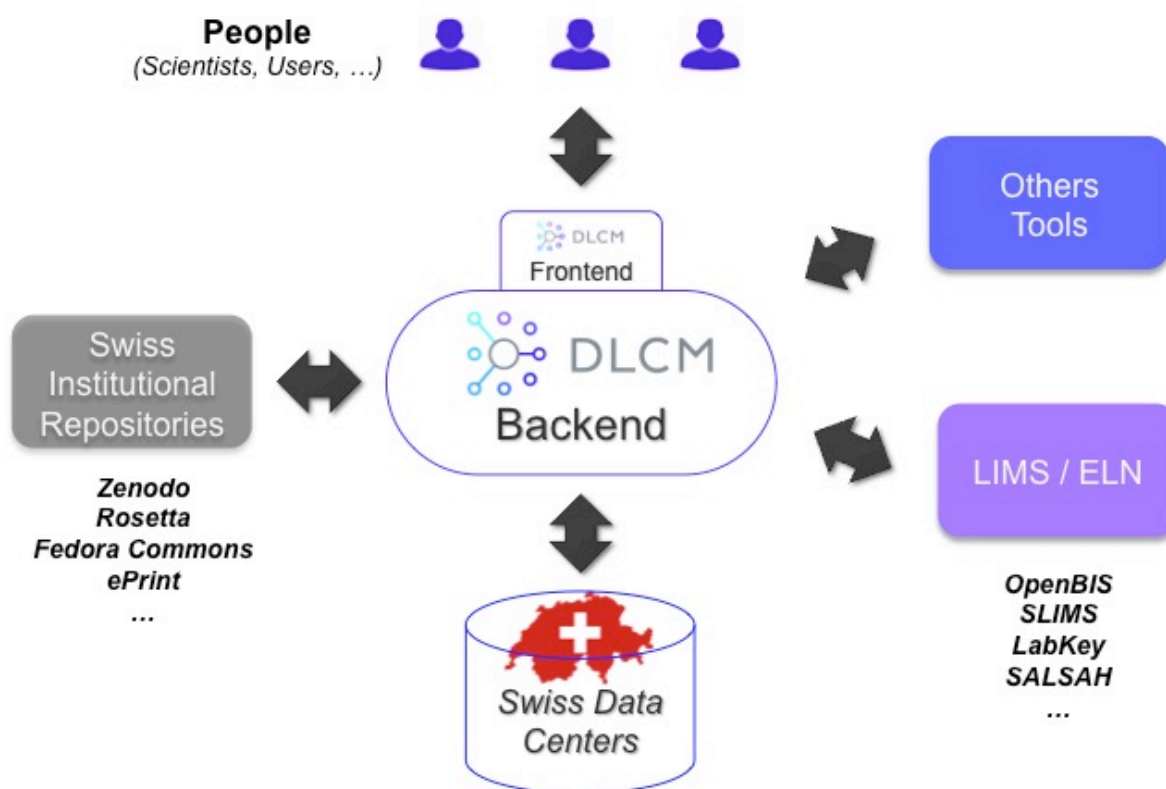


Figure 4: The DLCM LTP Ecosystem.

¹⁴ <http://www.genohm.com/slims>

¹⁵ <https://wiki-bsse.ethz.ch/display/bis/Home>

¹⁶ <https://www.vital-it.ch/research/software/ViKM>

¹⁷ <http://www.labkey.com>

¹⁸ For more information on Salsah/Knora and related DH projects, see <https://dhlab.unibas.ch>

Consulting, Training and Teaching

Computer engineers and information specialists in different institutions have already dispensed Research Data Management (RDM) workshops and trainings to librarians. In parallel, an extensive catalogue of RDM training modules is being created along with a need analysis for Bachelor training in Swiss Universities of Applied Sciences. Another important aspect being worked on is the development of a generic consulting service for Swiss HEI. One expected outcome of this initiative is a central and focal coordination desk integrated into a large network of trained RDM specialists representing their respective academic institutions and scientific communities.

Dissemination

At this early stage of the project, one of the main outreach milestones was to organize an annual National event on the RDM thematic targeting the Swiss researchers community. This type of event (the first having taken place in November 2016) is intended to gather international keynote speakers, and representatives from academic direction boards (e.g., Swiss National Science Foundation, Swissuniversities, etc.), and leave space for discussions through parallel workshop sessions whose topics typically encompass the main DLCM facets (DMP, active data management tools, infrastructures, policies, etc.).

