Revue électronique Suisse de science de l'information

Swiss Research Data Day

DLCM 2.0

Préface

Le Swiss Research Data Day 2020 (SRDD2020) organisé par le projet Data Life-Cycle Management (DLCM, https://dlcm.ch) a eu lieu le 22 octobre 2020 dans les murs de la Haute école de gestion de Genève (HEG-GE) en Suisse.

Un mot de bienvenue et une présentation annonçant le lancement d'OLOS, la solution suisse de gestion des données de recherche, ont démarré la journée. L'essentiel du programme et ses différentes parties se sont déroulés en ligne moyennant des sessions parallèles durant le colloque.

Ajoutées à ces dernières, cinq conférences plénières ont été données par Patrick Furrer de swissuniversities, Christine Pirinoli de la Haute école spécialisée de Suisse occidentale (HES-SO), Nancy McGovern du Massachusetts Institute of Technology (MIT) et Hrvoje Stancic de l'Université de Zagreb (UNIZG). La conférence de clôture, sur la gestion des données massives, a été donnée par Alberto Pace, qui a explicité les enjeux de ce domaine au sein du Centre Européen de Recherche Nucléaire (CERN).
Cette troisième édition des SRDD a donné naissance à une édition spéciale de la présente revue, réunissant les papiers présentés par plusieurs intervenants du 22 octobre. Ces papiers sont structurés en cinq thématiques.

Nuage de mots tiré des titres des conférences de la SRDD2020

L'une de ces thématiques est les données de recherche ouvertes et l'enjeu de leur gouvernance. A cet égard, deux illustrations sont présentées : une stratégie de gestion des données de la recherche développée pour NCCR Robotics et un portfolio proposé par l'équipe du Swiss Institute of Bioinformatics comprenant des bases de données et outils en ligne pour gérer les données scientifiques.

Une seconde thématique est celle de la gestion des données de recherche et des dimensions éthiques, légales, financières et académiques afférentes. A ce niveau, plusieurs intervenants ont apporté des cas pratiques, tels que DMLawTool et ses fonctionnalités qui offrent un arbre décisionnel facilitant la prise de décisions pour les questions juridiques relatives à la gestion des données, les 12 projets pilotes conduits par la ZHAW dans le cadre du projet DLCM, et d'autres questions de copyright et problématiques liées aux licences ouvertes.

Une thématique plus stratégique donne lieu à des retours d'expérience pertinents de l'UNIL et l'ETHZ.

Les deux dernières thématiques couvrent d'une part les compétences qu'implique le traitement des données ouvertes et d'autre part des cas pratiques de plusieurs chercheurs de l'EPFL et notamment du projet européen FAIR4Health, rapportant diverses expériences en matière de préparation des données en vue de leur partage et ré-exploitation. Dans cet ordre d'idées, le MOOC de DLCM a été présenté.

Plusieurs institutions, chercheurs, professionnels et experts ont suivi et contribué à cette rencontre. Au total, 42 conférenciers de plusieurs pays (Suisse, Allemagne, France, Hongrie, Croatie et Etats Unis) ont retenu l'attention de plus de 311 inscrits, qui ont suivi par Zoom ou par YouTube Live nos 28 présentations.

Je tiens à remercier chaleureusement l'équipe RESSI qui a accueilli nos articles scientifiques issus du SRDD2020 dans ce numéro spécial. Mes remerciements vont aussi à tous nos partenaires, conférenciers et bénévoles pour leur confiance et leur contributions précieuses.
Bonne lecture !
Basma Makhlouf Shabou
Prof. Dr. Basma Makhlouf Shabou
OLOS, Présidente (https://olos.swiss)

The Swiss Research Data Day 2020 (SRDD2020) organized by the Data Life-Cycle Management (DLCM, https://dlcm.ch) took place on 22th October 2020 at the Geneva School of Business Management (HEG-GE) in Switzerland.

A welcome and a presentation announcing the launch of the swiss research data management solution, OLOS, started the day. The main points of the program and its different parts were held online through parallel sessions during the symposium.

In addition, five plenary speeches were given by Patrick Furrer from swissuniversities, Christine Pirinoli from the University of Applied Science and Arts Western Switzerland (HES-SO), Nancy McGovern from the Massachusetts Institute of Technologies (MIT) and Hrvoje Stancic from the University of Zagreb (UNIZG).

The closing conference, on massive data management, was given by Alberto Pace who reported the challenges of this field within European Organization for Nuclear Research (CERN).

This third edition of the SRDD gave birth to a special issue of this review, bringing together papers presented by several speakers from October 22. These papers are structured in five themes.

Word cloud from the titles of the conferences of the SRDD2020

One of these themes focused on open research data and the issue of their governance. In this respect, two illustrations were presented: a research data management strategy developed for NCCR Robotics and a portfolio proposed by the Swiss Institute of Bioinformatics team including online databases and tools to manage scientific data.

A second theme of research data management and the related ethical, legal, financial and academic dimensions was addressed. Several speakers provided practical case studies, such as the DMLawTool and its decision tree functionality, helping to take decisions for data management related legal issues, the 12 pilot projects conducted by ZHAW in the framework of the DLCM project, and many other copyright and open licensing issues.

A more strategic theme was addressed with relevant feedback from UNIL, ETHZ.

Then, the two last themes cover on the one hand the skills involved in handling open data and on the other hand practical cases from several EPFL researchers and notably from the European project FAIR4Health reporting various experiences in preparing data for sharing and re-use. In this context, the DLCM MOOC was presented.

Several institutions, researchers, professionals and experts attended and contributed to this meeting. In total, 42 speakers from several countries (Switzerland, Germany, France, Hungary, Croatia and the United States) captured the attention of more than 311 registrants who followed our 28 presentations via Zoom or YouTube Live.

I would like to warmly thank the RESSI team for hosting the scientific papers resulting from the SRDD 2020 in this special issue. My thanks also go to all our partners, speakers and volunteers for their trust and valuable contribution.

Enjoy your reading!
Prof. Dr. Basma Makhlouf Shabou
OLOS, President (https://olos.swiss)

Table des matières

Fostering diversity, gender parity and transparency in the evaluation process of an Open Science funding program

Dr. Aude Bax de Keating
Politique des hautes écoles swissuniversities
Bern, Switzerland
aude.baxdekeating@swissuniversities.ch

ORCID 0000-0001-7901-5130

Dr. Patrick Furrer
Politique des hautes écoles swissuniversities
Bern, Switzerland
patrick.furrer@swissuniversities.ch

ORCID 0000-0003-0671-2407

Abstract—*The Scientific Information program from swissuniversities focusing on questions related to research data management, digitalization and the best practices to foster open access and open science principles has greatly evolved from 2013 to 2020. This article puts into light the transition of this program from Scientific Information towards Open Access and Open Science, where diversity, gender parity and transparency are for instance at the heart of the new program's evaluation.*
Keywords—Open Science, Open Access, Transparency, Diversity, Gender Parity, Evaluation, Reviewers, Research

I. INTRODUCTION

This article articulates the transition into open science and how this process emulates the changes to implement diversity, gender parity and transparency in the evaluation process of this new program. In which direction did the program evolve over time? What is at stake behind this transition and why is it important for the Swiss academic landscape as a whole?

II. FROM SCIENTIFIC INFORMATION TO OPEN SCIENCE

A. A short history with regards to the Scientific Information previous programmes (2013-2020)

Scientific Information has been the focus of the two 4-years previous periods 2013-2016 (CUS-P2) and 2017-2020 (P5). The evaluation process has been set in 2013-2014 and maintained without major modifications until the end of P5. The objectives of these two programs was to provide researchers, professors and students at Swiss higher education institutions with an optimal environment for the use (search, consultation, processing, visualization, storage, dissemination, sharing, reuse) of all forms of scientific information needed for their work.

Near the end of the P5 program, the expert group met to share their lessons learnt. They made recommendations to the direction of the program on how to orient and adapt the assessment of proposals and projects in the context of open science. Open Science is calling for more transparency and integrity, as well as a more comprehensive and participative governance, in particular in the evaluation process, which is the focus of this article.

B. The Open Science Program from swissuniversities (2021-2024)

The new program focusing on open science has a first emphasis on open access thanks to the Open Access Strategy and Action Plan (2018-2024)[1]. In this latter document, the goal of 100% of publicly funded research publications freely available by 2024 is presented in more depth. One of the main objectives of the Open Access Action Plan is to foster and ensure synergies, economies of scale and collaborations among Swiss higher education institutions. The Open Science Programme of swissuniversities is now implementing this Open Access Action Plan for 2021-2024.

In order to do so, a call for projects was unveiled on October 19, 2020 in order to foster collaborations among Swiss higher education to tackle current challenges to concretely implement open access on a national level[2]. Some of the projects which are particularly welcomed in this bottom-up and top down approach call include the participation of Switzerland to international initiatives, the inclusion of open science criteria in research assessment, the setting up of shared services and e-infrastructures, as well as the promotion of alternative forms of publication promoting open

[1] https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Organisation/SUK-P/SUK_P-2/PgB_OpenScience_-_Implementation_Phase_A_2021-2024_published.pdf
[2] https://www.swissuniversities.ch/en/themen/digitalisierung/open-access/conference-open-access-in-action

access[3]. Diversity, gender parity and transparency are some of the elements that the program wanted to pay particular attention to during the process of implementing the new reviewing process for this Open Science Program. Another new dimension present in the Open Science Program includes the peer review process to foster transparency. All proposals' abstracts and assessments are now published on the website in order to highlight an Open Review approach[4].

In 2020, the Open Science Delegation has provided recommendations regarding potential candidates to compose the new reviewer's pool for this Open Science program. One of the main difference is the composition, profile and number of this pool of experts. Out of the 30 candidates, 8 are based internationally outside of Switzerland, 13 are women and 17 are men[5]. New members will in addition join this first initial selection by the end of 2021 helping to include an even more diverse profile and reach an equal gender parity balance. In order to comply with the Open Science Program "Proposal", the reviewers pool composition must respect at the very least the following criteria:

More than 25% women

- More than 25% international
- More than 50% users/researchers
- More than 25% with service or e-infrastructure management expertise (e.g. legal, financial, business models)
- At least 2 reviewers specialized in diversity questions

From librarians to researchers, from lecturer to lawyer, from head of open access to head of innovation management, the professional roles, expertise and ages of the reviewers is very eclectic reflecting the diversity of the Swiss academic landscape today. A new addition to their duties is not only to evaluate projects when the proposals are submitted, but also to provide intermediary and final reviewing to assess the quality of the implementation of the selected projects over time.

The election of the new President for this reviewers' pool reflects as well a transparent process. Once the diverse group of reviewers have participated to an introductory workshop and agreed to sign the reviewers' declaration, all were invited to present their applications as president if they desired to do so by highlighting in their cover letter their experience, expertise and motivation to fill this role for the year 2021. Two candidates presented themselves as strongly interested to fill the duty of this President's role. Therefore, the reviewers pool was able to vote confidentially for their favorite candidate[6].

Grant agreements, declarations and contracts are also part of this transition to an Open Science Program. This translates itself by the fact that all reviewers have to sign a declaration confirming their interest for their mission as reviewers, their engagement for the reviewing process at the different stages of the project while being very transparent regarding any potential conflict of interests. The President of the reviewers' pool has to sign a contract with the President of the Open Science Delegation in order to confirm the strong engagement and involvement with the Open Science program and its Open Science Delegation. In addition, leading house institutions from selected proposals must now sign a contract to express their strong commitment to meet the objectives of their co-funded project. This legal dimension will very likely bring a new kind of collaboration and engagement on the level of the applicants, their institutions, the Reviewers pool and its new President. The future will tell us how this transition has been beneficial and which aspects can be optimized in the following years.

III. CONCLUSION

This article has presented the major changes that took place in the transition from the Scientific Information program to the Open Science program from swissuniversities. Over time the program evolved into a more diverse, transparent and international evaluation process. What is at stake behind this transition is the reflection of the diverse and international Swiss academic landscape as a whole, which aspires to put into practice the principles of Open Science. Since this transition is currently being put into practice, the lessons learned from the previous experts' pool, the newly formed reviewers' pool and the community will continue to be taken into consideration in order to continue to optimize the evaluation process of this Open Science funding program.

[3] https://www.swissuniversities.ch/en/topics/digitalisation/open-science/oa-call-for-projects

[4] https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/Open_Science/Newsletter/211_Result_of_the_call_deadline_15.01.21.pdf

[5] https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/Open_Access/OS_Program_Reviewers_16122020.pdf

[6] https://www.swissuniversities.ch/en/topics/digitalisation/open-science/about-us

OLOS.swiss

Pierre-Yves Burgi
Division of Information Technologies
University of Geneva
Geneva, Switzerland
ORCID 0000-0002-4956-9279

Hugues Cazeaux
Division of Information Technologies
University of Geneva
Geneva, Switzerland
ORCID 0000-0002-5618-2670

Andréé Jelicic
Division of Information Technologies
University of Geneva
Geneva, Switzerland
ORCID 0000-0002-6687-1373

Abstract*— *During the P-5 program (2019-2020), the DLCM1 services transitioned from pilot to operational status: a data management training program and ad-hoc support were delivered throughout Switzerland, and OLOS (olos.swiss), the long-term solution for research data archiving and publication, became operational. Yareta, powered by the same DLCM technology than OLOS, was launched in June 2019, serving all the Higher Education Institutions of the Geneva Canton. Thanks to OLOS, the next step is to launch in January 2021 an equivalent service at the National level.

Keywords—research data, archiving, OAIS, architecture, digital preservation.

I. INTRODUCTION

« OLOS » is the name[7] given to the long-term solution for research data archiving and publication, intended to be deployed at Swiss level. OLOS, which is issued from the Swiss DLCM project[8] (Burgi, Blumer, & Makhlouf-Shabou, 2017; Burgi & Blumer, 2018), differentiates itself from other FAIR repositories in that it minimizes the constraints for its customers. Its architecture is highly flexible making it suitable to any kind of research environment. These features are detailed in Section 2. Section 3 presents the OLOS organization with information on how to adhere to it. Final conclusions in Section 4 provide some perspectives on OLOS' future developments.

II. CONCEPTS

A. Architecture

OLOS's architecture is open, modular and scalable. It can integrate to any active research data management (ARDM) solution, adapt to any metadata scheme, and type of storage such as tape, SSD, file systems, and object storage (Burgi, Cazeaux, & Echernier, 2019). The main competitive advantage of OLOS thus comes from its modular and distributed architecture, and its strict compliance to the ISO 14721 (2012) OAIS reference model (Fig. 1). To the three standard packages: submission information package (SIP), archival information package (AIP), and dissemination information package (DIP), we added a pre-ingestion module to facilitate data transition between ARDM and archiving. To further, allow easy interconnections with any research environment, the whole architecture is based on various open and international standards, such as REST for web services, DataCite metadata schema for default metadata, OAI-PMH for exchanging metadata, DOI for sustainable references, and ORCID for unambiguous authorships. The user interface is developed in Angular and has been subjected to user experience design to be more efficient and enhance its, which means entire or whole. intuitive use of the numerous offered functionalities.

[7] OLOS' name takes its origin from the Greek word "Holos", which means entire or whole.
[8] The DLCM project was mandated by swissuniversities.

Fig. 1. Architecture of OLOS

The core of OLOS (a.k.a. "backend") is written in JAVA and relies also on several opensource modules such as FITS for automatic format identification, ClamAV for virus checking, S3 for object storage, Elasticsearch as the search engine (based on the Lucene search engine library), and Shibboleth for the authentication, among others (Fig. 1). The whole architecture can be deployed either on premise, fully in the cloud, or a mix of these two modalities; for instance, the SIP and DIP could be deployed in the cloud, while the AIP would be installed on premise to confer more control on long-term preservation to the host institution. Yareta (yareta.unige.ch), powered by the same DLCM technology than OLOS, was launched in June 2019 and successfully operated since then by the University of Geneva to serve all the Higher Education Institutions of the Geneva Canton (Burgi, 2019), which provided a basis for a better understanding of the unique context of each individual research (Bezzi, 2020).

B. Institutional Benefits

OLOS is agnostic to the hardware infrastructure. Provided by default on Software as a Service (SaaS) mode as a generic repository suited for research data of any discipline, client institutions can use it without any prior investment. The portal is natively connected to existing storage infrastructures in Switzerland like SWITCHengines, with the aim to create economies of scale at the national level in order to lower the overall research data preservation costs. In some cases (i.e., very large datasets, sensitive information, excess of storage capacity, etc.), a client institution may want to connect its own storage infrastructure to OLOS. Beside the benefits that local data storage may procure, such integration to a compliant preservation service like OLOS would entitle these institutions to recover, through grants, the costs incurred for storing one copy of the archived research datasets for the entire preservation period (usually 15-20 years). Furthermore, OLOS differentiates itself from other FAIR repositories by allowing each research institution to define, implement and monitor its own preservation policy regarding, for instance, the number of copies, the duration of preservation, the copyright licenses, an eventual validation workflow, etc. in accordance to its infrastructure, and its specific institutional characteristics and constraints. A dashboard allows the monitoring of the different phases of the dataset archiving process, and provides key indicators for higher-level data management and research impact assessment.

