Data Life Cycle Management Pilot Projects and Implications for Research Data Management at Universities of Applied Sciences

*Fürholz Andreas
Research and Development Unit, President's Office
Zurich University of Applied Sciences (ZHAW)
Winterthur, Switzerland
fueh@zhaw.ch

*Jaekel Martin
Research and Development Unit, President's Office
Zurich University of Applied Sciences (ZHAW)
Winterthur, Switzerland
jaek@zhaw.ch

*On behalf of the ZHAW pilot project teams. Contributions to this paper by: Durrer, J.; Pothier, J.; Sommer, B.; Johner-Kobi, S.; Koch, P.; Robin, D.; Krasselt, J.; Schwarz, B.; Bernath, J.; Götzö, M.; Holzer, L.; Lobsiger-Kägi, E.; Kaiser, C.; Šimukovič, E.; Hauf, N.; Klaas, V.; Morger, J.; Schroeder, C.; Hausmann, I.

***Abstract*— *Publicly available research data (Open Research Data) are a main pillar of Open Science and can be considered as a good measure to increase the effectiveness, transparency and reproducibility of scientific research. However, the rather new scientific practice of Open Research Data sets new demands on best practices in research data management and raises questions regarding the data publication itself, for example finding a suitable data repository or the consideration of legal aspects. To investigate these practical questions, 12 pilot projects were carried out within the DLCM 2.0 project. Research data were published in a variety of disciplines and related processes where reflected in workshops within the project consortium. The pilot projects have provided an insight into the characteristics of individual research data life cycles. A key finding is that the path to open research data is very domain specific. Based on this experience, we think that the individual research communities – as predominant re-users of research data – must develop discipline-specific standards, best practices and data processing workflows. We believe that this is the most important success criterion for data exchange and should be promoted in parallel with meeting the FAIR data principles and an appropriate data curation. To promote this development, support measures are needed at various levels. On the one hand, there is a need for cross-border initiatives to support the communities developing their standards and best practices. On the other hand, researchers must have the appropriate infrastructure, training and support on local level. We consider the latter to be particularly important. That is why we have set up a data stewardship model at our university, where researchers can receive active support over the entire research data lifecycle.*

Keywords—Open science, open research data, research data management, data stewardship, data stewards, electronic laboratory notebook

I. Introduction

As part of the Open Science movement, research results are published increasingly and more comprehensively. In addition to publicly accessible publications (Open Access), the underlying research data are being published more often (Open Research Data, ORD). This development is driven in particular by funding agencies such as the Swiss National Science Foundation (SNSF) or the European Commission, which want to increase the effectiveness, transparency and reproducibility of scientific research (EU, 2017; SNSF, 2021). Both funding agencies require the writing of Data Management Plans (DMP), which aim to clearly define the handling and the publication of research data. Another driver of ORD is the movement towards new evaluation systems for research outputs such as the Declaration on Research Assessment (DORA[1]).

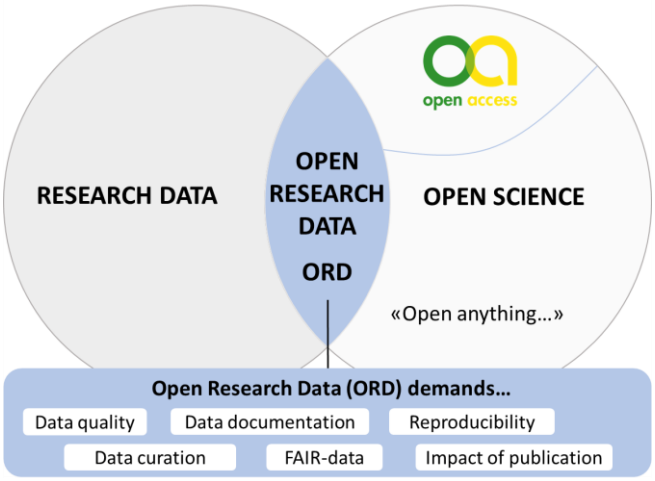

Fig. 1. Open Research Data (ORD) increases demand on best practices in Research Data Management (RDM).

From the perspective of universities, the publication of research data brings opportunities and challenges. Researchers and institutions can raise awareness of their research outputs and thus start more likely new collaborations or find new project funding partners. However, the publication of research data may set additional demands. In principle, ORD should meet the well-known FAIR data principles (Wilkinson et al., 2016). But the implementation of these principles currently often leads to additional work, e.g. in data preparation or data documentation. In addition, ORD are often only a part of the entire research/project data, which means that data curation is necessary prior to publication (Fig. 1).

One approach to mastering the above complexity is to actively manage research data over the entire life cycle (Fig. 2) and to consider discipline-specific best practices. In practice, however, some questions arise:

- How can efficient, comprehensible and reproducible data workflows be established?
- How can research data be published with impact?
- What support can institutions provide to their researchers?

[1] https://sfdora.org

II. Overview and Goals of Pilot Projects

To answer above mentioned questions, 12 pilot projects were carried out as part of the DLCM 2.0 project. The results and questions raised in practice have been reflected in workshops within the project consortium. We distinguished between two types of pilots.

A. *Open research data pilot projects ("ORD-Pilots")*

Research data in various disciplines were processed, published and archived within 10 pilot projects ("ORD-Pilots"). All except of one[2] of the research projects had been completed and related paper publications had already been done.

The first goal of the "ORD-Pilots" was to identify and evaluate suitable discipline-specific data repositories. This task was preceded by the assumption that discipline-specific data repositories allow a better reuse of research data. After the identification of suitable repositories, research data were post-processed and published. At the end of the pilot projects, the impact of the data publication was analysed. Table I gives an overview of the pilots, the research projects behind and the data generated therein.

B. *Electronic laboratory notebook pilot projects ("openBIS-Pilots")*

To practice an active handling of research data, an Electronic Laboratory Notebook (ELN) was tested in two further pilot projects. Since the ZHAW was a project partner of the DLCM 2.0 as well as of the openRDM.swiss project, the focus was on the implementation and use of openBIS[3].

Fig. 2. Research data lifecycle. Research data should be actively managed throughout the entire research data lifecycle.

Two very different use cases were selected. openBIS was implemented at the Polymer Chemistry Laboratory of the Institute of Chemistry and Biotechnology. Another implementation was at the Movement Laboratory of the Institute of Physiotherapy. The tasks included the identification of laboratory workflows and configuring the tools for data capturing.