Business Models

The DLCM project faces different challenges, such as the viability and sustainability of the proposed National services, collaboration between Swiss data centers and other RDM initiatives, and decision making of the HEI steering boards. The viability of the services is being worked on since the very beginning of the project in order to develop a business culture among the partners (which in an academic environment is not given), and to get enough time to analyze, validate the hypotheses and adapt iteratively the business models by interviewing the users (or “customers”) on their usages of the new services.

Conclusions

Targeting rational and optimized management of research outcomes is challenging, particularly when it means being able to use securely, fluently, promptly and durably research data during the whole research life cycle. Researchers face several demanding situations regarding data management, but tend to use ad hoc solutions. Based on face-to-face interviews with researchers from a variety of disciplines, the present work identified the major needs, which reduce to: guidelines and support in managing data, computing and analysis solutions, and solutions for storage of research progress and repositories. At the term of the project, Swiss-wide services are expected to respond to those needs and beyond them through the implementation of RDM best practices. Key to the viability of the future proposed services is the elaboration of sound business models, and for that the project members have been imparted knowledge on a methodology, usually confined to economists, based on two concepts, the Business Model Canvas and the Value Proposition Design (Osterwalder and Pigneur, 2010; Osterwalder et al., 2015). Further advantages of applying this methodology stems from the fact that the DLCM project members develop a common culture and use the same vocabulary beyond their institutional limits and collaborate in a creative and constructive way.

The experience gathered so far has shown that there are additional challenges in addition to providing suitable and sustainable services. Indeed, the national dimension of the project, and the attributes and needs of specific partner institutions require good coordination and proper governance in an academic context in which researchers are not necessarily accustomed to the application of binding rules. Consequently, finding the right incentives for having researchers comply with a minimum set of RDM policies will determine the success of this enterprise.

☐ The near future will show how the proposed services and the project results will be used. In any case, the increasing pressure from publishers and funding agencies necessitates the sharing of best practices and resources across all HEI to meet these challenges.

References

Abrams S, Cruse P, Kunze J, Mundrane M (2012) Total Cost of Preservation (TCP): Cost Modeling for Sustainable Services. UC Curation Center, California Digital Library

Bairavasundaram LN, Goodson GR, Schroeder B, Arpaci-Dusseau AC, Arpaci-Dusseau RH (2008) An analysis of data corruption in the storage stack. FAST'08: 6th USENIX Conf. on File and Storage Technologies, 223–238

Ball A (2010). Review of the State of the Art of the Digital Curation of Research Data (version 1.2). ERIM Project Document erim1rep091103ab12. Bath, UK: University of Bath

Beagrie N, Chruszcz J, Lavoie B (2008) “Keeping research safe. A cost model and guidance for UK universities” Final Report, Charles Beagrie Limited

Beagrie N, Lavoie B, Woollard M (2010) “Keeping research safe 2” Final Report, Charles Beagrie Limited

Bihouix P (2014) L'Âge des low-tech. Paris, Seuil

Borgman CL (2015) Big data, little data, no data. MIT Press, Cambridge MA

CEOS (2011) Data Life Cycle Models and Concepts. Working Group on Information Systems and Services Data Stewardship Interest Group. CEOS.WGISS.DSIG.TN01, 26 September 2011

Digital Curation Centre (2007-04-26). What is Digital Curation? <http://www.dcc.ac.uk/digital-curation/what-digital-curation> (accessed 2016-09-15)

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. PLoS Comput Biol 10(4) doi:10.1371/journal.pcbi.1003542

Kim J, Warga E, Moen WE (2013) Competencies Required for Digital Curation: An Analysis of Job Advertisements. The International Journal of Digital Curation 8(1) 66-83

Klindt M, Amrhein K (2015) One core preservation system for all your data. No exception! Proc. of the 12th International Conf. on Preservation of Digital Objects, 101-108

Lamport L, Shostak R, Pease M (1982) "The Byzantine generals problem" ACM Trans. on Programming Language and Systems 4(3) 382-401

Lee CA (2011) A framework for contextual information in digital collections. Journal of Documentation, 67(1) 95-143

Maniatis P, Roussopoulos M, Giuli TJ, Rosenthal DSH, Baker M (2005) The LOCKSS Peer-to-Peer Digital Preservation System. ACM Transactions on Computer Systems, 23(1) 2-50

NSB (2005) Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century. NSB-05-40, National Science Foundation
<http://www.nsf.gov/pubs/2005/nsb0540/nsb0540.pdf>

OECD (2007) OECD Principles and Guidelines for Access to Research Data from Public Funding <http://www.oecd.org/science/sci-tech/38500813.pdf>