C. Key Features For Researchers

The pre-ingest module provides high flexibility in data management by offering researchers the possibility to manipulate the datasets before final submission. Pre-ingestion thus comes after the ARDM phase, which usually involves intensive data manipulations, but comes before the archiving phase, which prevents any further modifications. Any postarchiving data modification would imply either a new archived dataset, or when permitted, a new version of the original dataset. In OLOS, versioning is not possible, but can be substituted by the use of collections, which could regroup several successive versions of the original dataset. Another key feature akin to the OLOS' modular architecture is the possibility to access all functionalities (deposit, download, search, etc.) from any environments able to activate web services. For instance, Jupyter Notebook connectors allow to search and retrieve data from OLOS, to subsequently process them based on a variety of languages and libraries. For large data volumes (over 100 GB) or high number of files (over hundred), we provide assistance to the researchers so that ingestion is automatized through batches making use of web services. Fig. 2 illustrates such a process involving large volumes.

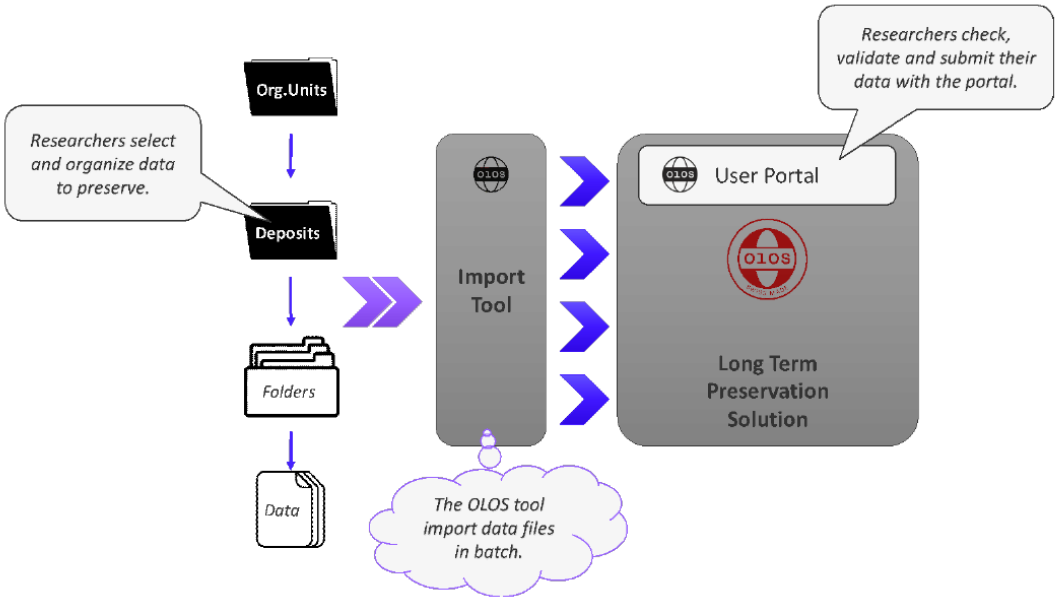


Fig. 2. Ingest process for large data volumes

An additional key feature stems from a concept very specific to OLOS: the organizational units. Datasets are organized within units whose granularity can be set at the project, laboratory, department, or institutional level. Such an organization can be a powerful instrument to monitor key indicators (see previous subsection B), and is also convenient to logically structure a lot of datasets. Finally, predefined roles (Fig. 3) provide the possibility to define different user groups, for instance giving the rights to co-authors to edit the dataset while restricting to viewing only for a specific range of users (e.g., visitors). The roles also make possible to setup a quality check, performed either by managers, stewards, or approvers through a workflow. The activation of such a workflow remains optional, and would not make sense if the institution/department/laboratory has no data quality strategy.

III. ORGANIZATION

OLOS is a non-profit association with headquarters in Switzerland, governed by institutional members of the research community. The OLOS association relies on several streams of revenue to avoid a long term dependance on public subsidies. The first one is a preservation fee charged on a per project basis to researchers. Its amount depends on the volume of data associated to preserve for the project, the number of copies (2 is recommended at minimum) and the preservation duration. Default preservation plans are 5, 10, and 15 years. If requested, a special quote can be issued to preserve the dataset for-ever. This is made possible thanks to a close collaboration with industry experts monitoring the developments of storage technologies.

Fig. 3. Predefined roles available in OLOS

Annual memberships are another pillar of OLOS financial sustainability. Several membership categories (i.e., bronze, silver or gold) are available to institutions to better suite their needs. Depending on its category, a member institution can, for instance, influence future developments of the portal, suggest further integrations to ARDM solutions or other tools used by its researchers, vote at the general assembly, or ensure a higher level of support for its researchers. To ensure high storage reliability OLOS has partnerships with trusted infrastructures and geographically distant providers, in Switzerland for more security. The integration of storage infrastructures abroad is planned to better suite the needs of international research projects.

IV. CONCLUSION & PERSPECTIVES

OLOS conception represents one of the main outcomes of the DLCM project, which benefited from the expertise of many librarians and IT professionals in the field of data management (Burgi, Blumer, & Makhlouf-Shabou, 2017). Since January 2019 (phase 2 of the DLCM project), we transitioned from a prototype to a functional long-term preservation service, whose technology has already proven itself at cantonal level with the instance called Yareta. OLOS is thus the logical step to extend the offer at National level and is due to operate starting in January 2021.

More than just a tool, OLOS is a service intended to help researchers better manage and organize their data from within their research environments. Next on the roadmap is the Fig. 3. Predefined roles available in OLOS Fig. 2. Ingest process for large data volumes development of the preservation planning module and new dashboard functionalities to provide institutions with fuller control on their assets.

ACKNOWLEDGMENT

OLOS was conceived and developed within the swissuniversities P-5 programme. We are grateful to all DLCM partners who contributed directly or indirectly to this product.

REFERENCES

Bezzi, M. (2020). Préservation des données de recherche : proposer des services de soutien aux chercheurs du site Uni Arve de l'Université de Genève. Mémoire de master : Haute école de gestion de Genève.

Burgi, P.-Y., Blumer, E., & Makhlouf-Shabou, B. (2017) Research data management in Switzerland: National efforts to guarantee the sustainability of research outputs. *IFLA Journal 43*. doi: 10.1177/0340035216678238

Burgi, P.-Y. & Blumer, E. (2018). Le projet DLCM : gestion du cycle de vie des données de recherche en Suisse. In A. Keller & S. Uhl (Eds.), *Bibliotheken der Schweiz: Innovation durch Kooperation. Festschrift für*

Susanna Bliggenstorfer anlässlich ihres Rücktrittes als Direktorin der Zentralbibliothek Zürich (pp. 235-249). Berlin : De Gruyter. doi: 10.1515/9783110553796

Burgi, P.-Y, Cazeaux, H., & Echernier, L. (2019) A versatile solution for long-term preservation of research data: Data Life-Cycle Management: the Swiss Way. In: *iPRES - 16th International Conference on Digital Preservation*. Amsterdam (The Netherlands).

Burgi, P.-Y. (2019) Le Projet de Loi 12146 : Infrastructures et services numériques pour la recherche. *Revue électronique suisse de science de l'information, 20*. http://www.ressi.ch/num20/article_168

ISO 14721:2012 (2012). Space data and information transfer systems - Open archival information system (OAIS) – Reference model

Managing the life cycle of a portfolio of Open Data Resources

Chiara Gabella PhD
Swiss Institute of Bioinformatics
Lausanne, Switzerland
https://orcid.org/0000-0002-7104-5025

Séverine Duvaud
Swiss Institute of Bioinformatics
Lausanne, Switzerland
https://orcid.org/0000-0001-7892-9678

Christine Durinx PhD
Swiss Institute of Bioinformatics
Lausanne, Switzerland
https://orcid.org/0000-0003-4237-8899

***Abstract*—*Data resources are essential for the long-term preservation of scientific data and the reproducibility of science. The SIB Swiss Institute of Bioinformatics provides the life science community with a stable and reliable data infrastructure. The Institute provides a portfolio of high-quality databases and software platforms, which vary from expert curated knowledgebases such as UniProtKB/Swiss-Prot (part of the UniProt consortium), neXtProt, STRING and Bgee, to online tools such as SWISS-MODEL and SwissDrugDesign. SIB aims to ensure that these SIB Resources are available on the long term, i.e. as long as their scientific return-on-investment and impact are high. SIB's vision is that data and research results should be freely accessible to all. For this reason, the Institute promotes the adoption of open licenses. This paper describes the processes that support the identification, evaluation, and development of SIB Resources. For over ten years, SIB employs a set of indicators that reflect the multiple facets and complexity of data resources. These indicators set quality standards, and monitor usage trends and scientific impact. They guide and inform both the SIB Scientific Advisory Board in its evaluation, and the managers of the databases and tools in the development of their respective resources towards scientific excellence and Open Science. Through a professional management framework with central services such as user-centric design and a license advisory committee, SIB supports the promotion of excellence in resource development and operation.*

Keywords*—*Bioinformatics, Open Data, Database, Software tools, Infrastructure, User Experience (UX), High-quality integrated resources, Sustainability.

I. Introduction

The SIB Swiss Institute of Bioinformatics (www.sib.swiss) is an internationally recognized non-profit organization, dedicated to biological and biomedical data science. It is present in the main academic institutions in Switzerland (Fig. 1) and leads numerous national and international projects with a major impact on life science research and health.

SIB's scientists are passionate about creating knowledge and converting complex questions into solutions in many fields, from biodiversity and evolution to medicine. They provide essential resources, such as databases and software platforms, as well as data management, software engineering, biocuration services, computational biology know-how and training in bioinformatics. The Institute delivers this expertise to academic groups and clinicians as well as to private companies.

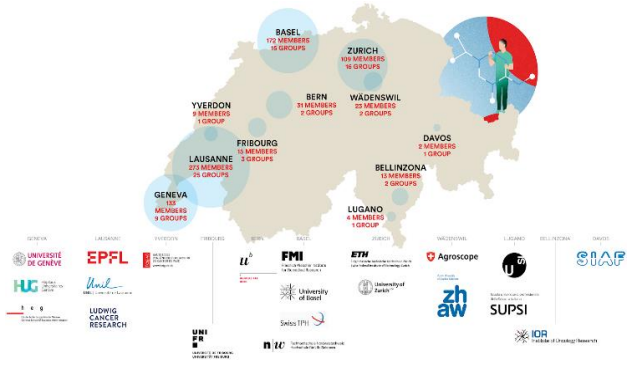


Fig. 1: Map of the SIB partner institutions as of December 2020. In January 2021, EMPA (the Swiss Federal Laboratories for Materials Science and Technology) in St. Gallen and the University Hospital of Zurich became new institutional members of SIB.

SIB fosters collaboration and knowledge sharing among some 800 scientists across Switzerland and represents the Swiss bioinformatics community, collaborating with international institutions on global research infrastructures. The Institute contributes to keeping Switzerland at the forefront of innovation by promoting progress in biological research and enhancing health.

A study by Attwood et al. looked at the 18-year survival of 326 publicly available biological databases (Attwood et al., 2015). Over 60% "died" within that time period, and a further 14% were archived, i.e. no longer updated. The instability of the existence of major data resources is associated with the risk of losing an immense wealth of biological information, and the associated investments. The study shows that a viable, sustainable framework for long-term data stewardship is sorely needed. Indeed, the databases that survived were, for the most part, important to their institution's main focus, and had core institutional support. Database longevity depends on the existence of infrastructures that are underpinned by long-

term strategies. Moreover, although data resources play an essential role in scientific research, a sustainable funding model that ensures their maintenance and development remains a critical challenge (Gabella et al., 2018). Within the limit of available funding, SIB's commitment is to ensure the long-term existence of the SIB Resources in order to provide a stable environment for the development, enhancement and maintenance of high-quality databases and software tools to support the life science community.

II. THE LIFE CYCLE OF A BIOINFORMATICS RESOURCE

A new database or software tool typically starts with a research project, leading to a proof-of-concept. Through further development, the resource evolves towards maturity and may become part of the research infrastructure of the scientific community (Fig. 2).

The SIB Groups develop and maintain numerous cutting-edge resources that are made available to the community. Among them, several key resources - the SIB Resources - benefit from the Institute's specific support after a careful selection process and are included in the SIB portfolio.

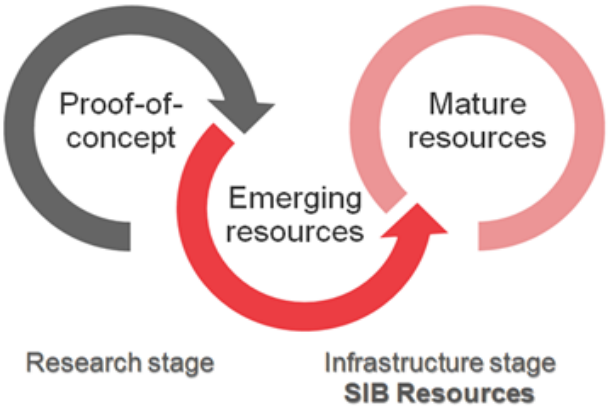


Fig. 2: The life cycle of a bioinformatics resource.

SIB's resource portfolio contains promising emerging resources, such as SwissLipids (Aimo et al., 2015) or V-pipe (Posada-Céspedes et al., 2020), as well as fully mature ones. To promote new developments at the highest level of scientific excellence, SIB is committed to identifying and supporting additions to its resource portfolio among the emerging or already well-established resources developed by SIB groups, and which have a (potentially) high impact on the life science community. These resources must have passed the proof-of-concept stage and reached a sufficient level of development and usage to be considered as infrastructure. They should demonstrate their uniqueness, a strong demand from the target community and alignment with Open Data practices, in addition to high scientific standards.

SIB's high-quality infrastructure, i.e. databases and software tools, as well as the associated services, are driven by excellence in research. SIB Resources are typically linked to research work and often embedded in a research group, in order to remain state-of-the-art (SIB Swiss Institute of Bioinformatics Members, 2016).

Among the SIB Resources we count the UniProtKB/Swiss-Prot database, which is part of the UniProt Consortium (Poux et al., 2017; The UniProt Consortium, 2019). The knowledgebase contains a reviewed collection of high-quality annotated and non-redundant protein sequences, bringing together experimental results, computed features and scientific conclusions to provide information related to a protein's function, structure and subcellular location, specific features and interactions. The UniProtKB/Swiss-Prot database contains over half a million expert-curated protein sequences. A team of highly qualified scientists —called biocurators — who are based in Geneva, select, review and annotate the information. The biocurators are supported by advanced machine learning models that automatically identify and classify relevant publications for review (Lee et al., 2018). With a million unique users per month, UniProt is the most widely used protein information resource in the world.

Another example of a SIB Resource is the SWISS-MODEL Workspace (Waterhouse et al., 2018). This fully automated web-based service assists and guides the user in building a three-dimensional structure of a protein, based on its homology with proteins for which experimentally determined structures are available. SWISS-MODEL receives over a million model requests every year.

STRING (Szklarczyk et al., 2019), another SIB Resource, is a knowledgebase and software tool for known and predicted protein-protein interactions. It includes direct (physical) and indirect (functional) associations derived from various sources, such as genomic context, high-throughput experiments, (conserved) co-expression and the literature. STRING networks cover over 5,000 different organisms with over 25 million high-confidence links between proteins.

The full list of SIB Resources is available in appendix A.

III. A COMPREHENSIVE SET OF INDICATORS AS GUIDE ALONG A RESOURCE'S LIFE CYCLE

SIB is bound to the State Secretariat for Education, Research and Innovation SERI through a Service Level Agreement (SLA), which covers the funding to SIB for the provision of bioinformatics resources to the life science

community. Every four years, the SIB Board of Directors (BoD) is responsible for selecting best-in-class resources (i.e. SIB Resources), as well as for the allocation of the federal funds to said resources[9]. Decisions are based on the recommendations of the Institute's external Scientific Advisory Board (SAB) that, in some cases, also takes input from evaluations by external reviewers. Members of the BoD do not take part in decisions for which they have a conflict of interest.

Three criteria define a best-in-class resource at SIB: scientific impact, scientific return-on-investment, and its fit within the resource portfolio. Scientific impact is defined as a combination of the scientific state-of-the-art, utility, and use. The (expected) scientific ROI of funding is estimated, i.e. the impact in terms of serving more users, of filling an important unmet need of the scientific community, and of accelerating science. And last but not least, careful attention goes also to the alignment of SIB's portfolio of databases and software tools with the Institute's core competencies and strategic focus.

The SAB assesses whether best-in-class criteria are met by using a set of 28 indicators that are grouped in six categories (see below). The responsible resource is required to submit a workplan based on these indicators (a template of the workplan for candidate SIB Resources, with the full list of indicators, is available in Table I. For more details see Gabella et al. (2020)).

Databases and software tools being very diverse, it is therefore crucial to take into account their many different facets into the indicators for evaluation. Providing precise information and figures allows then the panel of external reviewers and the SAB to make an objective recommendation.

These indicators can also be helpful for the scientists developing a resource to guide the development process. Indicators are assessed differently depending on the type of resource: it is the whole body of indicators together that reflects the quality and impact of a specific bioinformatics resource.