TABLE I. Overview Open Research Data Pilot Projects ("ORD-Pilots")

Department which run the pilot(s)	Pilot abbr.	Research project (URL to ZHAW project data base)	Funding of research project	Data description (published data only)
Architecture, Design and Civil Engineering	A	Criteria and strategies for the densification of settlement structures in the post-war period	Federal Office of Culture, Foundations	Digitized physical architectural models
Health Professions	H	Digital Parent Advisor	SAMW/ASSM, Käthe-Zingg-Schwichtenberg-Foundation	Survey, Focus groups
Applied **Linguistics**	L	various	various	Textual data (XML)
Life Sciences and Facility Management	N	Strategies to develop effective, innovative and practical approaches to protect major European fruit crops from pests and pathogens (DROPSA) Diagnostic and epidemiological tools for the *Xanthomonas hortorum* species-level clade based on OMICs technologies (XhortOMICs)	EU (FP7, No. 613678) SNSF (No. 177064)	Genome and transcriptome sequence data
Applied Psychology	P1	The impact of family stress on children in transition into puberty: The interplay of social and emotional processes	SNSF (No. 132278)	Survey (longitudinal study)
	P2	Preschool children, their media use and health aspects	Swiss Health Observatory OBSAN	Survey
Social Work	S	Educating children to the world? An ethnographic study on conceptions of social order of practitioners in care institutions for children and adolescents.	SNSF (No. 169727)	Interviews, Observation protocols
Engineering	E1	Nanoporous diaphragms for electrochemical sensors (NanoDiaS)	CTI (No. 16851.1 PFNM-NM)	Tomography data, Property data
	E2	NRP70 joint project: Renewable fuels for electricity production	SNSF (NRP 70, "Energy Turnaround")	Survey, Code, Calculation tool, Tabular data
Management & Law	M	Data Monitoring Local Communities in Switzerland	SNSF (No. 162948)	Survey

[2] Due to the incomplete paper publication of the research project behind pilot "P2" (see Table I), no research data were published. Instead, additional focus was placed on the handling of sensitive data and data anonymization.

[3] openBIS is developed by the Scientific IT Services of ETH Zurich. The tool is used for digital note taking, inventory management and data management.

III. KEY FINDINGS "ORD-PILOTS"

A. *Identification and evaluation of (discipline) specific repositories*

The process of identifying and evaluating discipline-specific data repositories was strongly depending on the pilot project and the domain. However, many of the pilots started to gain an overview over the available repositories by looking at existing studies/recommendations [4] or by browsing on re3data.org, a registry of research data repositories. FAIR data repositories with certificates (e.g. CoreTrustSeal [5]) were preferred. This approach provided an initial selection of data repositories. In most cases, this was followed by a search for comparable data sets to check the matching of research subject and discipline. This was widely considered as one of the most important criteria to increase the outreach of the data publication. In domains of social sciences and humanities, emphasis was placed on ensuring that the language and geographic scope match. For example, it was assumed that the publication of a German-language dataset with a strong study reference to Switzerland should be published in a national repository if possible.

Table II shows how the pilots assessed various criteria to evaluate suitability of data repositories. Our pilots confirmed that not only a matching research domain is important, but also specific metadata schemes that allow a suitable description and cataloguing of the data. This was widely considered as essential to find, assess and reuse data sets.

Based on the criteria described above, the choice of a suitable repository in the field of social sciences and humanities was relatively clear (Pilots P1, P2, S, M). This fell on FORSbase [6]. The geographical scope, sophisticated metadata schemes and an established community spoke for it. The same applies to the area of genomics, where data sets from two projects were published and the choice fell on established repositories (see Table III). There are two interesting aspects to be mentioned here: first, the data from three commonly used repositories – including our chosen repositories – are mirrored as part of an international collaboration (INDSC[7]). This leads in practice to a better findability and data redundancy. Second, data publications are in the field of genomics often mandatory. A data accession number must be provided before peer review. Furthermore, journals often specify data repositories which are to be used.
In the field of health sciences (Pilot H), the choice fell on the also established Harvard Dataverse [8], which, with its international community, represented an interesting contrast to the publication in FORSbase.
The interdisciplinary team behind the project of the national research program NRP 70 (Pilot E2) has decided – with one exception (Mendeley data) – to publish on the generic repository Zenodo.
In three pilot projects (architecture, applied linguistics, engineering) it was considered that data publication requires specific developments or special data discovery features. In the case of Architecture (Pilot A), digitized architectural

TABLE II. EVALUATION CRITERIA OF DISCIPLINE-SPECIFIC DATA REPOSITORIES

Criteria			Pilots										Practical procedure
			A	H	L	N	P1	P2	S	E1	E2	M	
General	Compliance to standards and FAIR principles		▒	■	▒	■	■	■	■	■	■	■	Check for trustworthiness and certificates (e.g. CoreTrustSeal)
Peer-group & outreach	General subject & discipline match		▒	■	■	■	■	■	■	■	▒	■	Check for similar data sets. Check language and geographic scope of other data sets.
	Geographic match & language	Swiss region	▒		■		▒	▒	■		▒	▒	
		German language					■	■	■			■	
		International		■	▒	■				■	▒		
Discipline-specific properties	Discipline-specific metadata scheme		▒	■	■	■	■	■	■	■	▒	■	Discipline-specific metadata schemes were considered as valuable to discover research data and evaluate reuse
	Data discovery features		■		■					■			
	Project specific developments		■										
Support			■	▒		▒	■	■	■			■	Support was generally considered as valuable. E.g. for data protection and licensing questions.
Other	Download approval						■	■	■			■	Check if other features are needed
	Data set versioning			■			■	■	■	■	■	■	
	PID/DOI reservation		■	■		■	■	■	■	■	■	■	
	Special submission workflows and API					■				■			

■ considered as relevant criteria for the pilot
▒ considered as less relevant criteria for the pilot
□ considered as not relevant criteria for the pilot or not discussed

PID: Persistent Identifier
DOI: Digital Object Identifier

[4] e.g. Milzow et al. (2020); von der Heyde (2019)
[5] https://www.coretrustseal.org
[6] https://forsbase.unil.ch
[7] http://www.insdc.org
[8] https://dataverse.harvard.edu