Osterwalder A, Pigneur Y, (2010) Business Model Generation: A Handbook for Visionaries, Game Changers, and Challengers. John Wiley and Sons

Osterwalder A, Pigneur Y, Bernarda G, Smith A, Papadakos T (2014) Value Proposition Design: How to Create Products and Services Customers Want. Wiley

Pouchard L (2015) Revisiting the Data Lifecycle with Big Data Curation. International Journal of Digital Curation 10(2) 176-192

Rosenthal DSH (2010) Bit Preservation: A Solved Problem? The International Journal of Digital Curation 5(1) 134-148

Schumacher J et al (2014) From Theory to Action: "Good Enough" Digital Preservation Solutions for Under-Resourced Cultural Heritage Institutions. A Digital POWRR White Paper for the Institute of Museum and Library Services

Tenopir C, Dalton ED, Allard S, Frame M, Pjesivac I, Birch B, Pollock D, Dorsett K (2015) Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide. 10(8):e0134826. doi:10.1371/journal.pone.0134826

Wallis JC, Rolando E, Borgman CL (2013) If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. PLoS ONE 8(7). doi:10.1371/journal.pone.0067332

Wallis JC (2014) Data Producers Courting Data Reusers: Two Cases from Modeling Communities. International Journal of Digital Curation, 9(1) 98-109

Appendix

Table 1 : Interview results: DLCM obstacles and challenges

	Initial data and workflow	Analysis and data exploration	Publication, archiving and long-term management	Research data in the future: challenges, risks, perspectives
Theology	No DMP Back-up	Freedom of methodology to organize data	Freedom of organization	No local storage Copyright issues
Informatics	No DMP	Lack of description	Lack of description	No storage for sharing and versioning Incentive
Linguistics	No DMP Description rules Back-up Lack of documentation	Freedom of methodology to organize data	Conservation rules Data is obsolete	Format issues
German	No DMP Description rules Back-up	Freedom of methodology to organize data	Conservation rules	Citation rules No archive Incentive
Cognitive Neuroscience	DMP for US-projects Data loss Local back-up and storage	-	Conservation rules No reproducibility	Versioning Standardization
Educational Sciences	No DMP Local storage and back-up	Freedom of methodology to organize data	Copyright issues	Institutional storage on local level with access rules Standardization
Geomatics	No DMP	-	-	Need of a data archive
Archaeology	Description rules Data storage No DMP	Freedom of methodology to organize data	Conservation rules Copyright issues	Interdisciplinarity

Vulnerability	No DMP No description rules Local storage	-	Data loss Data publication	-
Political Sciences and International Relations	Naming Preserving	-	-	data repository Twice production of same data Incentive
Medicine (Child Cancer Research)	-	-	-	-
Quantum Optoelectronics	Local storage	-	-	Link between data and publication Cost Guidelines Storage Incentive
Transport and Mobility Laboratory	Local backup	Dropbox	Conservation rules	Checklists Institutional repository Incentive Lack of documentation Copyright issues
Supramolecular Nanomaterials and Interfaces Laboratory	No DMP Back-up	-	-	Standardization Metadata
Genetics	Standardization in description	Dropbox	-	Guidelines Reproducibility
Audiovisual communication laboratory	-	-	-	Cost National perspective Metadata Incentive
UZH	Missing standards	-	-	Reproducibility Best Practices

				Incentive
Unispital ZH	Missing standards	Ethics	No long-term preservation strategy	Cost Centralized storage with access management Incentive
Proteomics	-	Centralized sharing tool		DMP Storage cost Long-term preservation
Photon Science/Scientific Computing	-	-	-	Cost Incentive Politics
Biosystems Science and Engineering	-	-	Long-term preservation Conservation rules	Easily accessible web front end for data publication Infrastructure
Seismic Networks	-	-	Sharing (others can see, but not process)	Automatic snapshots Storage facilities
Social Sciences, Modelling, Simulation	No DMP	-	-	Indexing large volumes of data including hidden annotations Framework for management support Capture defined states of data periodically Differential record of algorithms' versions Comprehensive Swiss cloud solution Real time streaming of big data Reliable citation counts

				Real time data input Support and staff
Consumer Behaviour	No DMP Data loss Reproducibility	-	-	Managing data all in one Local storage Incentive
Quantum Optics Group	-	-	-	Periodical snapshots of specific data Persistent link to published data Time
Physics, IT Services Group	No DMP Local Storage	-	-	Formats Incentive
Biocenter	No DMP	-	-	Repository
Epidemiology	-	-	-	Interoperability Long term storage is not funded
Digital Humanities	Data loss	-	Copyright	-
UNIL				