Indicators for SIB Resources are grouped into the following six categories:

Category I, Scientific focus and quality of science: Demonstrate high quality of data and metadata, respond to a clear scientific need, and be unique. This implies benchmarking against other resources, and being an authority in its field compared to the major competitors.

TABLE I. INDICATORS FOR A CANDIDATE SIB RESOURCE. MORE INFORMATION ARE AVAILABLE IN Gabella et al. (2020)

I: Scientific focus and quality of science		
a	Definition	The Resource is a ☐ Database ☐ Software tool ☐ Database & Software tool
b	Scope statement	Describe the scope and scientific coverage of the Resource (for example, all species or a subset of species, families, outputs from a particular experimental method), as well as the scientific need to which it responds.
c	Uniqueness of the resource	Indicate who are the major competitors of the Resource and how the Resource compares to them including benchmarking efforts.
d	Potential synergies and collaborations with other SIB Resources	Describe the potential synergies and collaborations with other SIB Resources that could increase the impact of the (respective) Resource(s).
e	Objectives and implementation plan for the next funding period	Describe the objectives and implementation plan for the next funding period.
II: Community		
a	User community	Whom is the Resource addressed to? Describe the current user community and the size of the potential user community. Are there other user communities that are currently not yet reached? Include quantitative measurements, if possible.
b	Overall usage - web access	Web access as measured by Google Analytics. Please join the following extract from Google Analytics (period: from 2018 to present, or from more recently if not yet available in 2018; view by month): - Audience > overview - Acquisition > overview

[9] Article 7.2 of the SIB Statutes

<table>
<tr><td></td><td></td><td>- Behaviour > overview</td></tr>
<tr><td></td><td>Overall usage - additional access methods</td><td>Please give figures to quantify any additional access methods: visits, unique visitors, hits, and downloads (includes FTP downloads and programmatic access)</td></tr>
<tr><td>c</td><td>Overall usage (software tools / service platforms)</td><td>If available, provide statistics about the number of jobs submitted and/or the number of downloads of the software tool and/or any other measurements that assess the usage of the Resource.

Example:
<table>
<tr><td>Type of measurement</td><td>Value at 31/12/2017</td><td>Value at 31/12/2018</td><td>Value at 31/12/2019</td><td>Value at 31/07/2020</td></tr>
<tr><td>Jobs submitted</td><td>100</td><td>200</td><td>400</td><td>500</td></tr>
</table></td></tr>
<tr><td>d</td><td>Usage in research as measured through citation in the literature</td><td>Go to https://europepmc.org/ and run a query using the relevant keywords, which best identify the Resource. Copy and paste the resulting URL, together with the number of citations (example: https://europepmc.org/search?query=resourceAAB%20%2B%20organism: 35 citations). If necessary, indicate more than one URL, but do not exceed 3.
If any extra explanation is needed, you can also add it to the text field. Moreover, if other citation counting tools are more representative of the usage of the Resource, please feel free to report the additional statistics.</td></tr>
<tr><td>e</td><td>Dependency of other resources</td><td>Indicate which resources depend on the Resource.

Examples:
● The UniLectin platform depends on SWISS-MODEL to display molecular interactions between small molecules and protein structures.
● PeptideAtlas uses information about single amino acid variants and post-translational modifications in neXtProt to enable processing and display of data</td></tr>
<tr><td colspan="3">III: Quality of Service</td></tr>
<tr><td>a</td><td>Unique ID</td><td>Indicate if the Resource provides persistent and unique identifiers and if yes, describe it.</td></tr>
<tr><td>b</td><td>Data entries/records (data resources)</td><td>Cumulative total number of entries or records to indicate the growth of the Resource. Please provide a description and explanation on what an “entry” is.

Example:
<table>
<tr><td>Entry</td><td>Description</td><td>Value at 31/121/2017</td><td>Value at 31/12/2018</td><td>Value at 31/12/2019</td><td>Value at 31/07/2020</td></tr>
<tr><td>Species</td><td>Number of Species in the Resource</td><td>100</td><td>200</td><td>300</td><td>400</td></tr>
<tr><td>Genes</td><td>Number of genes annotated in the Resource</td><td>1000</td><td>2000</td><td>3000</td><td>4000</td></tr>
<tr><td>Variants</td><td>Number of sequence variations present in the Resource</td><td>200</td><td>250</td><td>400</td><td>410</td></tr>
</table></td></tr>
<tr><td>c</td><td>Use of community-recognized standards for (meta)data</td><td>Which community-recognized standards are used for metadata and data (e.g. MIAME, JATS, INSDC features, ontologies)? Provide a link to documentation.</td></tr>
<tr><td>d</td><td>Data availability & access (data resources)</td><td>Data sharing services: list services through which data is shared (e.g. website, APIs, FTP, TripleStore)

Data sharing formats: list formats for available data (e.g. plain text, FASTA, XML, RDF, Dublin Core, tsv, JSON)</td></tr>
<tr><td>e</td><td>Customer service</td><td>Helpdesk: does the Resource run a helpdesk?

User feedback: how does the Resource seek and incorporate user input into service design decisions?

User training: which training is offered by the Resource? (face2face training, eLearning, etc.)</td></tr>
</table>

VI: Legal and funding infrastructure		
a	Scientific advisory board	Does the Resource have an external independent Advisory Committee with scientists and/or users (other than the SIB SAB)?
b	Legal framework supporting Open Science?	Does the Resource have terms of use or a licence that enables the reuse and remixing of data? (see Open Definition for a list of open licenses) If yes, please include a link to terms of use or state license designation (e.g. CC0, CC-BY, CC-BY-SA). Is the access to, and/or usage of, the Resource restricted in any way to the user or certain categories of users? If yes, please explain why. If the current legal framework does not support Open Science, do you plan to adopt an open licence?
c	Licensing code: is it published on open access?	Is the source code of the Resource published in open access (e.g. in GitHub)? If yes, please indicate under which licence and the relevant URLs.
d	Sustainable support and funding	Describe what has been, and what will be, undertaken to seek additional funds from other sources.
e	Estimation of funds	Please indicate an estimation of funds that will be needed from SIB (in CHF) for the next funding period (4 years). Detail (in person-months) for what kind of work SIB funding will be used (for example: XX person-months for biocuration, YY person-months for code development, ZZ person-months for dataset selection):
V: Impact and visibility		
a	Counterfactual	Description of how science would be affected if the Resource had not existed or was to disappear and not be replaced.
b	Accelerating science	Description of how much the Resource accelerates science: present one or more selected examples of how the Resource has been used by its user community, showcasing the importance of the Resource for advancing science. These "translational stories" help SIB to show its impact to audiences such as policy makers and funders and help the SAB to assess the Resource.
c	Visibility	What actions has the Resource taken to increase its visibility within its potential user community?
VI: SIB		
a	Benefits	How does your Resource contribute to SIB in terms of scientific credibility, visibility or any other aspect?
b	Participation	Describe your contributions to SIB in the last 2 years, and how you see your involvement in the coming years
c	SIB Portfolio of Resources	How does the Resource align with the current portfolio of SIB Resources?

Category II, Community: Know the community to whom it is addressed, its size and the usage: web statistics, user reach, and community size. Candidate resources with a valid track record of usage, responding to a clear need within the scientific community are more likely to be included as SIB Resources. However, emerging resources are encouraged to submit as well. The scientific context in which the resource operates should be taken into account. A resource that serves a small scientific community may not have as many users as a resource serving a broader interest, and yet it may reach 90% of the community it supports (coverage) and be crucial for the scientific work of that community.

Category III, Quality of the service: Demonstrate a high level of service and reliability with the integration of features such as persistent and unique identifiers, community-recognized standards, user support and training, as well as the integration of user feedback.

Category IV, Legal and funding infrastructure and governance: Have a sound legal framework supporting Open Science and seek complementary funds from other sources in order to ensure sustainable long-term funding.

Category V, Impact and translational stories: Have a significant impact on the life science community and be impact driven.

Category VI, SIB: Group Leaders who manage a SIB Resource are expected to show strong involvement in SIB. Therefore, this aspect is also taken into consideration in the evaluation.

For each category and for each of the new candidate resources, the external reviewers assign a score according to a 9-point scale, going from Poor (1) to Exceptional (9), supplemented by a short commentary. At the SAB meeting, both the existing and the short-listed candidate SIB Resources are evaluated, based on these categories. The SAB scores their workplans and makes recommendations regarding their funding level.

SIB's support to the identified best-in-class Resources takes two forms. On the one hand, SIB offers a portfolio of services that support professional infrastructure provision, including User Experience (UX) studies and design, hosting, best practices and knowledge sharing, a security audit, user training, as well as licensing and legal advice. On the other hand, SIB provides funds that allow hiring skilled personnel to develop and maintain the resource. The mission of SIB is to maintain a portfolio of SIB Resources over the long-term and fund the SIB Resources within this portfolio as long as they have a high impact in the life science community. It is the role of the SAB to evaluate this impact and give recommendations to the Board of Directors regarding the level of funding.

In principle, SIB commits to supporting resources along the entire funding period of 4 years. However, the SIB Resources are also evaluated by the SAB at mid-term. This mid-term review includes an evaluation of the progress made since the beginning of the period. Based on the outcome, the BoD can decide to adjust the funding level or stop funding the resource.

Conclusion

Thanks to its coherent portfolio, including both emerging and already well-established resources, SIB is a key driver of innovation in bioinformatics. The indicators developed for the evaluation and selection process, through continuous monitoring of usage trends and scientific impact of the Resources, inform their life-cycle management by providing strategic recommendations for mature resources and allowing promising resources to develop to their full potential.

The provision of a professional solid infrastructure ranging from user-centred design, user research, licensing consulting and funding, enables the SIB's resource portfolio to be at the forefront of scientific excellence and ensures its long term sustainability in a context of Open Science.

Acknowledgment

We thank the SIB Scientific Advisory Board for their precious and expert advices and the SIB Board of Directors for their commitment within SIB in taking all the decisions necessary to achieve the aims of the Institute. Overall, we want to thank all the SIB Group Leaders and the SIB Resources managers for their commitment in making a better science.

References

Aimo, L., Liechti, R., Hyka-Nouspikel, N., Niknejad, A., Gleizes, A., Götz, L., Kuznetsov, D., David, F. P. A., van der Goot, F. G., Riezman, H., Bougueleret, L., Xenarios, I., & Bridge, A. (2015). The SwissLipids knowledgebase for lipid biology. *Bioinformatics*, 31(17), 2860–2866. https://doi.org/10.1093/bioinformatics/btv285

Attwood, T. K., Agit, B., & Ellis, L. B. M. (2015). Longevity of Biological Databases. *EMBnet.Journal*. https://doi.org/10.14806/ej.21.0.803

Gabella, C., Durinx, C., & Appel, R. (2018). Funding knowledgebases: Towards a sustainable funding model for the UniProt use case. *F1000Research*. https://doi.org/10.12688/f1000research.12989.2

Gabella, C., Durinx, C., & Appel, R. (2020). Selection of SIB Resources for the period 2021-2025. *Zenodo*. https://doi.org/10.5281/zenodo.3937334

Lee, K., Famiglietti, M. L., McMahon, A., Wei, C.-H., MacArthur, J. A. L., Poux, S., Breuza, L., Bridge, A., Cunningham, F., Xenarios, I., & Lu, Z. (2018). Scaling up data curation using deep learning: An application to

literature triage in genomic variation resources. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1006390

Posada-Céspedes, S., Seifert, D., Topolsky, I., Metzner, K., & Beerenwinkel, N. (2020). V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput sequencing data. *BioRxiv*, 2020.06.09.142919. https://doi.org/10.1101/2020.06.09.142919

Poux, S., Arighi, C. N., Magrane, M., Bateman, A., Wei, C.-H., Lu, Z., Boutet, E., Bye-A-Jee, H., Famiglietti, M. L., Roechert, B., & UniProt Consortium, T. (2017). On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics*, 33(21), 3454–3460. https://doi.org/10.1093/bioinformatics/btx439

SIB Swiss Institute of Bioinformatics Members. (2016). The SIB Swiss Institute of bioinformatics' resources: Focus on curated databases. *Nucleic Acids Research*, 44(D1), D27–D37. https://doi.org/10.1093/nar/gkv1310

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Von Mering, C. (2019). STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research, 47*(D1), D607–D613. https://doi.org/10.1093/nar/gky1131

The UniProt Consortium. (2019). UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Research, 47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., De Beer, T. A. P., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: Homology modelling of protein structures and complexes. *Nucleic Acids Research, 46*(W1), W296–W303. https://doi.org/10.1093/nar/gky427

V. Appendix: The current SIB Resource Portfolio

Name of the SIB Resource		***Type***	***Description***	***Highlights***
Bgee	Gene expression expertise	Knowledgebase with expert curation and software tool	Gene expression data (including all types of transcriptomes), allowing retrieval and comparison of expression patterns between animals, humans, model organisms and diverse species of evolutionary or agronomical relevance.	Only resource to provide homologous gene expression between species.
EPD	Eukaryotic Promoter Database	Knowledgebase with expert curation and software tools	Quality-controlled information on experimentally defined promoters of higher organisms, as well as web-based tools for promoter analysis.	Over 180,000 promoters download- able, analysable over a web interface and viewable in the UCSC genome browser.
neXtProt	Human protein knowledgebase	Knowledgebase with expert curation and associated tools	Information on human proteins such as function, involvement in diseases, mRNA/protein expression, protein/protein interactions, post-translational modifications, protein variations and their phenotypic effects.	High data coverage through integration of multiple sources. Advanced semantic search functionalities. Tools specifically designed for the proteomics community.
STRING	Protein-protein Interaction Networks and Functional Enrichment Analysis	Knowledgebase and software tool	Resource for known and predicted protein-protein interactions, including direct (physical) and indirect (functional) associations derived from various sources, such as genomic context, high-throughput experiments, (conserved) co-expression and the literature.	An ELIXIR Core Data Resource. STRING networks cover over 5,000 different organisms with over 25 million high-confidence links between proteins

SwissDrugDesign	Widening access to computer-aided drug design	Software tools	Web-based computer-aided drug design tools, from molecular docking (SwissDock) to pharmacokinetics and druglikeness (SwissADME), through virtual screening (SwissSimilarity), lead optimization (SwissBioisostere) and target prediction of small molecules (SwissTargetPrediction).	Comprehensive and integrated web-based drug design environment
SWISS-MODEL	Protein structure homology-modelling	Software tool and repository	Automated protein structure homology-modelling platform for generating 3D models of a protein using a comparative approach, and database of annotated models for key reference proteomes based on UniProtKB.	Easy-to-use web-based platform processing over two million model requests per year, providing model information for experts and non-specialists.
SwissOrthology (OMA + OrthoDB)	One-stop shop for orthologs	Phylogenomic databases and software tools	Web portal of resources to infer orthologs, i.e. corresponding genes across different species, a key aspect to predicting gene function or reconstructing species trees. It includes OrthoDB, BUSCO as well as OMA and the Quest for Orthologs benchmark service.	World-leading orthology and comparative genomic resources.
SwissRegulon	Tools and data for regulatory genomics	Software tools and knowledgebases	Web portal for regulatory genomics, including genome-wide annotations of regulatory sites and motifs, the webserver ISMARA for automated inference of regulatory networks and CRUNCH for automated analysis of ChIP-seq data, and REALPHY for reconstructing phylogenies from raw sequence data.	ISMARA and Crunch web servers allow users to upload raw microarray, RNA-seq or ChIP-seq data to automatically infer the core regulatory networks acting in their system of interest.
UniProtKB/SwissProt	Protein knowledgebase	Knowledgebase with expert curation	Hundreds of thousands of protein descriptions, including function, domain structure, subcellular location, post-translational modifications and functionally characterized variants.	Expert-curated part of UniProt, the most widely used protein information resource in the world, with over six million pageviews per month. An ELIXIR Core Data Resource.
SwissLipids	A knowledge resource for lipids	Knowledgebase	Information about known lipids, including knowledge of lipid structures, metabolism and interactions, providing a framework for the integration of lipid and lipidomics data with biological knowledge and models.	Contains information on more than 590,000 lipid structures from over 640 lipid classes.
V-pipe	Viral genomics pipeline	Software tool	Pipeline integrating various open-source software packages for assessing viral genetic diversity from next-generation sequencing data.	Enabling reliable and comparable viral genomics and epidemiological studies and facilitating clinical diagnostics of viruses.

NCCR Robotics Research Data Management Strategy: A Workflow Application

Valeria Di Cola
NCCR Robotics - EPFL
Lausanne, Switzerland
valeria.dicola@epfl.ch
0000-0003-2925-8366

Abstract— *In compliance with the SNSF requirements, a Research Data Management strategy has been developed for NCCR Robotics in synergy with the EPFL Library, and the EPFL Legal and Ethics Departments. The strategy includes guidelines and a workflow for making research data related to NCCR Robotics journal publications and proceedings publicly available, including the special cases of sensitive and protected data.*
The article focuses on the coaching program that was established to support NCCR members in the implementation of this strategy. Concrete resources and tools – such as tailored presentations answering the specific needs of each laboratory – have been developed in order to offer practical solutions, focusing on three main aspects of research data management (RDM): (1) Data lifecycle and FAIR principles; (2) A proposed practice within NCCR Robotics, with emphasis on Zenodo as a recommended data repository; and (3) General Data Management Best Practices.