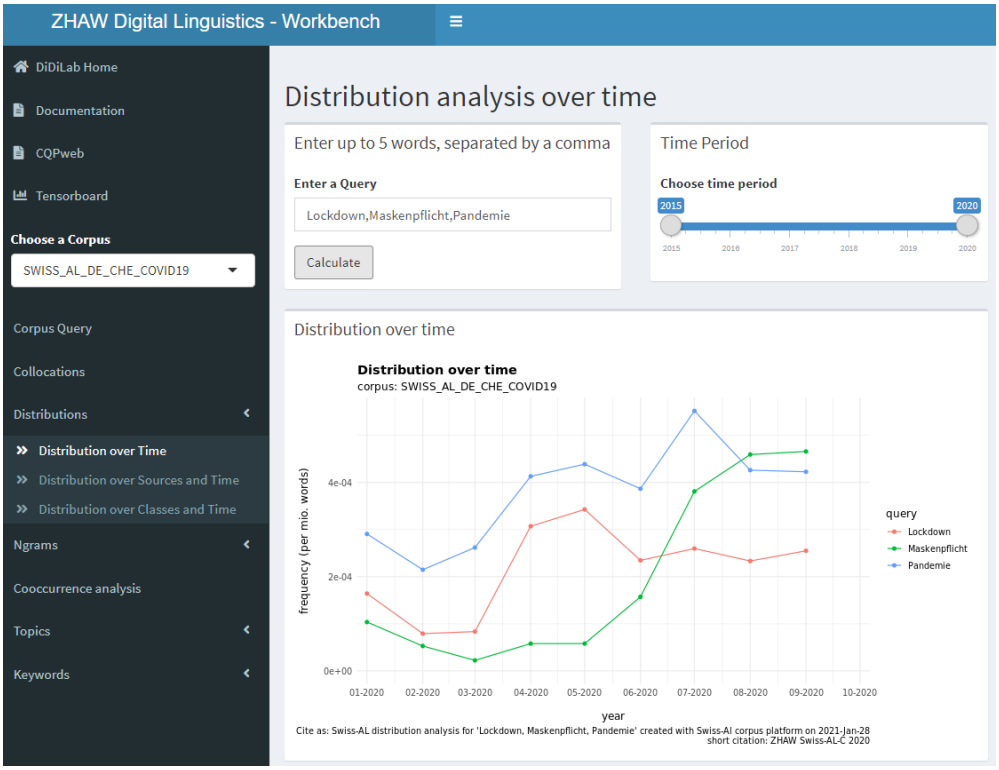

Fig. 3. Corpus linguistic workbench (accessible under https://swiss-al.linguistik.zhaw.ch)

models were to be embedded three-dimensionally in a landscape model. The technologies required for this are now being used for the first time by the Data and Service Center for the Humanities (DaSCH [9]). Due to this novel implementation, this data publication is still in progress. In the case of Applied Linguistics (Pilot L), extensive corpus linguistic data were published on an analysis workbench (Fig. 3). The corpus data consist basically out of pre-processed and aggregated texts that come from publicly available websites (e.g. from federal administration, politics, education, social media). However, due to copyright reasons, not all derivatives can be published (e.g. derivates from newspapers articles). The workbench includes various analysis tools, so that also non-linguists can now perform data based linguistic analysis.

The appropriate publication of 3D tomography data in the field of materials science was rather challenging (Pilot E1). Due to large amounts of data (several gigabytes) and the advantageous coupling of image data with material property data, a suitable portal with dedicated discovery features would be useful in this area. For the time being, the data was published on Zenodo. Parallel to the pilot projects, activities have now been started within the community with the aim of developing a portal for 3D material data.

When publishing research data in discipline-specific repositories, support from the repository operators were also mentioned as an important criterion. In some cases, support was appreciated when it came to questions related to data licensing, data anonymization and data access control. Further technical criteria considered important are the possibility of a download approval, versioning of data sets and a reservation of Digital Object Identifier (DOI). When uploading or downloading large amounts of data, such as in genomics, programming interfaces (API) can be useful. Some pilot teams also mentioned an appropriate user-experience, data accession metrics and the connection to a long-term preservation system as important.

B. Impact and practical experiences of data sharing

One of the aims of the pilot projects was to determine the impacts and benefits that emerged from the data publications. Several studies have shown benefits of data publication, for example by indicating an overall potential increase in scientific efficiency through reuse of data (Pronk, 2019). Christensen et al. (2019) found researchers to get more citations if they publish research data. Our setting of the pilot projects allowed to have a very practical approach to find answers on this question. However, a comprehensive statement about the effects of the data publication of all pilot projects does not seem trivial. Mainly because the quality criteria for such an assessment are unclear. Further, some of the data sets had only been published for a few months by the end of these pilot projects. Some of the data publications were also only downloadable after specifying the purpose of the data reuse and after approval by the researchers. Finally, we collected several indicators to assess the impact our publications (Table III).

Our practical and pragmatic finding is that a data publication is successful if the target community is reached and there is demand for the supplied data. In our case, data publications on FORSbase contributed to networking activities and potentially new partnerships in several cases. The publication of a scientific article in Elsevier, as well as the associated source code of a fuel cell model (published on Mendeley Data), have probably even paved the way for a successful submission of a new EU-funded R&D project and a new business idea. Finally, the publication of corpus linguistic data was already gaining some popularity. Since August 2020, the team of digital linguists has been holding workshops that enable researchers to explore the linguistic

[9] https://dasch.swiss

TABLE III. OVERVIEW PUBLICATIONS OF "ORD-PILOTS" AND IMPACT

Pilot short name	Repository	Digital Object Identifier (DOI)	Publication date	months online	authorisation? [a]	views	downloads	citations	Contact requests	Notes on the impact
A	DaSCH	Data publication in progress								
H	Harvard Dataverse	10.7910/DVN/JI9GIJ	24.09.19	16	N	n/a	71	n/a	N	
L	ZHAW	Link to Digital Linguistics Workbench	Sept. 2019	16	N	n/a	-	n/a	Y	Gained already popularity; valueable basis for new transdisciplinary projects. Workshops are hold for researchers to exploit data
N	European Nucleotide Archive (ENA) Gene Expression Omnibus (GEO/NCBI)	Accession PRJEB25730 Accession PRJEB27248 Accession PRJEB38812 Accession GSE150636	22.04.18 14.06.18 28.07.20 25.08.20	33 31 6 5	N	n/a	n/a	n/a	N	
P1	FORSbase	10.23662/FORS-DS-1086-1 10.23662/FORS-DS-1089-1 10.23662/FORS-DS-1090-1	07.01.20	12	Y	n/a	n/a	n/a	N	
P2	FORSbase	No data publication within this pilot. Preferred repository was specified.								
S	FORSbase	10.23662/FORS-DS-1129-1	29.04.20	9	Y	n/a	2	n/a	Y	Requested twice for education purposes
E1	Zenodo	10.5281/zenodo.4049960	25.09.20	4	N	28	2	0	N	
E2	Zenodo	10.5281/zenodo.3365919	12.08.19	17	N	38	75	0	0	
		10.5281/zenodo.3740888	06.04.20	9	N	37	22	0	0	
		10.5281/zenodo.3744301	08.04.20	9	N	42	12	n/a	0	
	Mendeley Data	10.17632/2msdd4j84c.1	24.08.18	29	N	1418	294	0	Y	Supported submission of EU-funded project and a new business idea
M	FORSbase	10.23662/FORS-DS-1116-1	04.12.19	14	Y	n/a	10	n/a	Y	10 data requests from education, research & media

Numbers from January 2021

[a] FORSbase allows to set a download authorization of max. 3 years

data. This laid the foundation for new and transdisciplinary research projects, as for example within COVID-19 research (ZHAW, 2021).

Based on the findings of the pilot projects, we propose the following recommendations:

- *Data curation is important*. Only the part of research data for which a demand can be expected should be published[10].
- *Publication of research data in discipline-specific repositories is a key factor for impact.* Discipline-specific repositories contribute to the quality and findability of research data by offering support and specific metadata schemes.
- *Linking paper publications and ORD increases outreach.* Use a DOI to refer to ORD from the paper.

C. *Implications for discipline-specific research data management/workflows*

Our pilot projects showed a variety of types of research data as well as different ways in which they are collected and methodically and technically processed (Fig. 4). This statement can be made even within similar research domains.

A major challenge has been dealing with sensitive data in the social sciences and humanities. We perceived a rather narrow line between maintaining reusability and a reasonable degree of anonymization. For example, when anonymizing qualitative data, it has been difficult to maintain the context and heuristic value for appropriate data reuse. One of the main difficulties was that the reuse of data and the corresponding data anonymization processes were not sufficiently considered in the project planning. This is illustrated by the fact, that in some cases informed consent was not available electronically. We state that publishing sensitive data requires a lot of background and process knowledge. In principle, knowledge and frameworks are available (e.g. Bambey et al., 2018; Elliot et al., 2020), but an efficient and pragmatic implementation remains a challenge. We think that the researchers should be given targeted support here.

Our researchers confirmed that they often use software tools which contribute to the highest productivity and which they had been using previously in their professional work. The effective practice is subject to high inter-individual variability. A common denominator in our pilot projects, however, was that the tools used were often commercial and data processing was done using non-open data formats. This meant that the data had to be converted and partially (again) documented

[10] We are aware that some funding agencies advocate the publishing of all research data. We believe that data exchange is more successful if open data sets have a well-defined scope and are demand-orientated.

before publication. The following recommendations result from this experience:

- *Consider Research Data Management (RDM) as an essential part of the project.* The collection, processing and publication of research data must be actively and in detail discussed with all stakeholders (e.g. tools & toolkits, data formats, data set language, anonymization, licenses & Intellectual Property, IP).
- If possible, give preference to *open file formats*[11] and *open source software*. The publication of the data will be easier and in accordance with the FAIR data principles.
- Try to *standardize* and *automate data processing*. This will most likely improve processing efficiency, comprehensibility and reproducibility

Based on our pilot projects, the use of open file formats and standardized data processing workflows are among the most important criteria for successful data sharing. Because of this, a culture of data sharing has established itself in disciplines such as genomics or geoinformatics (Brodeur et al., 2019; Byrd et al., 2020).

These conclusions suggest that standards and discipline-specific workflows for data collection and data processing should be developed wherever possible. Some concepts, frameworks and platforms have already been proposed: one approach is the development of so-called Domain Data Protocols (DDP), which represent a practice-oriented addition to DMPs, which are perceived as somewhat bureaucratic (Science Europe, 2018). DDPs contain specific building blocks for DMPs and for the discipline-specific management of research data. DDPs are developed by the community itself and adopted by the funding agencies. Furthermore, innovative technical workflow frameworks (e.g. Canonical Workflow Frameworks) could fundamentally change data processing in the future (Hardisty & Wittenburg, 2020). This approach basically involves the fragmentation, reassembly and automating of data processing workflows with the aim of making data processing more efficient and reproducible. Finally, new platforms such as RENKU[12] could also offer the technical basis for mapping data workflows as completely as possible and making the data and results publicly accessible in the sense of Open Science.

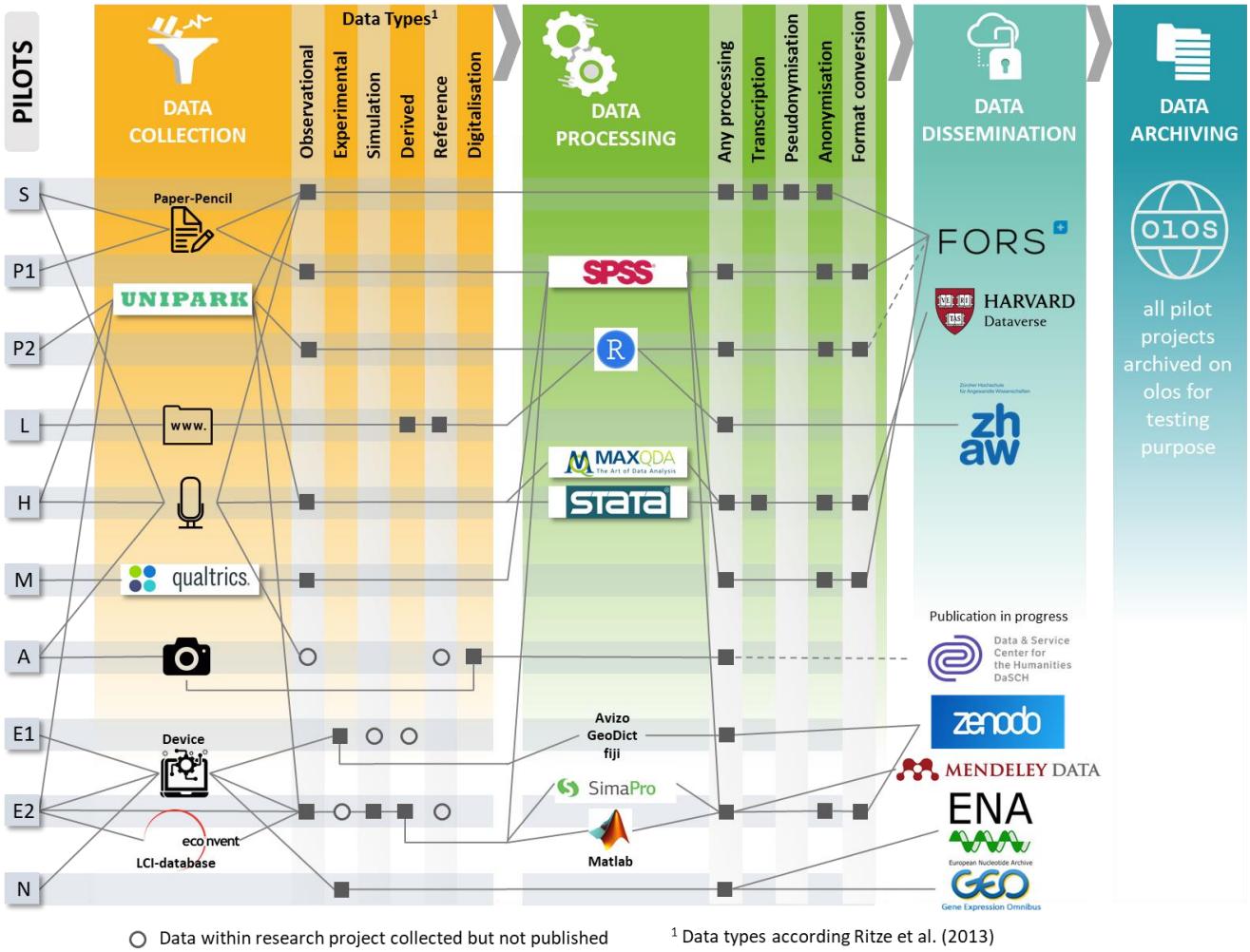

Fig. 4. Data processing workflows of ORD pilot projects.

[11] Possible source of information is the UK Data Service (UK Data Service, 2021).