Keywords—RDM Strategy, FAIR, data repository, workflow.

I. INTRODUCTION

The National Centre of Competence in Research (NCCR) Robotics is a Swiss nationwide organization funded by the Swiss National Science Foundation, bringing together more than 100 top researchers from all over the country with the objective of developing new, human-oriented robotic technology for improving quality of life. The Centre binds together experts from seven world-class research institutions; École Polytechnique Fédérale de Lausanne (EPFL) (leading house), Eidgenössische Technische Hochschule Zürich (ETH Zurich) (co-leading house), Universität Zürich (UZH), Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) Lugano, University of Bern (UNIBE), the Swiss Federal Laboratories for Materials Science and Technology (Empa) and the University of Basel (UNIBAS). In addition to standard research data collected, the research units that are part of the consortium produce sensitive personal data as well as protected data, which may include - but is not limited to - information relating to data received by a third party under confidentiality or a specific agreement; data being subject to be protected via a patent or any other intellectual property title or subject to be licensed to a third party for commercial purposes (e.g. start-up).

Studies have shown that keeping research data freely available is crucial for open science — and the research funding could depend on it (Schiermeier, 2018). In 2019 we developed a Research Data Management Strategy (RDMS) as part of our contractual obligation with SNSF in close collaboration with related offices of the home institutions (EPFL, ETHZ and UZH Libraries, EPFL Legal and Ethics Departments).

The RDMS aims at contributing to recognize research data as valuable academic resources that need to be managed, shared and preserved to foster research and science. It provides relevant tools and guidance to better manage the data that our scientists handle throughout its whole life-cycle, from the planning stage of the project to the long-term preservation strategies. Hence, contribute to making FAIR Science a reality as well as help researchers be more productive for themselves and their collaborators (ELIXIR, 2021).

The strategy includes guidelines and a workflow for making research data related to NCCR Robotics journal publication and proceedings publicly available, including the special cases of sensitive and protected data. Here, we present the workflow application of the implementation of the RDMS over the 28 Laboratories that are part of our Centre of competence.

II. General section

A. *Internal Organisation, Roles and Responsibilities*

All information and training activities for NCCR members regarding SNSF/NCCR data management policy and requirements are managed by the NCCR Robotics RDM Officer, supported by the EPFL Library.

Maintenance of the data management infrastructures, data backup and preparation, curation and documentation of datasets, submission of datasets on repositories are under the responsibility of each individual laboratory.

B. *General Strategy for Data Storage*

Each laboratory defines its own data storage and preservation strategy based on its needs and safety concerns. Support is provided by the IT services related to the laboratory, that install and support the data storage of each laboratory according to its needs.

However, these common principles must be followed:

• The data (or at least a copy) must be stored on the institutional hard-drive; cloud-based storages (such as Google Drive, Dropbox, or Switch Drive) are not considered institutional hard-drives.

• In case of sensitive data, the storage must guarantee that the data are well protected and only authorized people can have access to them.

C. Public Data Sharing

Only data for journal publications and proceedings that resulted from activities that have been funded through NCCR Robotics are concerned by this strategy.

NCCR Robotics recommends the use of Zenodo (zenodo.org) , a data repository that respects the FAIR Data Principles and is maintained by a non-profit organisation (CERN). Every publication made on Zenodo is publicly available and has its own DOI for unique identification and citation.

In case a NCCR member does not want to use Zenodo, they have the freedom to use another data repository. In this case they have to make sure the chosen data repository respects the SNSF requirements (it must respect the FAIR Data Principles and must be maintained by a non-profit organisation), and also has to communicate manually the DOI of the data publication to the NCCR Robotics RDM Officer.

The workflow application of the RDMS (with Zenodo, or with another data repository) is summarized in Fig. 1 in order to help NCCR Robotics members with the data publication. The cases of sensitive and protected data are also included in the workflow.

Within every published dataset, a README text file should be included that contains at least the following meta-data (as required by the SNSF FAIR Principle) :

- a title (a name given to the dataset or the research project that produced it);
- an abstract of the project (description);
- creator (the name of the person who collected or contributed to the data) identified by ORCID;
- the date of collection;
- a short description of each file;
- a persistent identifier (ISBN, or DOI);
- the license;
- If unique tools or proprietary software are used, this will also be documented in the metadata when appropriate. If possible, the tools or links to the provider will be made publicly available.

In the case of developed code or scripts, they are considered as data and should be also published:

- the programming language used
- the versions used (libraries, compiler, packages, etc.)
- the machine used
- the license of the code

In case the data has no license already attached and there is no claim by any third party on such data, NCCR Robotics recommends the application of the CC-BY license to the published data. The CC-BY license authorizes the access, usage (commercial or not) and the modification of the published data but requires the citation of the author by the user of the data.

III. CONCLUSION

It is clear that keeping research data available is essential for open science. Not only does it ensure transparency and reproducibility, increasing data visibility and number of citations, but fulfilment of funders' requirements. Following best research practices leads to saving time and avoiding risks of data loss. Lastly, comprehensive RDM allows our researchers not only to produce new knowledge and make more discoveries just by re-using data, but also to archive, retrieve and re-use their own data. Nowadays, we are witnessing an increase in demand and use of big data, so having an up-to-date RDMS is part of maintaining modernity with the world scale digital research.

With our RDMS we took the first steps towards a paradigm shift in providing FAIR RDM together with training and education for our consortium on this topic. The strategy was implemented virtually in 2020 during 4 months, over 28 Labs through one-to-one trainings) and reached more than 100 NCCR Robotics researchers.

A. Final recommendations

- Share any data that is relevant for re-use.
- Data underlying publications must be made available at the time of the publication.
- Wider data can be made available after the project ends.

ACKNOWLEDGMENT

This research was supported by the Swiss National Science Foundation through the National Centre of Competence in Research Robotics (NCCR Robotics).

We are thankful for the invaluable contribution of Eliane Ninfa Blumer, Antoine Masson, and Francesco Varrato from the EPFL Library for helping us develop this strategy and for the day-to-day support implementing it. We appreciate the inputs from Matthias Töwe (ETH Library), and Florian Steurer and Stefanie Strebel (UZH Library) to adapt the document to the different institutions.

REFERENCES

ELIXIR (2021) Research Data Management Kit. A deliverable from the EU-funded ELIXIR-CONVERGE project (grant agreement 871075). URL: https://rdmkit.elixir-europe.org

Schiermeier, Q. (2018). Data management made simple. *Nature*, *555* (7696).

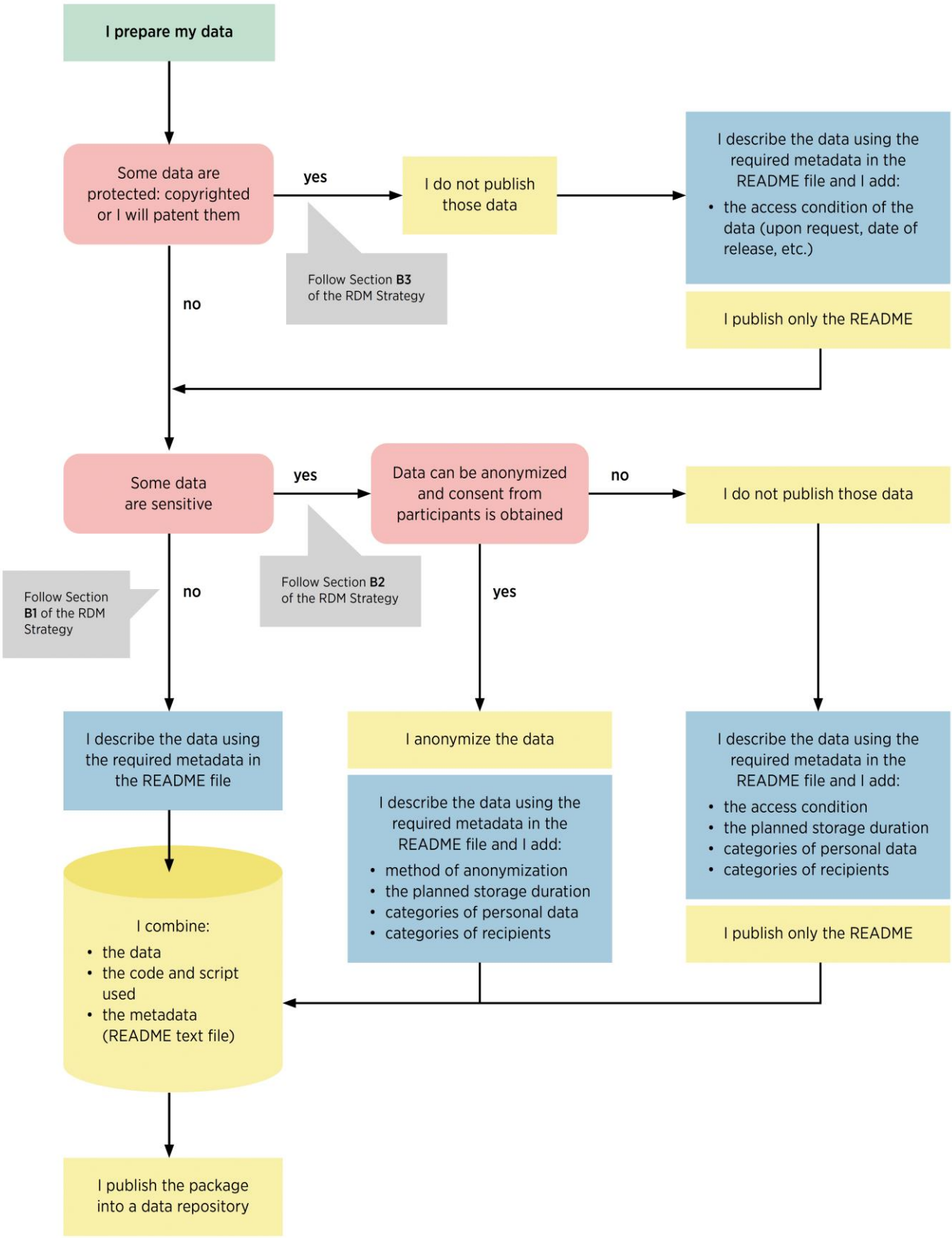


Fig. 1. Workflow application of the NCCR Robotics RDMS

Preserving Large Quantities of Data and Maintaining Digital Sovereignty

Alberto Pace
CERN, IT department
Geneva, Switzerland
alberto.pace@cern.ch

Abstract— *Corporate data are often the most valuable assets that companies need to access, analyze and therefore preserve to ensure business continuity. As technologies are evolving rapidly and the volume of data is increasing exponentially, the straightforward response to this challenge is to outsource the problem to external IT companies that can provide attractive costs and effective solutions. However, this does not come without the risk of creating uncontrolled external dependencies and vendor lock-in that turn to be irreversible and can endanger the core business itself and even threaten its survival.*
This paper presents some issues and mitigation strategies that could be adopted in designing proper solutions.

Keywords—RDM Strategy, FAIR, data repository, workflow.

I. INTRODUCTION

In recent years, the evolution of storage technology has improved beyond the best expectations. In the early '90s corporate servers had capacities up to 1 GB. Nowadays, corporate servers can exceed the one PB capacity, i.e. one million times more, representing an increase of six orders of magnitude.

All the other components of computing witnessed similar evolutions. CPU performance has increased by three orders of magnitude while the multicore, distributed computing approach has increased the number of CPU cores available to a data processing application by 2 or 3 additional orders of magnitude.

Networking has also experienced similar improvements. Again, in the early nineties, computers could be connected with networks that were reaching few Mbit/s of data transfer rates. Today, it is possible to have connections that exceed 100 GB/s, which represents a five orders of magnitude improvement.

This trend is not finished yet. Actually, we are just at the beginning and there are clear hints that this trends will continue for several years. For instance, today you can purchase a micro-SD card of 1TB capacity. If you scale the volume of a micro-SD to the volume of a 3.5" hard disk, you could expect more than a 2.8 PB capacity in the volume of a single hard drive.

II. PERFORMANCE AND RELIABILITY OF HIGH DENSITY STORAGE

Another surprising evolution on large flash capacity is that we can expect both performance and reliability to increase. This reason is that the usual techniques used in large servers to increase both performance and reliability can be embedded into a single storage device.

The most known technique to increase performance is striping , where the data are split and then read and/or written in parallel to multiple and independent storage locations. This offers the possibility of an arbitrary performance increase that can be obtained by demultiplexing a data stream to an arbitrary number of parallel streams that access different storage locations. Clearly, this technology can be embedded into the single storage media.

Similarly, there is a wide set of error correction techniques that can be used to increase reliability. This is generally obtained by dedicating a small part of the storage capacity in the media to store error correction information that can be used in case of a media read and/or write error. In this scenario, traditional errors that would in the past lead to data loss, would be automatically corrected by the redundant information and the software embedded in the media, thus significantly increasing the media reliability. In addition, these techniques allow predicting data loss with high accuracy, because the amount of redundancy left available can be used to precisely estimate the probability of data loss.

Therefore, we can expect in the coming years, the appearance of storage devices with very high capacity, very high performance in terms of both latency and throughput as well as an outstanding reliability.

III. THE CHALLENGE OF DATA PRESERVATION

The cost reduction that we have observed in storage media should ease the problem of long term data preservation. When the cost drops, it becomes more affordable to buy additional media to store additional copies of the data that needs to be preserved. Therefore, the naïve conclusion is that this evolution significantly facilitates the data preservation.

Unfortunately, this statement is only partially true, as there is another aspect to take into consideration: If storage is cheap, the amount of data that is produced increases. In this case, the challenge of data preservation becomes more difficult because the amount of data to preserve increases exponentially over time and because data need to be moved from older media to newer ones, the constraints on the surrounding architecture is constantly increasing.

Let's take the example of the networking requirements for moving hundreds of petabytes compared to what was required, few years ago, to move hundreds of terabytes. One could take the simplistic approach to address this challenge by deploying a network that is 3 orders of magnitude faster. Could this approach be effective ? Probably not: ten or twenty years ago, moving hundreds of terabytes would take a couple of months with, on average, one error generated for every terabyte copied. This was producing approximately a hundred of errors to be manually resolved in a couple of months. This was a realistic process to implement at that time.

If you would use the same architecture today, with a network 3 orders of magnitude faster to transfer 3 orders of magnitude more data, you would certainly fail to achieve it in a couple of months. Why? Because as you have the same architecture you can expect the same error rate, and therefore instead of some hundreds of errors, you will have more likely some hundred thousand errors (3 orders of magnitude more) to be manually addressed before you succeed. This highlights why the architecture must be reviewed and, in the particular example, it becomes evident that a specific development is essential to process known errors in a complete automated way, as the manual approach does not scale with the growth.

IV. PRESERVING DIGITAL SOVEREIGNTY

Digital sovereignty is a fashionable word used today in describing computing architectures where external dependencies are minimal. This means that there is no external body that can either de jure or de facto exercise pressure or constraints on the ability to take the necessary decisions to improve an existing architecture.

A more practical viewpoint to define digital sovereignty is related to the identification of which (digital) activities can be outsourced while maintaining the authority to self-govern.

The comparison with outsourcing is important because this approach makes you lose digital sovereignty. On this point, the industry defines few clear criteria that must be addressed before outsourcing, such to be effective.

Namely:

a) The activity is not strategic nor it is core business

b) The activity has clear established standard interfaces or protocols that are used to define the outsourcing contract

c) There are multiple independent vendors implementing these standard interfaces.

If any of these three requirements is not satisfied, you are exposed to problems, in particular vendor lock-in, business or service failure, even blackmailing. The vendor lock-in has a critical impact when data are involved because the ultimate goal to preserve the access to your data is at stake, given that companies and contracts can fail, law changes and contracts can be subject to remote jurisdictions.

There are two levels of external dependency that can severely limit your independence from a particular vendor:

a) the fact that the vendor has your data

b) the fact that you use licensed software from your vendor and that the license can be revoked.

These two aspects will be discussed in the following paragraphs.

V. STORING YOUR DATA IN THE CLOUD

If you store your data in the infrastructure of a cloud provider, you rely on the vendor to implement all the processes necessary to ensure that the data are securely stored, preserved and not accessed by third parties. Of course, you can also store the data yourself on premises. However, when data are on premises, you will have to implement the same processes, but with the advantage that you will be able to audit your infrastructure to know where you are standing.

One question is: how can you verify that all processes you expect to be in place on the vendor's side to properly manage your data are really there? this verification is difficult. Often it relies on a blind trust in the vendor, or it is just "because it is written in the contract".

With data stored on the cloud, you have some indicators and statistics that can help you. For example, the fact that you pay a higher price for a service may give you the confidence that the service is better managed. Or the fact that a vendor claims to have a large number of satisfied customers may also reassure you, as a trouble shared is a trouble halved. Unfortunately, all these arguments are just marketing statements that are unrelated to real facts. Another qualitative indicator is the fact that the vendor claims several years of successful operation without major incidents or data loss. But also in this case, this is another good example where the disclaimer "Past performance is no guarantee of future results" fully applies.