[12] https://renkulab.io

IV. KEY FINDINGS OPENBIS-PILOTS

The overarching goal of the "openBIS-Pilots" was to practice Active Research Data Management (ARDM). By ARDM we mean the use of tools and skills that go beyond storing research data in a file/folder-based system. Options to practice ARDM are the use of Electronic Laboratory Notebooks (ELN) or Electronic Data Capture systems (EDC). As mentioned, we basically used the ELN/LIMS-system[13] openBIS in two laboratories at ZHAW. One laboratory additionally tested REDCap[14] for data capturing.

A. *General conclusions of the "openBIS-pilots"*

ELNs have become useful and in some cases indispensable tools in experimental research. This is due, for example, to the fact that many tools support the principles of Good Laboratory Practice (GLP) and measures to ensure data integrity (e.g. audit trail feature). Additionally, the tools offer many other features as well as interfaces to third-party applications. Numerous products with comparable properties are available [15]). Another conclusion is that the implementation, configuration and user training require considerable effort. Various handouts have been published to facilitate entry into the world of ELNs and their implementation (DLCM, 2017c; Kwok, 2018; ZB MED, 2020). However, we consider the introduction of ELNs to be a complex process that require careful planning and approaches that are known from Requirements Engineering. Based on the experience gained in the openBIS-Pilot projects, the following aspects and practical procedures seem particularly important to us:

- *Involve end users.* Consider aspects that favour on-boarding and sustainable use (e.g. usability).
- *Identify the benefit of ELNs at different workflow levels.* For this purpose, we recommend sketching the laboratory processes and defining the requirements (e.g. features, data protection standards)
- *Features are used repeatedly when they increase productivity and quality.* Users will fall back into traditional file/folder-based (or paper) documentation and data storing if they cannot take advantage of available ELN/EDC features (e.g. integration of data, scripts, annotation).
- *Consider an iterative implementation.* It may be difficult to capture all data processing workflows and user needs from the beginning: start small and expand.
- *Support & community.* Implementation, configuration and application should be supported by qualified staff. Establishing contact to the user community and to developers is also important.

When implementing an ELN, close support seems to be the most important success factor. We recommend the implementation of pilot projects to be able to transfer best practices.

B. *Practical experiences from the implementation at the ZHAW's Polymer Chemistry Laboratory*

The Polymer Chemistry Lab of ZHAW focuses on the synthesis, functionalization, and characterization of nanostructured polymeric materials. Because of the processes and research methods used, there was early evidence that using an ELN might be beneficial.

The implementation started with a recording of the inventory and the usual research work steps. This included, for example, the sample archive, Standard Operating Procedures (SOP), the device infrastructure and data processing. This was followed by an initial and rather targeted configuration of openBIS. The templates were then iteratively improved and

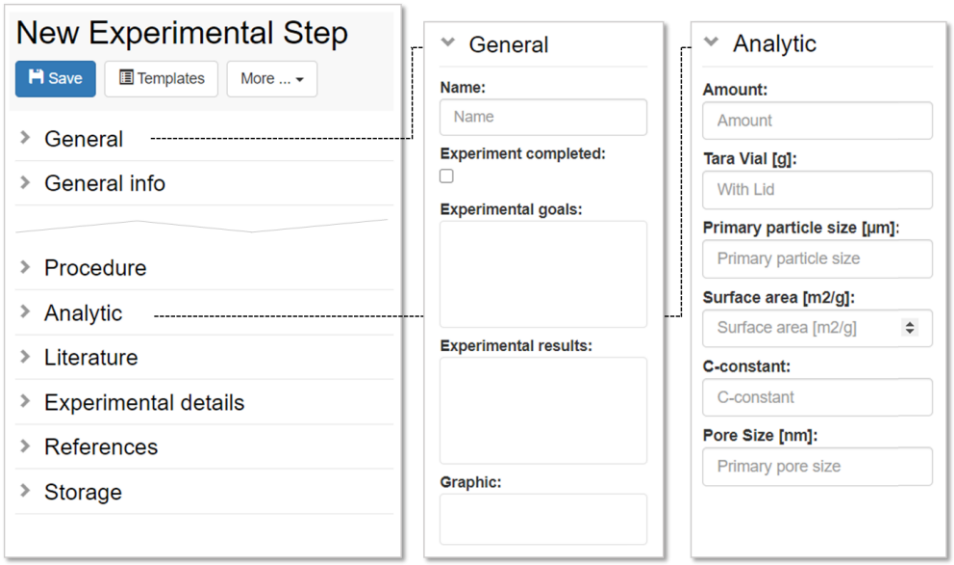

Fig. 5. Generic openBIS template with additional sections and specific metadata fields to improve searchability.

[13] openBIS is a combination of an ELN and a Laboratory Information Management System (LIMS).

[14] REDCap is an Electronical Data Capture System (EDC) which is developed by the Vanderbilt University (https://projectredcap.org).

[15] Overviews to be found e.g. in DLCM (2017a, 2017b); Harvard Medical School (2021).

additional features added. From this experience we draw the following conclusions for our use case (which we consider a classical use case):

- *Low-level, generic templates are preferred.* Over-structuring the template hinders flexibility. Instead, a generic template respects the diversity of projects. Additional, optional sections with documentation that build on each other can be added as required (Fig. 5).
- *Searchability of projects and experiments is a key feature.* Laboratory-specific metadata fields should be added to improve the findability and reuse of experimental data, information and knowledge (Fig. 5).

C. *Practical experiences of the implementation at the ZHAW's Movement Laboratory*

The Movement Lab at ZHAW focuses on the analysis of movement sequences and muscle activities using state-of-the-art technology. The research projects mostly include an *in situ* recording of measurements from test persons. This leads to the need for clearly structured and efficient process flows as well as increased requirements for data protection. For these reasons, an ELN had to meet special requirements:

- Very high usability for easy, secure and time-efficient data capturing.
- Enable validation and plausibility check of data during input.
- Compliance to data protection rules for personal data when storing and accessing data (e.g. including track changes and activity logs).

The implementation started with a definition of a standard project procedure (Fig 6). This contains, among other things, an Informed Consent Form (ICF), SOP, a Case Report Form (CRF) and data processing in Matlab. The goal was to implement the SOP and the CRF with participant data in openBIS. The implementation of the SOP in openBIS was easy to accomplish. On the other hand, the implementation of our sophisticated CRF resulted in insufficient flexibility in data entry and no direct validation. For this reason, a Jupyter Notebook was used and coupled with openBIS.

This solution combining openBIS, Jupyter Notebooks and Matlab for data processing basically works. In practice, however, it was found that the solution is rather complex and further development or adaptation to new projects is difficult. This is also due to the interfaces and the two different programming languages used in Jupyter and Matlab (Python in Jupyter). For these reasons, REDCap was tested as an alternative to implement the CRF. As part of a user study, these two approaches (openBIS/Jupyter vs. REDCap) were compared; REDCap turned out to be the more user-friendly solution in our case. Based on this experience we draw the following conclusions for our use case:

- *Data security is the first hurdle.* The requirements must be carefully checked.
- *A guided data entry and immediate validation can be a difficult task for ELNs.* Consider EDC-systems.
- *Dependencies on specific tools and programming language are problematic.* Universal programming languages such as Python as well as stand-alone executables offer better flexibility.

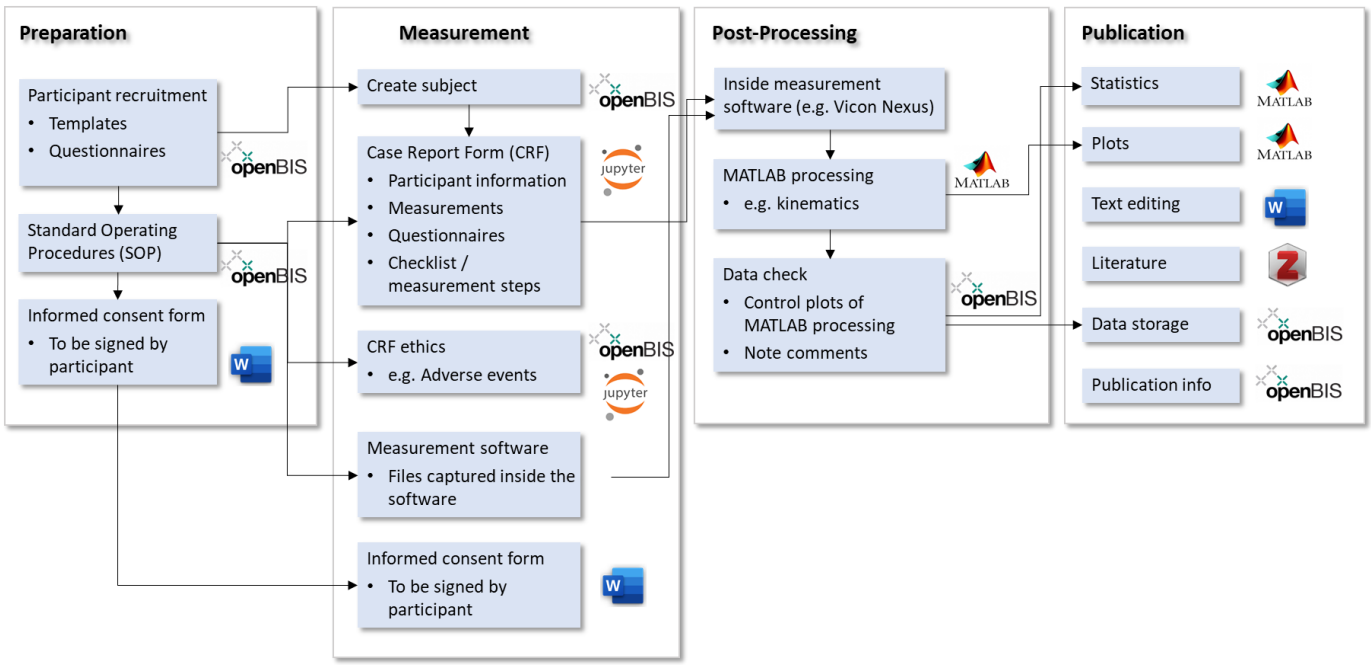

Fig. 6. Standard project procedure of the Movement Laboratory at ZHAW with pilot implementation of openBIS and Jupyter.

V. IMPLICATIONS FOR RESEARCH DATA MANAGEMENT AT UNIVERSITIES (OF APPLIED SCIENCES)

The pilot projects have made it clear that the publication of research data places additional demands and (time) expenditure on the management of research data. Although the reuse of openly available research data is largely defined by the research community and subsequent users, universities and research institutions have a crucial role in fostering good RDM and data publication practices, services and infrastructures.

For this reason, a working group has developed a concept/framework for research data support services in parallel to the ongoing pilot projects. This working group consists of members of the central research support, the university library and the ICT and will be given a permanent mandate after finalisation of the pilot projects. The cooperation of different university units in the development of services appears to be advantageous for the purpose of bundling resources and competencies. Other universities are successfully pursuing similar models (Sesartic Petrus & Töwe, 2019).

In our opinion, our concept can at least partially be transferred to other universities (of applied sciences). At the ZHAW, three basic levels of action regarding RDM were identified:

1. Normative level
2. Tool and infrastructure level
3. Support level

A. *Normative level*

The normative level contains top-level, institutional regulations and policies regarding ORD. At ZHAW, the strategic positioning and implementation of Open Science was integrated into the top-level institutional R&D policy (ZHAW, 2019). This policy contains the approaches for the implementation of open R&D processes and urges the consideration of legal and ethical obligations (protection of sensitive data), as well as contractual obligations with application partners (e.g. IP). More precise specifications in relation to ORD may be integrated at a later stage depending on the strategic development at the tool and infrastructure level as well as on legal considerations.

B. *Tool and infrastructure level*

The tool and infrastructure level includes the provision of ICT tools for the (active) management of research data. The aim is:

- to be able to offer appropriate tools and professional support over the entire research data life cycle
- to streamline the use of RDM tools across research groups and disciplines (to improve the ability of university IT and support services to handle new tools)
- to identify the potential to build standardised, automated data processing workflows

It is considered important that tools are open source, or at least support open formats. Researchers at ZHAW already use a portfolio of applications, which is now being continuously expanded according to the specific needs of the different departments (Fig. 7).