This is where the standard outsourcing approach is important when moving data to the cloud and the cost analysis and the risk analysis are both essential.

First, the cost analysis should go well beyond the recurrent cost per byte stored. The network transfer cost must be carefully evaluated as vendors typical offer free data ingestion, but expensive data retrieval. In addition, the time required to execute a complete data retrieval is important, as you must define from the beginning a possible exit strategy to avoid lock-in. Finally, it is also important to ensure that costs are guaranteed over a certain number of years and that you know your future costs sufficiently early so that you have enough time to review and adapt your strategy.

Then comes the risk analysis, which is by far more delicate, because it requires a subjective judgment to measure the probability of a bad event and its impact. Having said this, here is a list of possible bad events that must not be underestimated:

a) the loss of access to the data due to a technical incident in the vendor's premises

b) the loss of access to the data due to a contractual disagreement with the vendor

c) the loss of access to the data due to a decision from the vendor's jurisdiction

d) the fact that you will need significant investment to either take your data out, or to change vendor (you are locked-in)

e) the sudden (or planned) increase of cost for the vendor service that you are paying for

As there are many things that can go wrong, the single cloud provider approach should only be considered as a short-term solution, where data is not strategic and long-term preservation is not required.

On the other hand, if long term solutions are needed, there are several approaches that can mitigate the risk, the most obvious is to have multiple cloud providers. With multiple cloud providers you are able to constantly verify the interoperability among the cloud providers which will ensure that you are not locked in. You can also store your data multiple times across providers so that any access loss to one copy with one vendor will not affect the other copy.

However, be aware that all these mitigation strategies are effectively transferring back to you the workload that you thought you had outsourced. With multiple vendors, you are now in charge of the data management, allocating and moving the data and archiving it to locations that are now considered ephemeral or unreliable. Therefore, you have insourced the process of data preservation that you initially wanted to outsource.

So far, we have discussed only the data preservation requirement. However, there are many other risks to take in account, when outsourcing cloud storage. A significant example is the risk of disclosure of confidential data due to a technical incident in the vendor's premises. Also here, there are mitigation strategies. In this particular example, the usage of client-side data encryption voids the risk, as the vendor has only access to blobs of encrypted data, and any disclosure of this data would have no negative impact. But … now you have to manage the encryption keys and to provide keys to decrypt the data to any relevant stakeholder who need to access the data. De facto, you have insourced all your security, despite the goal of outsourcing it.

Finally, another mitigation strategy is to keep an additional copy on premises, but again, this comes back to insource what you intended to outsource.

VI. SOFTWARE LICENSES AND DIGITAL SOVEREIGNTY

Similar to the usage of cloud providers, another critical component to preserve digital sovereignty is the software and the licenses that are needed to use it.

The available options go well beyond the simple decision between open source vs proprietary software, and the best approach is, again, to apply the outsourcing strategy and identify costs and risks involved, knowing that the licensing horizon offers countless possibilities.

Whenever:

a) the data being processed does not require high confidentiality,

b) the processing required is based on standard functionalities,

c) there are (tested) interoperable vendors implementing the desired algorithms,

then the need to have source code access is reduced, and a proprietary license may be the most appropriate option. In this scenario, especially when multiple vendors are available, you can expect to be in a strong position to be able to negotiate a cheap license in a sustainable partnership with the vendor that will last several years.

As for the use of storage cloud providers, whenever any of the previous conditions is missing, then a risk analysis becomes necessary. In this case, among all possible risks to be evaluated, one can mention:

a) the risk when using closed source solutions that the software leaks information to the vendor or contains unwanted hidden features that can compromise your security (for example: backdoors or maintenance interfaces)

b) that some or all functionalities of the software can be suddenly remotely disabled by the vendor

c) that license renewal conditions can be unilaterally imposed on you at the end of the current licensing period

Points a) and b) can be mitigated by giving preference to the open source approach that allows the software running in your premises to be scrutinized. On point c), it is important to have licensing conditions that allow some reduced use of the software beyond licensing expiration in case of non-renewal to ensure to have the necessary time to migrate to alternative solutions. In the case of a computing infrastructure that hosts a large amount of data, migrating to alternative infrastructures can take years, and this time must be included in the risk analysis.

These points demonstrate that if the software plays a role in delivering your business, being able to control its strategy can really give you a significant competitive advantage. In this case, it is obvious that software provides a huge flexibility and that is where you invest. However, the more you invest, the more complex it becomes, and this is the area where you need top-level skills in your staff.

As mentioned, using proprietary software under license offers the advantage that costs are easy to estimate and customization can be purchased. In this scenario the competitive advantage remains effective while your license is kept cheap. It is equivalent to outsource an activity that is not important to your business, in order to reduce the cost.

On the other hand, when you move to open source software you may save in licensing cost but you need high skilled personnel that maintain the strategic software, and this can be expensive. However, with this approach, you gain in flexibility and you only pay the customization cost.

Finally, one last option which gives you total sovereignty but also the highest cost is the full stack development, done internally in closed source. This approach is often taken when the software is so strategic that the software becomes the business.

The process to define the software strategy is very similar to the one used to define the extent of cloud storage usage: it is entirely based on a cost and risk analysis approach.

One final important aspect to be aware of is that, when choosing the open source approach, the critical mass to build a successful infrastructure is beyond what a small institute or company can usually afford. In this case, the consortium approach is the best practice to collaborate on well-focused projects that guarantee maintaining ownership of the critical activities at a minimum cost. However, you will not have the exclusiveness competitive advantage of the developments produced by the consortium.

VII. CONCLUSION

If you have various vendor relationships, you should not be surprised that the more critical a component is to your business, the more marketing pressure you will receive to outsource it. The general recommendation is to outsource only standard activities that are well defined and interoperable. This means that you should insource what is specific to you, and your critical activities. Do not outsource your own business! The open source approach remains the best practice to insource your critical activities at a minimum cost, and when you cannot afford the cost, it is preferable to have a consortium approach rather than to accept a vendor lock-in. Both the consortium and the open source approaches can guarantee a fixed cost for software. When you have reached an architecture where software accounts only as a fixed cost, you have reached the perfect scale-out solution, where the marginal cost of your growth will be minimized as you pay only for the cost of the additional hardware and the additional energy consumption. This approach is particularly valid for storage. If you manage to have no variable cost for the software, your cost of adding additional storage will be minimal and you can expect huge savings compared to cloud storage. The only drawback is that the critical mass that you need is large, but this can be addressed with the consortium approach.

Why data copyright and open licensing matter

Anouk Santos
Informational Resources and Archives Service (UNIRIS)
University of Lausanne
Lausanne, Switzerland
ORCID 0000-0002-1836-0835

A Abstract— *Research data are often not protected by copyright because they lack originality, but are generally made available under a copyright license. This situation is problematic firstly because research data are not always protected, and secondly because the licenses chosen are often restrictive and closed (e.g. those prohibiting modification and/or commercial use); or open, but not very suitable for Open Research Data (e.g. licenses requiring attribution or share alike). For those reasons, and to fully achieve the objectives of the Open Science movement with the least restrictions to impact data reuse, public domain licenses are recommended for sharing research data (e.g. CC0).*

Keywords—copyright, research data, data license, open license, Creative Commons, CC0, public domain.

I. Introduction

I devoted my master's thesis in Information science from the Haute école de gestion of Geneva to the legal framework of research data and data licenses (Santos 2020), mandated by the DLCM project. In this paper, I will present some considerations taken from my thesis. Those are not new findings, but rather reminders of what is needed to understand in terms of copyright and open licensing so as to fulfill the requirements for Open Research Data and, more generally, Open Science.

II. Data copyright

According to the Swiss Federal Act on Copyright and Related Rights (Swiss Confederation 2020), a work must comprise three conditions in order to be protected: to be a creation of the mind, to have an individual character, and to be expressed in one form or another (CCdigitallaw 2020). Having an individual character means that it is impossible for someone performing the same task to create an identical work. Data are not explicitly mentioned in the Swiss copyright law, so their protection must be assessed on a case-by-case basis according to the three aforementioned conditions. The following generic data examples can be mentioned: database, software, data visualizations, metadata or any other data. Thus, to assess if research data are protected by copyright or not is a difficult question to answer.

On the contrary, we know for sure that factual scientific data are not protected. Facts, information, ideas, formulas, algorithms, scientific measurements, etc. are not eligible to copyright protection because they are not considered individual works of authorships. They are discovered and compiled by a researcher's methods and this is something that copyright does not reward (Pantaloni 2017). So, in a dataset, some parts can be protected by copyright while others are not, and sometimes, the entire dataset is not protected. This is something that is important to keep in mind when thinking about data licensing.

III. Open data licensing

Here I will be referring to the Creative Commons (CC) licenses. They are the recommended licenses to use because they are compatible with data, widespread within the scientific field, quite simple to use and well-known. However, CC license are not always fully understood, notably some of their consequences on scientific research. Note that CC licenses are not designed for software or computer code and that there are open source licenses more suitable for this type of data.

One very important definition in order to comprehend open data is *The Open Definition* from The Open Knowledge Foundation (2015) that says: "Open data and content can be freely used, modified, and shared by anyone for any purpose (subject, at most, to requirements that preserve provenance and openness)". In the terms of the CC licenses, to preserve provenance means to give attribution, and to preserve openness is to require that the derivative

work is shared under the same license as the original work (Shared Alike). In fact, according to the open definition, there are only three CC licenses that are truly open and suitable for open data: CC0 (a public domain dedication), CC-BY and CC-BY-SA. All three allow data to be freely used, modify and shared. All the others CC licenses are therefore considered as not open.

CC licenses with the No Derivatives element (ND) are a problem for scientific research because the ND requirement forbids data to be modified, corrected, translated, combined or enriched with any other data, or shared partially (only the full dataset can be shared) (Ball 2014, Kreutzer 2014). Overall, those licenses prevent the creation of derivative works, which is not useful for Open Science because in the research field data generally exist in order to be crossed-referenced with other data, which is impossible to do with such a license. Furthermore, research is very often based on previous scientific works and thus there is a strong need to be able to combine data (Hirschmann 2020).

When a dataset is made available under a CC license with a Noncommercial element (NC) it is not open either. The main problem is to establish what is a noncommercial use because interpretations differ. In fact, a NC license could prevent some common reuse of research data: in a work for which the author receives a financial retribution (for example a published book or the publication of an article in a journal owned by a commercial editor), or also public-private partnerships, which occur regularly. As a result, data must be allowed to be used for any purposes, even commercial ones, to be truly open.

We saw that according to The Open Definition (The Open Knowledge Foundation 2015), CC licenses with Attribution (BY) and Share Alike (SA) are open. But unfortunately, those requirements can be problematic too regarding Open Research Data.

The problem with CC-BY-SA, which is a copyleft license, is the incompatibility with any other copyleft licenses. For example, it is impossible to combine a dataset which is under CC-BY-SA with another one under CC-BY-NC-SA, because both require that the same license is kept afterwards, which is impossible (Ball 2014). For this reason, the SA element affects the interoperability of data and increases the incompatibility of licenses, already initially caused by their proliferation.

As for attribution, there are two main problems that raise voices against it for research data. The first one is known as “attribution stacking”: when reusing or combining a lot of datasets that have a lot of authors, you must cite each of them correctly. It can be time consuming and difficult to achieve as the datasets are reused (Ball 2014). The second one is more an ethical consideration: can researchers articulate community norms, here peer citation, as a legal form? The answer is no: attribution cannot be legally binding by a license if the data are not protected by copyright… and as a result not eligible to licensing (Ball 2014). Thus, for instance, with factual scientific data that are not protected, or works that are into the public domain, someone doing that would be overriding his rights to that content.

Consequently, we are left with CC0, a public domain dedication, as the best choice. Here are some of the reasons why:

- CC0 solves the problem of licenses’ incompatibility: placing data into the public domain means that anyone can reuse them for any purpose. It avoids creating data silos that are incompatible with each other (Lämmerhirt 2017).
- CC0 achieves legal interoperability as it is an answer to the ambiguity of data copyright: there is no need to know which data are protected or not because all of them are placed into the public domain (Fortney 2016). CC0 allows legal interoperability by waiving patrimonial and moral rights of the data that are protected (to the extent allowed by law).
- There is a certain logic to put publicly funded data into the public domain. It is also coherent with the general sharing and reuse ethics which prevail normally within the scientific community (Murray-Rust et al. 2010).
- Open Science is easier to achieve with the least restrictions to impact data reuse (Labastida, Margoni, 2020).

IV. CONCLUSION

In conclusion, if non-open restrictions are put in place in licenses for research data, they must be carefully considered and justified because of their consequences. Such licenses can be very concrete barriers to the reuse of the data, and more globally the sharing of scientific research. Therefore, in my opinion, there is a need to raise awareness about data copyright and advocate for open licenses, and preferably for the public domain, in order to ensure that the principles of the Open Science movement are preserved. For the researcher sharing data openly also has its advantages: if data are reused, the data creator will be cited for his work, thus the visibility and discoverability of his

research will increase. He will potentially create opportunities for new research collaborations and demonstrate his integrity and the robustness of his work (Leeming 2017).

REFERENCE

Ball, A. (2014). *How to License Research Data.* Digital Curation Centre. https://www.dcc.ac.uk/resources/how-guides/license-research-data.

CCDigitalLaw. (2020, April 1st). *2.1 Protected work.* CCdigitallaw. https://ccdigitallaw.ch/index.php/english/copyright/2-what-work/21-urheberrechtlich-geschuetztes-werk.

Fortney, K. (2016, September 15). CC BY and data: Not always a good fit. *Office of Scholarly Communication of the University of California.* https://osc.universityofcalifornia.edu/2016/09/cc-by-and-data-not-always-a-good-fit/

Hirschmann, B. (2020, January 20). *Creative Commons Licenses*. Manual Research Collection (ETHZ Library). https://documentation.library.ethz.ch/display/RC/Creative+Commons+Licenses.

Kreutzer, T. (2014). *Open Content: A Practical Guide to Using Creative Commons Licences.* German Comm. for UNESCO. https://meta.wikimedia.org/wiki/File:Open_Content_-_A_Practical_Guide_to_Using_Creative_Commons_Licences.pdf.

Labastida, I. & Margoni, T. (2020). Licensing FAIR Data for Reuse. *Data Intelligence*, 2(1-2), 199-207. https://doi.org/10.1162/dint_a_00042

Lämmerhirt, D. (2017, December). Avoiding data use silos: How governments can simplify the open licensing landscape. *research.okfn.org.* https://research.okfn.org/avoiding-data-use-silos/

Leeming, J. (2017, June). Ask not what you can do for open data; ask what open data can do for you. *Nature jobs*. http://blogs.nature.com/naturejobs/2017/06/19/ask-not-what-you-can-do-for-open-data-ask-what-open-data-can-do-for-you/

Murray-Rust, P., Neylon, C., Pollock, R. & Wilbanks, J. (2010, February 19). *Panton Principles: Principles for Open Data in Science.* Panton Principles. https://pantonprinciples.org/

Open Knowledge Foundation. (2015). *Open Definition 2.1*. http://opendefinition.org/od/2.1/en/.

Pantaloni, N. (2017, Decmber 12). Copyright and data curation. *Indiana University Libraries Blogs.* https://blogs.libraries.indiana.edu/scholcomm/2017/12/12/copyright-and-data-curation/.

Santos, A. (2020, August 14). *Données de la recherche : cadre juridique et licences.* Geneva: Haute école de gestion. Master's thesis. https://doi.org/10.5281/zenodo.3967402.

Swiss Confederation. (2020, April 1st). *Federal Act on Copyright and Related Rights (Copyright Act, CopA) of 9 October 1992*. https://www.admin.ch/opc/en/classified-compilation/19920251/index.html

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.*

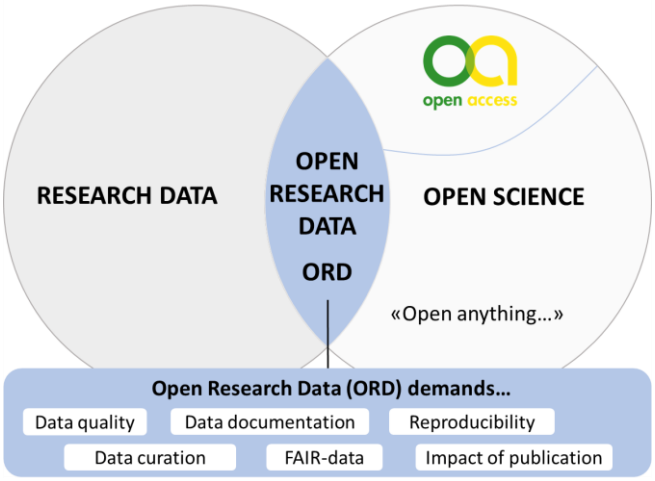


Fig. 1. Open Research Data (ORD) increases demand on best practices in Research Data Management (RDM).

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[10]).

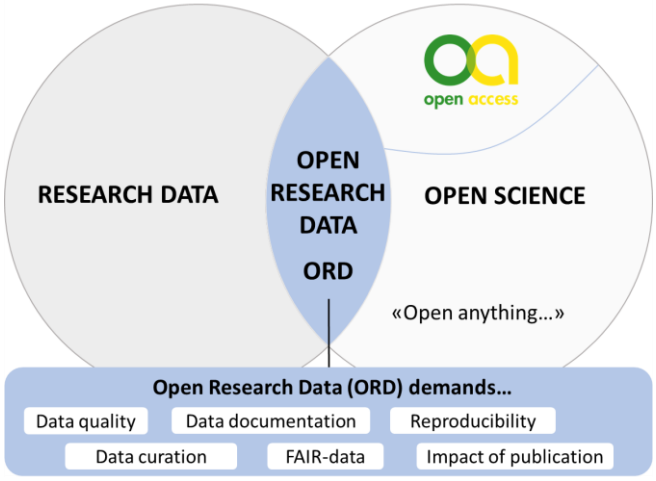


Fig. 1. Open Research Data (ORD) increases demand on best practices in Research Data Management (RDM).

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?

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 (www.dlcm.ch) project. The results and questions raised in practice have been reflected in workshops within the project consortium. We distinguished between two types of pilots.

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

[10] https://sfdora.org

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[11] of the research projects had been completed and related paper publications had already been done.

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

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

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.

[11] 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.

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

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[12] or by browsing on re3data.org, a registry of research data repositories. FAIR data repositories with certificates (e.g. CoreTrustSeal[13]) 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[14]. 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[15]). This leads in practice to a better findability and data redundancy. Second, data publications in the field of genomics are often mandatory. A data

[12] e.g. Milzow et al. (2020); von der Heyde (2019)
[13] https://www.coretrustseal.org
[14] https://forsbase.unil.ch
[15] http://www.insdc.org

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[16], 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.

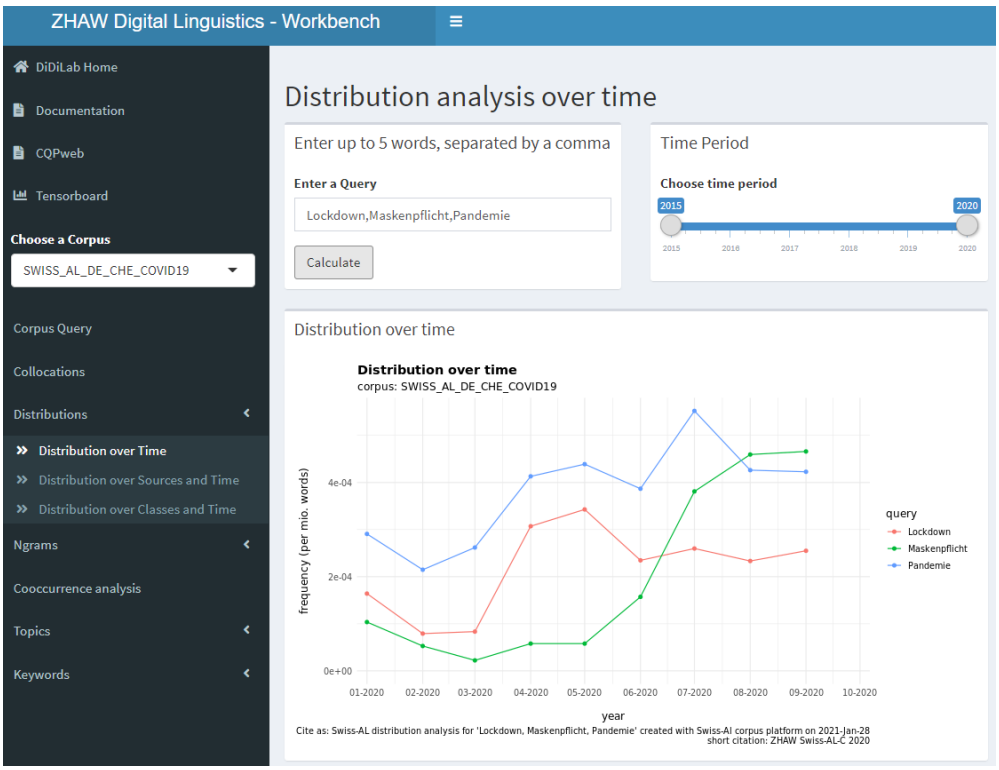


Fig. 3.Corpus linguistic workbench (accessible under https://swiss-al.linguistik.zhaw.ch)

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 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[17]). 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. derivatives 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.

[16] https://dataverse.harvard.edu
[17] 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

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 of 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 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[18].
- *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 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*[19] 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

[18] 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.
[19] Possible source of information is the UK Data Service (UK Data Service, 2021).

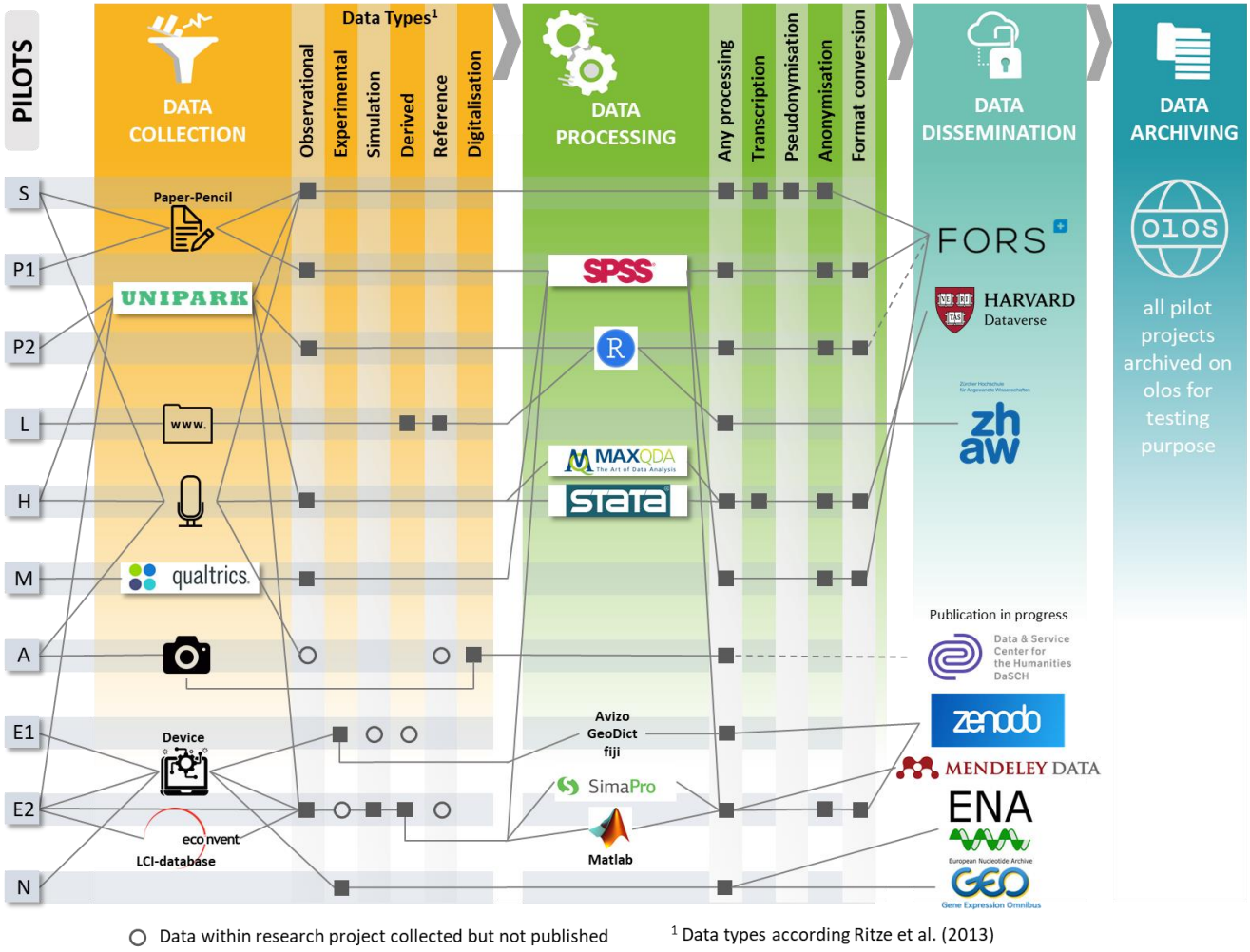


Fig. 4. Data processing workflows of ORD pilot projects.

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, this latter being 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[20] 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.

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).

[20] https://renkulab.io

As mentioned, we basically used the ELN/LIMS-system[21] openBIS in two laboratories at ZHAW. One laboratory additionally tested REDCap[22] 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[23]). 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 requires 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.*

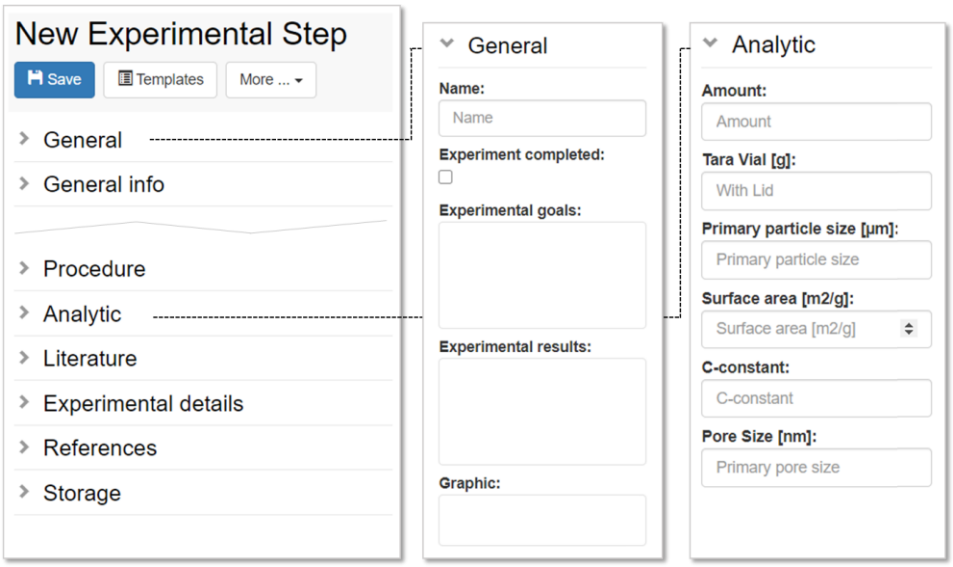


Fig. 5. Generic openBIS template with additional sections and specific metadata fields to improve searchability.

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.

[21] openBIS is a combination of an ELN and a Laboratory Information Management System (LIMS).
[22] REDCap is an Electronical Data Capture System (EDC) which is developed by the Vanderbilt University (https://projectredcap.org).
[23] Overviews to be found e.g. in DLCM (2017a, 2017b); Harvard Medical School (2021).

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 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.

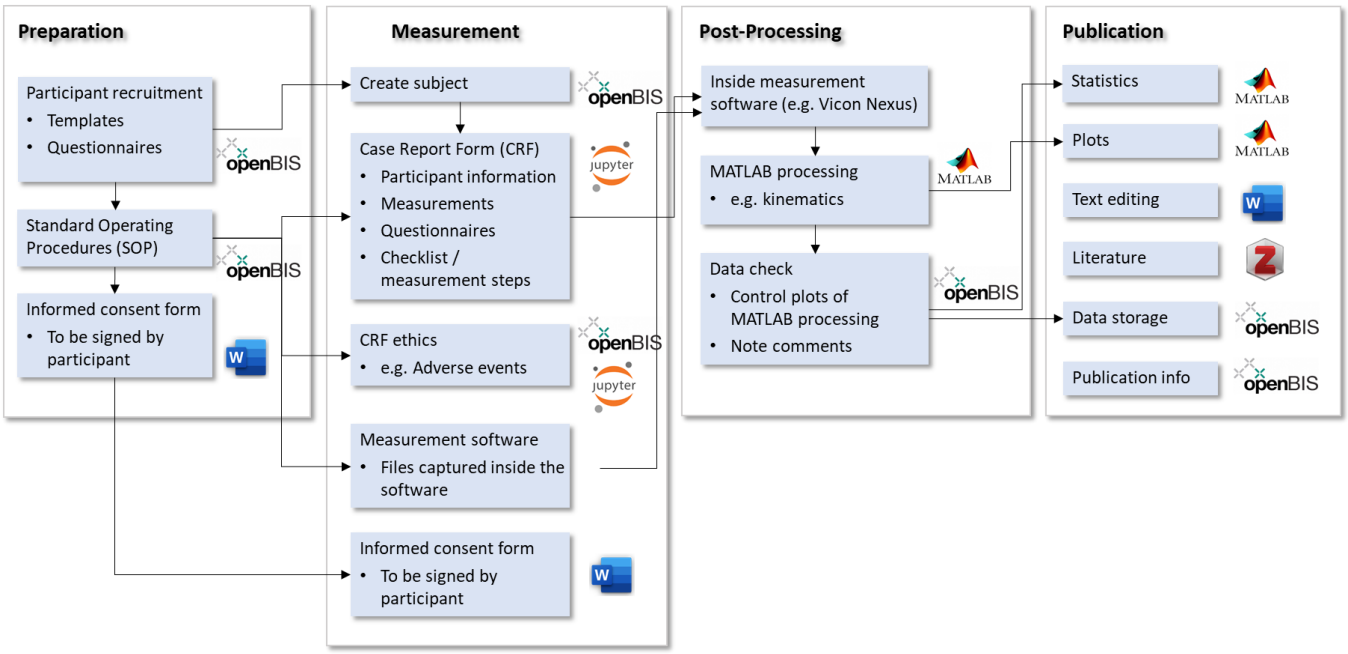


Fig. 6. Standard project procedure of the Movement Laboratory at ZHAW with pilot implementation of openBIS and Jupyter.

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.

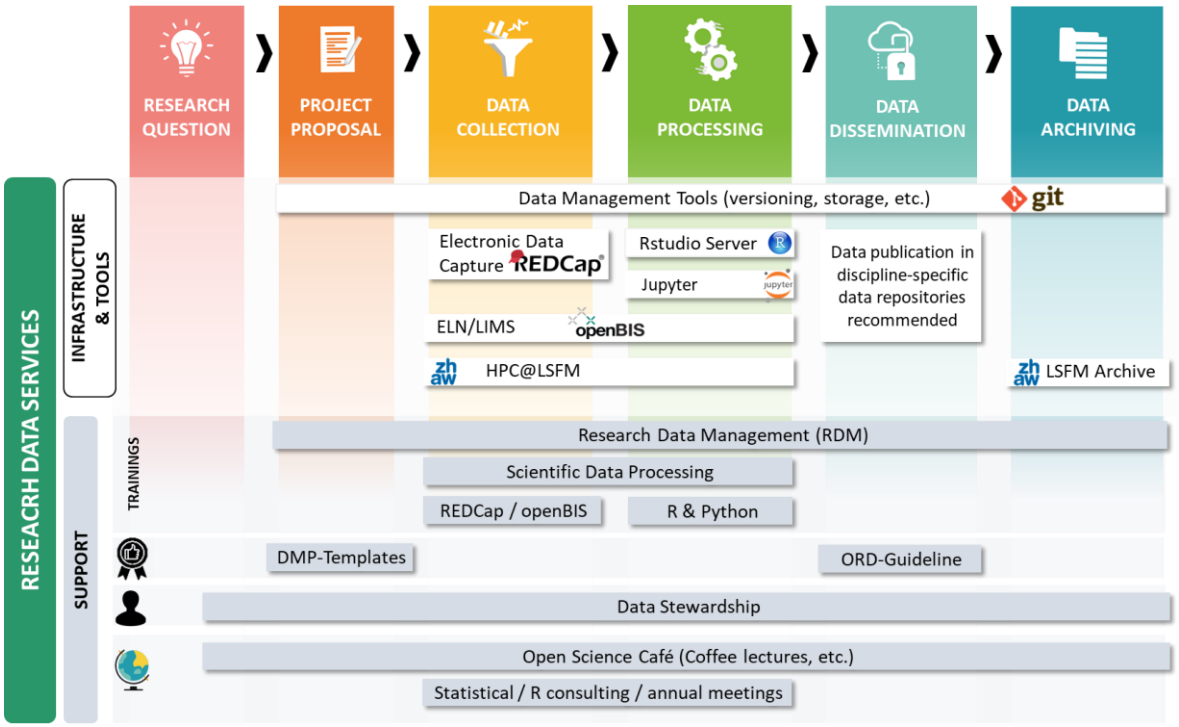


Fig. 7. Starting portfolio for Research Data Management services at Zurich University of Applied Sciences (from an ORD perspective; further tools and services are available).

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).