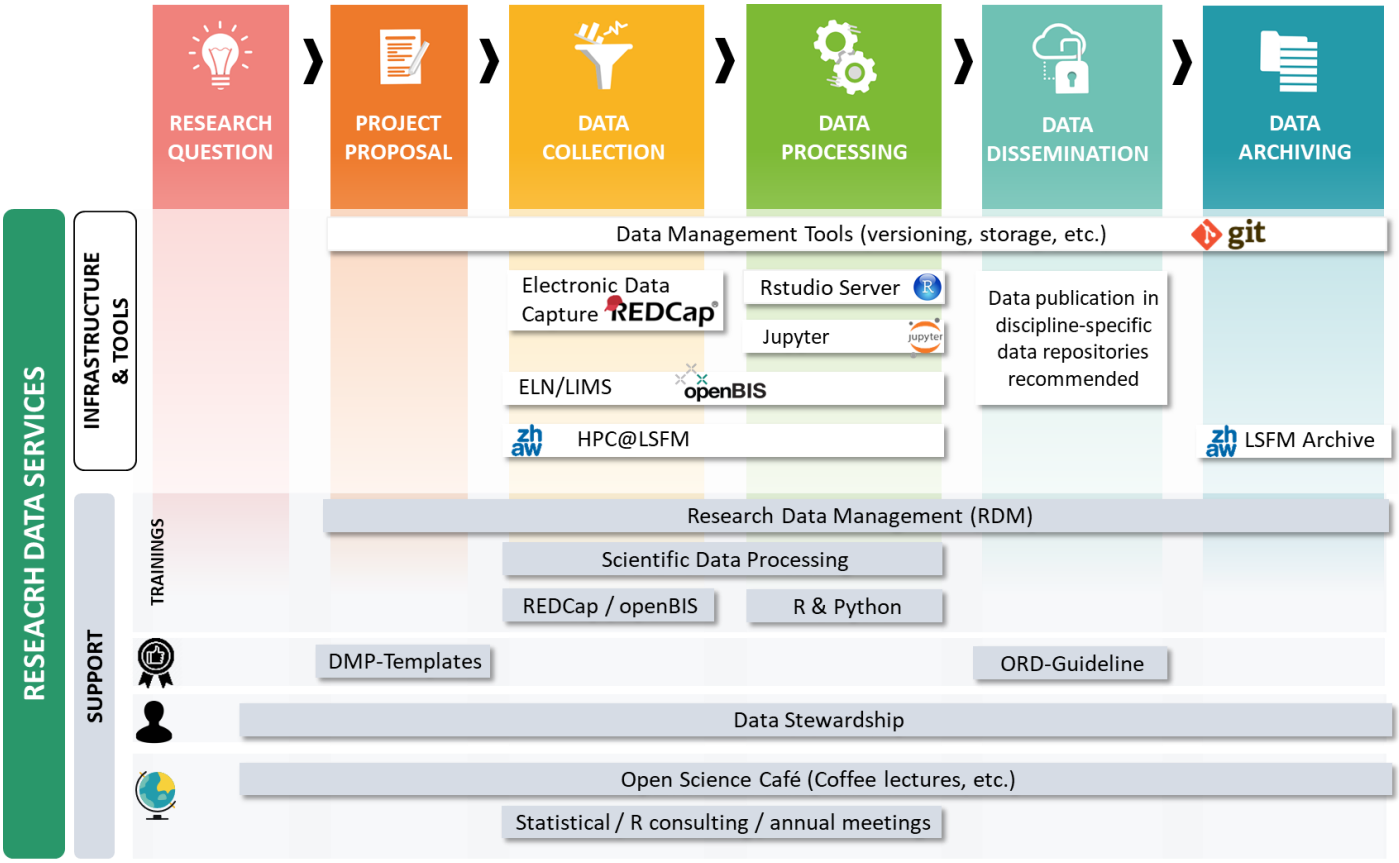

Fig. 7. Starting portfolio for Research Data Management services at Zurich University of Applied Sciences (from an ORD perspective; further tools and services are available).

At the infrastructure level, the aim is to provide more complex systems for managing research data. This may include an institutional data repository (or a solution for institutional management of research data), a long-term archive for research data or even infrastructure for domain-specific data repositories. The development in this regard is still dependent on the availability and design of national infrastructures, legal considerations and institutional requirements.

C. *Support level*

The support level includes training, development of best practices and support in the field of Research Data Management and scientific data processing. As already mentioned, the aim is to support researchers along the entire research data life cycle. We use the term “data stewardship” for this purpose. Data stewardship models have already been previously proposed or are already being used successfully (Dunning & Teperek, 2019; Mons, 2020; Swiss Academies Of Arts And Sciences et al., 2019). The core of our data stewardship model consists of several professionals ("data stewards") who have different technical and disciplinary backgrounds (data-, information-, computer scientists). The philosophy is to support the researchers “hands-on” and aims at helping researchers to make the most of their data (e.g. in terms of public resonance, accessibility, interactivity etc.). This is facilitated by defining one clear central contact for researchers.

Another focus of the data stewards is the development of data processing workflows suitable for ORD. From our point of view, script-based programming languages such as Python and R play a key role here. The data stewards help to develop high-level skills in this area, including, for example, curating and making program libraries available.

Furthermore, our pilot projects show that data anonymization plays a central role in data publication. The data stewards also offer advice here or liaise with other bodies[16]. The data stewards could also take on tasks in the generation of synthetic data sets, which might be published increasingly due to data protection reasons[17]. If necessary, other departments are involved: for example, the legal service or the data protection officer. This is the case, for example, when it comes to the design of Informed Consent Forms, data protection issues or data licensing.

The data stewards, or the newly created unit “ZHAW Services Research Data” as a whole, also takes over coordinating activities, for example, when it comes to the integration of existing or new (national) services into the RDM service portfolio.

Our data stewards also support existing or new communities in the implementation of discipline-specific research data management. The main goal is to provide platforms and opportunities for exchange of good practices in RDM. The support of some already existing events such as the annual statistics meeting at ZHAW or the exchange among our internal statistical and R consultants have been integrated into the RDM service portfolio (Fig. 7). More such communities are likely to form soon, e.g. for data anonymization or the handling of qualitative data. Finally, we recently founded the Open Science Café[18], which uses various formats to provide individual researchers the opportunity to exchange ideas.

VI. Conclusion

Our pilot projects gave an insight into the immense diversity of research data as well as into very individual and partly complex data processing workflows. Research data must therefore be documented in detail to ensure the comprehensibility and reproducibility of the data. We have also observed that a large part of the pilots' research data was generated on a project-specific basis. For these reasons, we believe that research data should be understood as a highly complex and project-specific product of research. We consider this basic understanding of the characteristics of research data to be important as it determines how research data can or should be shared.

We have illustrated here that the rather new practice of data sharing is successful if the available data sets meet the requirements and demands of a particular (scientific) community. To fulfil these requirements and demands, research data must be collected, processed, documented, curated and published according to discipline-specific best practices. Such best practices and standards are only established in a few domains (e.g. in genomics or for geodata). In our opinion, other areas of science should follow suit. Concepts for the implementation of such practices (such as Data Domain Protocols DDP) and initiatives (such as the Research Data Alliance RDA) have already been proposed. Therefore, we consider it essential that research communities are supported on a local, national and international level to implement such concepts.

A key factor of ORD is a comprehensive and professional management of research data throughout the research process. Hence, RDM must be part of the project and included in the project planning. Our pilot projects showed that Active Research Data Management and the use of appropriate tools (e.g. ELNs) make an important contribution to the quality, comprehensibility and efficient handling of research data. Wherever possible, open source tools and open data formats should be used in RDM to better meet FAIR data principles and enhance flexibility within the RDM ecosystem of a higher education institution.

In our opinion, the best way to publish research data is to use discipline-specific repositories. These offer specific metadata schemes, which contribute to data quality and significantly increase the findability of the research data. Repository operators can also react to the individual needs of the communities and provide domain-specific features and support.

The availability of support, tools and infrastructure is another condition for the success of ORD. Supporting researchers is the responsibility of the universities. For this reason, a data stewardship model was introduced at the ZHAW that supports researchers throughout the entire research data life cycle according to a "hands-on" philosophy. We have learned that this task is complex and requires

[16] Such as with FORSbase or Qualiservice (https://www.qualiservice.org)
[17] see Burgard et al. (2017)
[18] Our (public) Open Science Café (https://bit.ly/39o5TCb) is a virtual space hosted by wonder.me

cooperation between several organizational units (e.g. library, research support, ICT) and specialists (e.g. data curators, data scientists, computer scientists, data protection officers). On the one hand, this is due to the diversity and complexity of the research data and data processing steps; on the other hand, the publication of research data and its re-use are usually opposed to other interests (e.g. data protection, IP). Providing researchers the necessary tools and technical support throughout the entire research process is also the responsibility of the universities. Given the diversity of tasks, universities could potentially also cooperate in the support of their researchers.