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[24]. 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[25]. 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

[24] Such as with FORSbase or Qualiservice (https://www.qualiservice.org)
[25] see Burgard et al. (2017)

recently founded the Open Science Café[26], 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 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

[26] Our (public) Open Science Café (https://bit.ly/39o5TCb) is a virtual space hosted by wonder.me

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 - olos.swiss) 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/

DMLawTool – A Guiding Tool for Researchers to Address Legal Aspects in Data Management

Anna Picco-Schwendener
eLearning Lab
USI – Università della Svizzera italiana
Lugano, Switzerland
anna.picco.schwendener@usi.ch

Branislava Trifkovic
eLearning Lab
USI-Università della Svizzera italiana
Lugano, Switzerland
branislava.trifkovic@usi.ch

Suzanna Marazza
eLearning Lab
USI – Università della Svizzera italiana
Lugano, Switzerland
suzanna.marazza@usi.ch

***Abstract*—*DMLawTool is a web-based tool that guides researchers working mainly in the fields of humanities and social sciences through the most relevant legal issues related to data management. It has the form of decision tree that provides different solution approaches on how to correctly deal with research data during the whole data lifecycle. The released version of the DMLawTool will be available openly and free of charge as an open-source software on the CCdigitallaw.ch platform (www.ccdigitallaw.ch). This paper illustrates the different development steps of the tool, introduces its structure and main functionalities and provides a reflection on the faced challenges.*

Keywords*— *research data management, data protection, copyright, legal aspects;

I. INTRODUCTION

DMLawTool is currently being developed by the Università della Svizzera italiana in collaboration with the University of Neuchatel within the P-5 program of swissuniversities (swissuniversities, 2021).

The DMLawTool helps researchers to understand which legal issues they have to consider when dealing with research data. The two main topics it addresses are data protection and copyright. With the help of a decision tree, the tool guides the user through the most relevant legal aspects and finally provides suggestions on how to correctly handle research data throughout its whole lifecycle, from data collection to data archiving or erasure (elimination of data).

The idea to develop such a tool originates from the need of research data archives to address also legal aspects when archiving data. To date research data archiving platforms in the fields of humanities and social sciences in Switzerland mainly focused on the development of technically safe and secure solutions. However, with the introduction of the European General Data Protection Regulation (GDPR) (European Parliament and Council of European Union, 2016) the need of having complete Data Management Plans (DPMs) for projects of the Swiss National Science Foundation (SNF) (Swiss National Science Foundation, 2021), the upcoming revision of the Swiss Federal Act on Data Protection (Schweizer Eidgenossenschaft, 2019) and the recently revised Swiss Federal Act on Copyrights and Related Rights (CopA) (Schweizer Eidgenossenschaft, 2020), it has become essential to address legal aspects of data management in order to act in a legally compliant way. However, there is still much incertitude around legal aspects. Even though the legal liability of research data lies with the researcher - also during the archiving process - platforms understood the importance of raising awareness about legal aspects and providing researchers guidance and support.

The project thus closely collaborated with five Swiss research data archiving platforms in the fields of social sciences and humanities, in order to understand their needs and get an overview of the legal problems they have already come across.

The DMLawTool wants to act as a light bulb that turns on whenever researchers feel lost within the misty legal jungle and are unsure about the best approach to adopt in order to handle and archive research data in a legally compliant way. It addresses questions ranging from copyright (e.g. Is my research data considered a work and thus protected by copyright? Do I own the rights to make my data available for re-use?) to data protection (e.g. Does my data contain personal data? If yes, do I really need to anonymize it? How do I correctly anonymize it? What do I have to consider when working with personal data?), and licensing (e.g. which licenses should I use if I want to make data available for re-use in an open way?). The tool encourages open access practices and re-use of data wherever possible. Furthermore, it pays particular attention to the vulgarization of the legal language.

The tool is currently developed in English language and will be available as open-source software on the CCdigitallaw.ch platform (www.ccdigitallaw.ch) by the end of March 2021. In this way all platforms dealing with humanities and social science research data in Switzerland can use it either as a standalone instrument or integrate it in their archiving processes.

The following chapters will present the empirical basis of the project, introduce the DMLawTool and its functionalities, and provide a short reflection on the main challenges faced during the project.

II. Empirical Basis

A. *Data Collection*

In order to learn about how research data archiving platforms currently deal with legal aspects, and to identify legal issues and questions that they have already come across, the project partners closely collaborated with five Swiss repositories active in the fields of social sciences and humanities: 1) DaSCH (https://dasch.swiss), a platform for humanities research data 2) Yareta (https://yareta.unige.ch), the research data repository of Geneva's Higher Education Institutions, 3) FORSbase (https://forsbase.unil.ch), a platform that provides access to data about social sciences studies in Switzerland, 4) Dodis (https://www.dodis.ch), the archive for diplomatic documents of Switzerland, and 5) Historisches Lexikon der Schweiz (https://hls-dhs-dss.ch), an encyclopedia made out of research data.

Between January and May 2020, the project team conducted a total of six interviews with representatives of these platforms (three in person and three online).

The interviews revealed the central role of the researchers in the archiving process and also the fact that the legal liability of archived data actually lies with the researchers. In light of this, the project team decided to interview also three researchers active in the fields of humanities and a director of a university's research support center. This allowed to consider also the needs of researchers and considerably enriched the insights gained from the collected data.

B. *Data Analysis & Findings*

All information gathered during the interviews has been systematically organized in an Excel Table with a particular attention towards legal issues that have already emerged, solution approaches applied so far, and expectations towards the DMLawTool.

Subsequently, visual maps have been created to group, structure and link contents within single thematic clusters such as copyright, data protection and expectations towards the tool. Figure 1 shows the thematic map with clusters related to copyright. The map shows the importance of concepts such as re-use of data, licenses, image rights and ownership of copyrights. With regard to data protection, anonymization and the use of personal data resulted to be particularly hot topics. As far as the tool is concerned, interviewees expect to see practical examples of do's and don'ts. They would further appreciate templates for different kind of contracts and consent forms. They like the idea of a decision tree with a simple and straightforward structure that guides the user. Last but not least they suggest using a language which is accessible to a public at large and including some interactive elements as well as references and

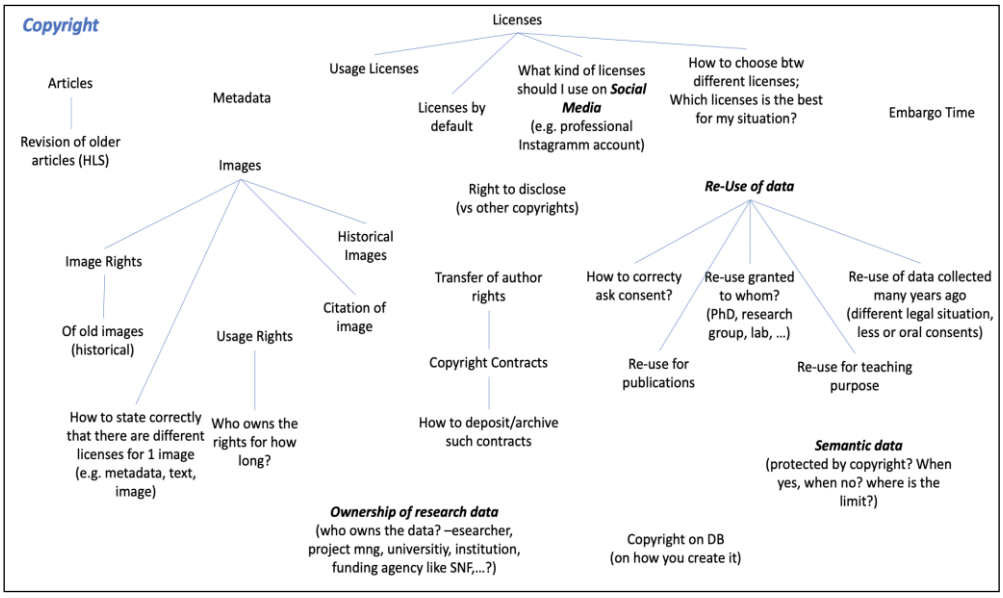


Fig. 1. Thematic map with cluster related to copyright

links to additional resources allowing to 1) further investigate certain aspects and 2) contact people who can provide further help.

III. DMLawTool and its Functionalities

This paper describes a beta version of the DMLawTool, which can be accessed at the following link: https://dmlawtool.web.app. It corresponds to the development stage of the tool at the end of January 2021 and might be subject to further changes.

A. *Decision Tree Structure and Behavior*

The DMLawTool is structured as a decision tree and as such is composed of branches and nodes.

When fully expanded, it provides an overview of the main legal issues related to data management, however, its main goal is to guide users through the various legal topics by providing the necessary explanations to move through the tree. This allows users to understand which legal aspects are relevant for their specific cases and to further deepen them. This is also the reason why the tree at first is not shown fully expanded but has to be opened-up step by step by the user. As shown in Figure 2, a "+" sign on the node allows to expand the next branches while a "-" sign allows to collapse nodes.

Each node has a meaningful name and is connected to two or more child nodes. Users are asked to choose among various available branches to proceed within the tree. At the end of each branch there is a so-called "end-node". These nodes are highlighted with a different color (currently yellow) and generally represent specific actions that the user can or should take once arrived at this point. In other words, the end-nodes provide solutions to the legal issues identified as relevant for a specific situation.

Beyond the name, each node contains knowledge notions that are necessary to advance through the tree. They can be accessed by clicking on the node. The node texts have a precise guiding role, and do not follow the rules of a traditional encyclopedia, which would simply explain the term of the node. In fact, each node text provides those explanations and definitions necessary to understand the following nodes, so that a user is able to decide which branch to choose next. This means, that you will not find a definition of data protection in the data protection node, but you will find this definition in the previous node, as you need to know what data protection is in order to decide whether it is relevant for your current research.

Each node text is structured in a similar way. Even though the tool is about legal issues, the storytelling is

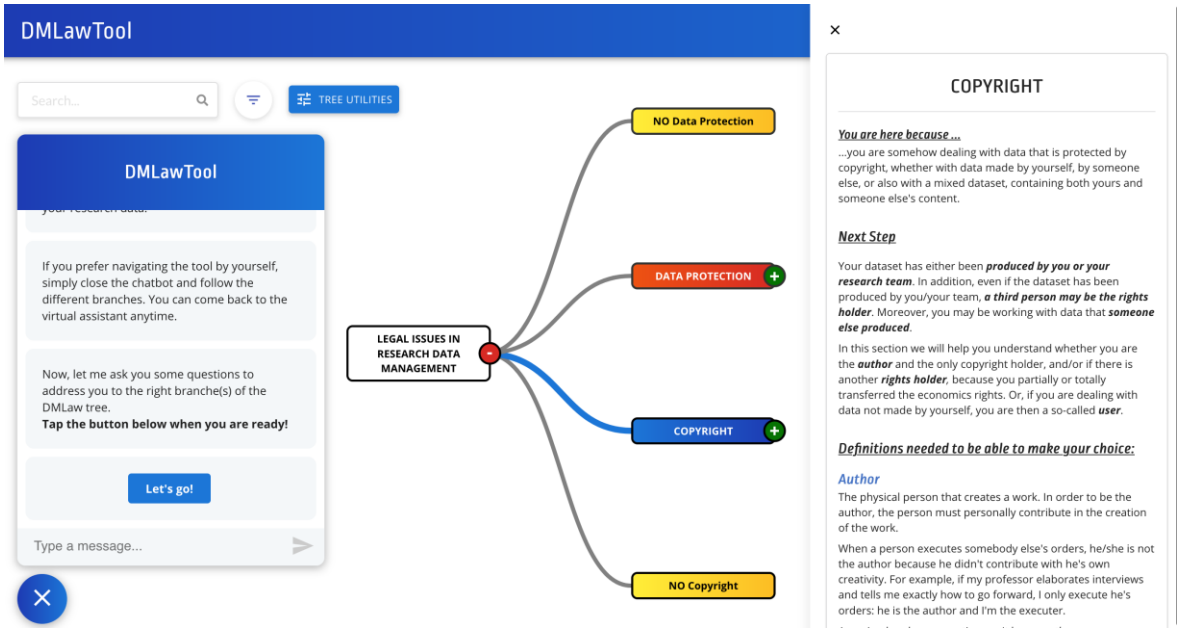


Fig. 2. DMLawTool with 1st branches of tree, text of copyright node and Chabot

fundamental. This is why each node text starts with a section called "You are here because …" followed by "Your next steps". Then the necessary definitions are provided and illustrated with practical examples. Each text ends with a section called "be aware of" and FAQs. End-nodes have a slightly different structure as they generally explain how to implement a specific action. For example, the end-node "anonymization" will illustrate different anonymization techniques.

Content wise, the decision tree is structured around data protection and copyright. Already at the first level there are also two end-nodes: “no data protection” and “no copyright”. These nodes show the potentials and pitfalls for your research data, if copyright or data protection regulations do not apply. In these cases, in fact, archiving, sharing and making data available as open access and for re-use is particularly easy as there are not many obstacles. These nodes thus allow favoring open access practices.

B. Chatbot

Another important element of the DMLawTool is the interactive Chabot, which simulates a conversation with the users. Its main function is to welcome users to the DMLawTool and introduce the decision tree, its behavior and functionalities. At first, it has been developed to lead the users to the first branches of the tree, however, after some user testing it emerged that in this way contents are provided in two different locations, which might be confusing. However, the designers are still thinking about other ways to take advantage of the Chabot in order to maximize its potential. The Chabot can be closed and re-opened at any time.

C. Search Functionalities and Tree Utilities

DMLawTool has a powerful text search function. When you enter a word, the tool highlights all nodes of the tree in which the word is present. Furthermore, each node has a series of tags assigned. This allows filtering contents based on chosen tags.

Tree Utilities allow to center and expand the tree and to zoom in and out. Furthermore, the tool works a lot with colors. Each of the two main legal topics have a dedicated color: data protection nodes are red while copyright nodes are blue. This allows to visually highlighting to which thematic area a node belongs.

IV. Main Challenges

A. Selecting the most Relevant Legal Aspect

DMLawTool focuses on legal aspects that are relevant for research data management. This means that the tool is a tradeoff between explaining every single legal detail or exception and providing best practices that are relevant for the people working with research data. The process of identifying aspects that could be neglected was not always easy. To guarantee a high quality of the tool, each item in question has thus been discussed extensively among the different legal experts involved in the project.

B. Vulgarization of legal language and storytelling

It is known that legal people speak their own language, which is often incomprehensible to non-experts. Using a simple vocabulary and creating a clear narrative, were thus among the main goals of the project. In order to find a balance between legal accuracy and general comprehension, legal and communication experts worked hand in hand: legal experts wrote the definitions while the communication experts elaborated the storytelling and included/vulgarized the definitions. In this way, all texts went through various iterations in order to be checked for their solidity from both a legal and communication point of view.

Another challenge was to avoid content duplications. Each legal aspect should have its specific place within the decision tree structure. This is important for the tool’s maintenance as it should be possible to implement changes in only one single place.

V. Conclusion

With the help of a carefully developed decision tree, the DMLawTool helps researchers to identify those legal aspects that are relevant for their projects and proposes them different solutions on how to handle their data in a legally compliant way. In this way, the tool wants to take away fears and favor open access, sharing and re-use of research data.

The focus of the tool has been widened during the project to cover not only the archiving stage but the whole life cycle of research data, from data collection to its archiving or cancellation.

Scalability is a very important aspect of the tool in order to allow adaptations or extensions. The tool has been developed in English, but translations in other languages are welcome and easily possible thanks to the fact that all contents will be published under an open Creative Commons license (Creative Commons, n.d.) and the tool will be available as open-source software.

Acknowledgment

The DMLawTool project is supported by the P-5 program of swissuniversities, under grant number 192-009.

We would like to particularly thank all the interviewees for their availability and insightful contributions in the data collection phase.

A warm thanks goes to all people involved in the project, in particular to Nathalie Tissot for the scientific supervision, Ivan Pavic for the software implementation, Yves Bauer for his legal inputs, and Mattia Pera for the graphic design.

References

Creative Commons. *Share your work*. Creative Commons. https://creativecommons.org/share-your-work/

European Parliament and Council of European Union. (2016). *Regulation (EU) 2016/679*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32016R0679&from=EN.

Schweizer Eidgenossenschaft. (01.03.2019). *Federal Act on Data Protection.* Fedlex. https://www.fedlex.admin.ch/eli/cc/1993/1945_1945_1945/en

Schweizer Eidgenossenschaft. (01.04.2020). *Federal Act on Copyright and Related Rights*. Fedlex. https://www.fedlex.admin.ch/eli/cc/1993/1798_1798_1798/en

Swiss National Science Foundation. (2021). *Data Management Plan (DMP) - Guidelines for researchers*. SNF. http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx

Swissuniversities. (2021). *P-5: Scientific information*. Swissuniversities. https://www.swissuniversities.ch/en/topics/digitalisation/p-5-scientific-information

Three years of publishing data in ETH Zurich's Research Collection: Lessons learned and new developments

Barbara Hirschmann
ETH Library
ETH Zurich
Zurich, Switzerland
http://orcid.org/0000-0003-0289-0345

Abstract – *In June 2020, ETH Zurich's Research Collection celebrated its third anniversary. The Research Collection serves as an institutional repository for ETH Zurich that can host both publications and research data and is operated by the E-Publishing team at the ETH Library. Publishing research data and advising customers on research-data-specific questions in the publishing workflow has emerged as a new field of activity for the team. With over 800 research data items published over the last few years, we have now gained a good understanding of the actual use cases for publishing data in an institutional repository at a large university for science and technology. We regularly talk to researchers about the incentives and requirements for publishing their data and monitor what kind of data they deposit. In this paper, we share and discuss our insights. We present statistics on the types of deposited datasets and explain how "FAIR" they are in terms of accessibility, licences and metadata. We also discuss our workflows for checking datasets for formal quality criteria and compliance with institutional policies and how to bridge publishing and preservation requirements in a research data repository. Finally, we give an overview of two ongoing development projects. The first one aims to enable ETH researchers to deposit datasets directly from the data management tool openBIS, while the second one will deliver a solution for publishing large datasets via the Research Collection.*

Keywords – institutional repository, data publishing, quality assurance.