Certain tools and infrastructure should be developed and made available at the national / international level – but only if they manage to gain sufficient support of the relevant user communities.

For the success of ORD – and Open Science – it is ultimately also decisive how researchers are assessed. Policy makers, third-party funders and universities must therefore go ahead and pay the same attention to ORD as to OA publications. For these reasons, a positioning on ORD was included in the general R&D policy at the ZHAW as well as corresponding measures to support open R&D processes. The most immediate and essential measure consisted in establishing a new service unit (ZHAW Services Research Data) which implements the data stewardship model and provides researchers with infrastructure, tools and support. The overarching goal is to release the potential of the researcher's data in the sense of Open Science.

Remarks

The results of this paper reflect the practical experience of Research Data Management over the entire life cycle of the research data obtained by the ZHAW pilot projects. As part of the DLCM 2.0 project, the ZHAW contributed to the completion of DLCM services in various other ways. This included, for example, the co-development and testing of the DLCM archiving solution (OLOS) and the associated professional services.

Acknowledgment

We thank swissuniversities for funding the DLCM 2.0 project and the University of Geneva and HES-SO for leading the project. We would also like to thank the SIS team at ETH Zurich for the great openBIS support. Finally, we would like to thank all the pilot project members of the ZHAW who were not mentioned by name.

References

Bambey, D., Corti, L., Diepenbroek, M., Dunkel, W., Hanekop, H., Hollstein, B., Imeri, S., Knoblauch, H., Kretzer, S., Meier Zu Verl, C., Meyer, C., Meyermann, A., Porzelt, M., Rittberger, M., Strübing, J., Von Unger, H., & Wilke, R. (2018). Archivierung und Zugang zu Qualitativen Daten. *RatSWD Working Paper Series*. https://doi.org/10.17620/02671.35

Brodeur, Coetzee, Danko, Garcia, & Hjelmager. (2019). Geographic Information Metadata—An Outlook from the International Standardization Perspective. *ISPRS International Journal of Geo-Information*, *8*(6), 280. https://doi.org/10.3390/ijgi8060280

Burgard, J. P., Kolb, J.-P., Merkle, H., & Münnich, R. (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts- Und Sozialstatistisches Archiv*, *11*(3–4), 233–244. https://doi.org/10.1007/s11943-017-0214-8

Byrd, J. B., Greene, A. C., Prasad, D. V., Jiang, X., & Greene, C. S. (2020). Responsible, practical genomic data sharing that accelerates research. *Nature Reviews Genetics*, *21*(10), 615–629. https://doi.org/10.1038/s41576-020-0257-5

DLCM. (2017a). *Curated list of ELNs.* https://www.dlcm.ch/application/files/8315/1368/6983/ELN_list_December_2017.xlsx

DLCM. (2017b). *Curated list of LIMS.* https://www.dlcm.ch/application/files/4415/1368/6918/LIMS_list_December_2017.xlsx

DLCM. (2017c). *Guidelines for introducing an ELN/LIMS in academic research laboratories*. https://www.dlcm.ch/application/files/3915/1368/9573/DLCM_ELN_LIMS_guidelines.pdf

Dunning, A., & Teperek, M. (2019). *Strategic Framework for Data Stewardship at TU Delft 2020 to 2024*. https://doi.org/10.5281/ZENODO.3565506

Elliot, M., Mackey, E., & O'Hara, K. (2020). *The Anonymisation Decision Making Framework: European Practitioners' Guide (2nd edition)*. UK Anonymisation Network.

EU. (2017). *H2020 Programme—Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020*.

Hardisty, A., & Wittenburg, P. (2020). *Canonical workflow framework for research CWR - Position Paper, version 2.* https://osf.io/3rekv/

Harvard Medical School. (2021). *Electronic Lab Notebooks. ELN Comparison Grid.* https://datamanagement.hms.harvard.edu/analyze/electronic-lab-notebooks

Kwok, R. (2018). How to pick an electronic laboratory notebook. *Nature*, *560*(7717), 269–270. https://doi.org/10.1038/d41586-018-05895-3

Milzow, K., von Arx, M., Sommer, C., Cahenzli, J., & Perini, L. (2020). *Open Research Data: SNSF monitoring report 2017-2018*. Zenodo. https://doi.org/10.5281/ZENODO.3618123

Mons, B. (2020). Invest 5% of research funds in ensuring data are reusable. *Nature*, *578*(7796), 491–491. https://doi.org/10.1038/d41586-020-00505-7

Pronk, T. E. (2019). The Time Efficiency Gain in Sharing and Reuse of Research Data. *Data Science Journal*, *18*, 10. https://doi.org/10.5334/dsj-2019-010

Ritze, D., Eckert, K., & Pfeffer, M. (2013). Forschungsdaten. In P. Danowski & A. Pohl (Eds.), *(Open) Linked Data in Bibliotheken*. DE GRUYTER SAUR. https://doi.org/10.1515/9783110278736.122

Science Europe. (2018). *Science Europe Guidance Document—Presenting a Framework for Discipline-specific Research Data Management*.

Sesartic Petrus, A., & Töwe, M. (2019). Forschungsdatenmanagement an der ETH Zürich: Ansätze und Wirkung. *Bibliothek Forschung Und Praxis*, *43*(1), 49–60. https://doi.org/10.1515/bfp-2019-2002

SNSF. (2021). *Open Research Data*. http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/default.aspx

Swiss Academies Of Arts And Sciences, Roger, P., Gerhard, L., Donat, A., Appenzeller Claudia, Stéphanie, G., Daniel, H., Beat, I., Jérôme, K., Cécile, L., Gabi, S., & Yilmaz Aysim. (2019). *Open Science in Switzerland: Opportunities and Challenges*. Zenodo. https://doi.org/10.5281/ZENODO.3248929

UK Data Service. (2021). *UK Data Service. Recommended formats.* https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

von der Heyde, M. (2019). *Open Research Data: Landscape and cost analysis of data repositories currently used by the Swiss research community, and requirements for the future* (1.0.0). Zenodo. https://doi.org/10.5281/ZENODO.2643460

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

ZB MED. (2020). *Elektronische Laborbücher im Kontext von Forschungsdatenmanagement und guter wissenschaftlicher Praxis—Ein Wegweiser für die Lebenswissenschaften* [Application/pdf]. https://doi.org/10.4126/FRL01-006422868

ZHAW. (2019, November 1). *F&E Policy Zurich University of Applied Sciences*. https://gpmpublic.zhaw.ch/GPMDocProdZPublic/1_Management/1_04_Governance/1_04_01_Fuehrungsgrundlagen/Z_PY_F_und_E_Policy_ZHAW.pdf

ZHAW. (2021). *Digital Transfer Platform for COVID-19 Research*. https://www.zhaw.ch/en/research/research-database/project-detailview/projektid/3623/