I. Introduction

This article gives an overview of the functionalities of ETH Zurich's institutional repository Research Collection. It focuses on the repository's features for data publishing and illustrates how ETH researchers have actually used the repository during the last three years. It also explains what type of checks and workflows the ETH Library has set up to assure formal quality and policy compliance of the deposited datasets. Lastly, we report on the status of two ongoing development projects that will expand the repository's capabilities for data publication.

II. The Research Collection

A. Overview

The Research Collection is ETH Zurich's repository for publications and research data. It hosts research output produced by academic staff at ETH Zurich, one of the leading universities of science and technology in mainland Europe. The repository is operated by the ETH Library, which serves both as the main library of ETH Zurich and as a Swiss national centre for technical and scientific information.

The Research Collection is a publication platform that offers three main functionalities: it is a directory of all publications produced at ETH Zurich; it is an open-access repository; it is a research data repository. The platform was developed by the ETH Library from 2014 to 2017. The project included a tender process to select a service provider that led the technical implementation of the repository and provides ongoing maintenance support. It also involved the migration of data from two separate legacy systems into the new repository. While the previous systems were both based on the open-source software Fedora, the Research Collection runs on DSpace, an open-source repository tool widely used in academic libraries worldwide.

In order to integrate the Research Collection into the information landscape at ETH Zurich, and to fulfil all the requirements of a platform with the functional scope described above, the ETH Library implemented comprehensive customisations in DSpace. The repository now features various interfaces with internal and external systems. For example, in the ETH Zurich's academic reporting system, in the Annual Academic Achievements, and on researchers' institutional websites, Research Collection data are used to display publication lists (Hirschmann, 2018).

III. Features for publishing research data

Although the Research Collection hosts open-access publications and research data, certain features were designed and implemented especially with those users in mind that use the repository as a research data repository.

For example, research data can be published as supplementary material with a publication but also as a stand-alone publication. If the dataset is a supplement to a publication, there is a feature to link these two items together. When users upload their research data, they can choose a value from a list of resource types to categorise their data. The types offered are a subset of the resource types defined in the DataCite Metadata Schema (DataCite Metadata Working Group, 2019) and include dataset, image, model, software, sound, video and data collection.

Users define the access rights for their datasets themselves. The possible access rights for research data range from open access to closed access, the latter meaning that only repository staff can access the files. Options for restricting access to a dataset also include embargoes and granting access to all or selected users from ETH Zurich only. It is worth noting that the metadata of an item are always freely accessible, so even closed-access datasets will have a publicly visible landing page. If a dataset has restricted access settings, end users can still request access to the files from the landing page via a request form. Repository staff forward access requests to the submitter or rights holder of a dataset who then decides whether to grant the requester access to their data. For freely accessible datasets, submitters can choose an open content licence that will then be displayed on the landing page of the dataset from where end users download the files.

For each dataset, a digital object identifier (DOI) is minted. If users need the DOI before actually publishing their dataset – for example to include it in a manuscript – they can reserve a DOI. The Research Collection also displays download statistics for published datasets both at file and item level. For end users, there is a feature to preview the contents of ZIP and TAR containers before actually downloading the files.

In terms of file formats, there is no technical limitation to what users can upload. If a certain file format is known to repository staff and has therefore already been added to the DSpace file format registry, the Research Collection displays the support level for the uploaded files to the submitter. Users can then choose that the library should keep their data for an unlimited period of time or, alternatively, they can indicate a limited retention period of 10 or 15 years, for example if they already know that their file formats will not be usable over the long term. All uploaded files, independent of their retention period, are transferred to the library's preservation system, the ETH Data Archive, which is based on the software Rosetta by Ex Libris (see Töwe & Barillari, 2020).

IV. How ETH researchers use the repository

In this chapter, we present some insights into how ETH Zurich's researchers actually use the repository when it comes to publishing their datasets.

Depositing data in the Research Collection is not mandatory for researchers at ETH Zurich. While there is a strict requirement for researchers to report all their publications via the Research Collection so they can be listed in the annual academic reports and there is also an open-access policy (ETH Zurich, 2018) that requires researchers to deposit open-access versions of their papers in the Research Collection, there is no dedicated policy for depositing and publishing research data. The Guidelines for Research Integrity and Good Scientific Practice at the ETH Zurich (ETH Zurich, 2007) do require proper data management and contain a general expectation to share data, but they do not require deposit in the Research Collection.

In 2018, the first full year of operation of the repository, researchers deposited 191 datasets in the Research Collection. The number of published datasets has since grown each year, from 233 items in 2019 to 329 items in 2020. This brings the total number of datasets published in the Research Collection to 865 items at the end of 2020.

As mentioned above, there are various subtypes of research data to select from when uploading data. Around 80% of users categorise their data as either a dataset or data collection, while the more specific types such as image or video are not used as much (Fig. 1).

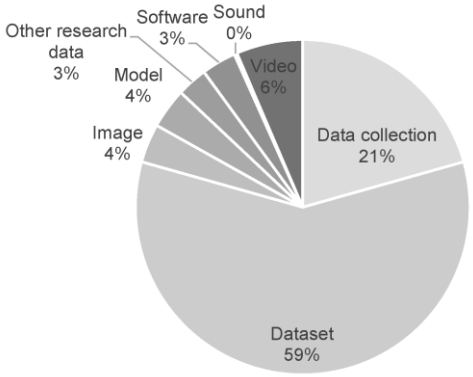


Fig. 1. Types of datasets in the Research Collection

When uploading datasets, users are only required to enter a few mandatory metadata fields in the submission form such as title or creator. However, there is also a range of optional metadata they can provide and this information is particularly important when it comes to making datasets findable in compliance with the FAIR principles (Wilkinson et al., 2016). Looking at the usage of these optional metadata fields, we can see that most users do not provide any information in the methods or software section and only a few datasets are linked to a grant. Less than a third of all research data items contain an abstract or subject keywords. However, around half of all datasets have an entry in the "Related publication" field, so the functionality to link publications and datasets is used quite often and plays an important role in making the data discoverable (Fig. 2).

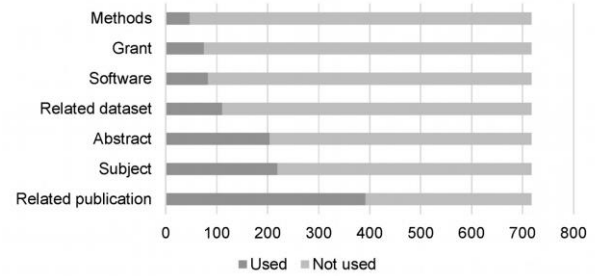


Fig. 2. Share of datasets for which submitters provided certain optional metadata.

Looking at the file formats, there is a large amount of datasets that users provide in ZIP containers. The individual files are often text files or CSV files, but also PDF files or other formats from a long list of other proprietary and non-proprietary file formats (Fig. 3).

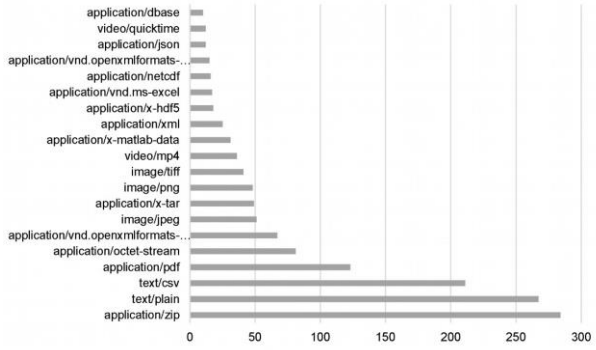


Fig. 3. Number of files of a certain format deposited in dataset items in the Research Collection.

In terms of availability, ETH researchers rarely use the option to restrict access to their datasets that they deposit in the Research Collection. There are datasets that are deposited in an external repository and then only linked from the Research Collection. These are the items described as "Metadata only" in Fig. 4. Apart from these items, almost all datasets are published open access.

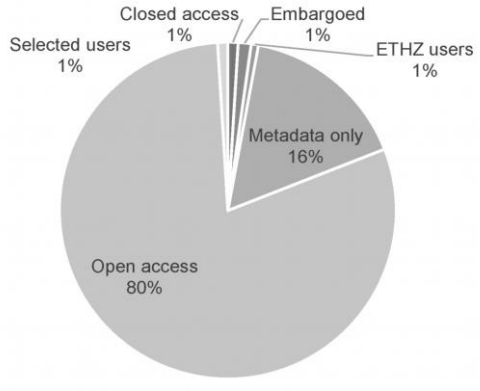


Fig. 4.Access status of datasets in the Research Collection.

When it comes to licensing, around half of the users decided to publish their dataset without an open content licence, instead using the repository's standard copyright statement that allows usage for non-commercial purposes but does not allow redistribution of the content. Among those that chose an open content licence, most users chose the Creative Commons Attribution licence (Fig. 5).

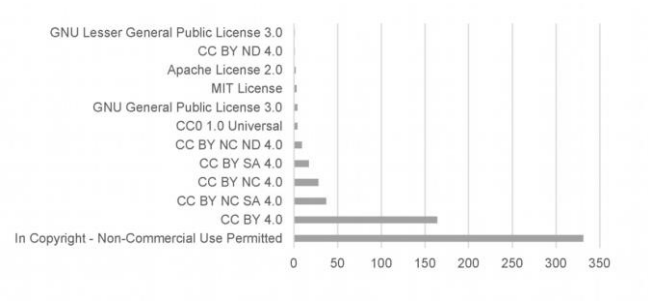


Fig. 5. Rights statements of datasets in the Research Collection.

V. QUALITY assurance AND COMPLIANCE CHECKS

Similar to the experiences of other academic libraries (e.g. Lafferty-Hess et al., 2020), one of the main challenges for the ETH Library when setting up and running a repository for research data has been the definition of the library's responsibility and the scope of its activities when it comes to data curation, quality assurance and compliance monitoring.

Prior to launching the Research Collection, the E-Publishing team at the ETH Library had mainly dealt with quality assurance and copyright compliance of open-access publications and metadata-only records. Working with research data items therefore posed some new challenges for the team, such as how to check the validity of metadata and file formats in research data items and how to deal with research-data-specific risks when it comes to compliance with institutional policies and legal norms. In this chapter, we describe the workflows that are currently in place for dealing with research data items in the repository. However, this a dynamic field and we expect to continuously review and adapt our processes in the coming years as we learn more about user requirements and best practices in data publishing.

A. Quality assurance

The repository has a quality assurance (QA) process in place. When there is a new submission, repository staff have a look at the item and perform some basic checks before they release the files to the public or the designated users. These checks involve various aspects regarding the metadata and files provided by the submitter.

As a first step, the repository staff check whether metadata are consistent with repository rules, correct spelling errors and check whether related publications or datasets are correctly linked to the item. The staff also add formal metadata such as MIME type and size of the dataset to the record. Since QA staff are mostly trained librarians and information professionals – not subject experts – they do not add metadata describing the content of the dataset.

The QA process also involves checking the files the submitter has provided. First, the dataset is downloaded in order to detect potentially virus-infected files. Then, QA staff try to check the readability of the files or a sample of files by opening them with a viewer or another tool. They also run the files through DROID – a tool from the UK National Archives – to perform format identification. Once this step is completed, QA staff check whether the

detected file formats are compatible with the retention period the submitter has chosen. If there are new formats that are not yet recorded in the repository's file format registry, they determine the support level and add them to the registry. Apart from file-format-related checks, staff also check whether the file names and folder structure are comprehensible.

It is important to note that if QA staff detect any problems with the submitted files during these steps in the QA process, they do not edit or manipulate the submitted files. Instead, they contact the researchers that have submitted the datasets and inform them about the potential problems they have detected. Researchers then have the opportunity to make suggested changes to their files and re-upload them.

B. Compliance checks

Another aspect often closely linked to the step of looking into the uploaded files is compliance with policies and legal norms. In principle, compliance with legal norms is the responsibility of the submitter. This is stated explicitly in the Research Collection terms of use. Users also need to confirm in the submission process that they are not violating third-party rights or institutional policies by submitting their dataset.

On the other hand, looking at the uploaded datasets, QA staff regularly detect cases that violate certain policies or legal norms. This is usually not because researchers knowingly decide to violate third-party rights, but rather because they are either not aware that certain norms exist or they simply forgot to delete certain files in their data collection before submitting them.

Observing this discrepancy between what researchers confirm in the submission form and what repository staff see in the submitted datasets, the ETH Library decided that repository staff should inform users if they detect violations of certain norms during the QA process and that they would not release the data to the public before these possible violations were cleared up or resolved. This process is both a service to the researchers but also a risk management measure for ETH Zurich as the institution that hosts the datasets and runs the repository.

Copyright compliance checks involve checking whether a dataset contains third-party copyrighted material and checking files and metadata for licence incompatibilities. When it comes to copyright, what happens on a regular basis is that users include third-party copyrighted material in their data collections without having obtained the copyright owner's permission to publish the material, or that the licence that users choose in the submission form contradicts the licence statement that they have included in their data collection. In order to resolve such potential copyright violations and contradictions, QA staff contact submitters and ask them to delete certain files, obtain permission from copyright holders and/or rethink their licensing choice.

Research data deposited in the Research Collection often contain scripts and software code. As defined in the university's Exploitation Guidelines (ETH Zurich, 2020), all software developed at ETH Zurich and made available to third parties – even under an open-source licence – must be registered with the technology transfer office ETH transfer. Since this policy is not yet well known among ETH researchers, repository staff regularly find software code within the submitted data packages that has not yet been registered with ETH transfer. In these cases, QA staff inform the researchers that they are required to register their code and publish it under an open-source licence.

Disclosure risk is another topic for which repositories must put in place policies and potential mitigation measures. In the Research Collection submission process, researchers have to confirm that they have anonymised all personal data and obtained written consent of the study participants for publication. However, there is still a remaining risk that potentially sensitive, personal data could be released to the public. To mitigate this risk, some specialised, disciplinary repositories have dedicated disclosure risk review workflows in place (see e.g. ICPSR, 2021). However, such a process is not feasible for most institutional repositories such as the Research Collection. The ETH Library has neither the expertise nor trained staff members to perform disclosure risk reviews on datasets. We therefore approach this topic by taking preventative measures. This includes offering training sessions with data protection experts from other ETH units and arranging individual consultation sessions so that researchers working with patient data or other sensitive data can discuss their specific use cases and datasets with a data protection expert before they submit the data to the Research Collection.

C. How to reconcile publishing and preservation requirements

One particular topic that has come up in discussions at the ETH Library about the QA process is the question of how to deal with conflicting requirements regarding file formats coming from users on the one hand and from the library's digital preservation experts on the other hand. The Research Collection itself is not a preservation system but a publication platform. There is a data export process that continuously transfers all data from the Research Collection (DSpace) to the ETH Library's preservation system, the ETH Data Archive (Rosetta). The Data Archive is a dark archive that hosts a copy of all Research Collection data and that can – if needed in the future – perform preservation tasks on the deposited files to keep them readable.

One requirement for being able to perform such preservation tasks is that the Research Collection deliver to the Data Archive only files in formats that are suitable for long-term archiving. This, however, is a non-trivial task to achieve in a research data repository because often research data are produced and stored in file formats not suitable for preservation. As described above, Research Collection staff generally do not edit the researchers' files and therefore also do not convert files to other formats. Implementing a file format conversion service would also require considerable additional human resources at the ETH Library. On the other hand, it is also usually not a top priority for researchers to invest time in file format conversions either. When researchers upload their data to the Research Collection, their priority is usually to have it published sooner rather than later.

At the same time, QA staff have noticed that even if researchers submit their data in non-recommended formats, they still often ask the library to keep these data for an unlimited period of time. Actually, three quarters of all datasets are deposited with the user choosing an indefinite retention period, rather than a limited retention period of 10 or 15 years, with a large part of these deposits coming in non-recommended formats.

Looking at this situation, the library has recently decided that it will change its approach to this topic. Originally, we had assumed that every dataset with an indefinite retention period must only contain files in formats suitable for archiving. However, since we have realised that it is not possible to achieve this in practice, we have decided that we will no longer use the retention period as the main indicator for long-term preservation. Instead, we are now working on implementing a separate checkbox in the submission form where we ask users to indicate if they are actually interested in keeping their data readable over the long term. Only if the submitter activates this checkbox will our team provide recommendations and work with the submitter to help them convert their files into suitable formats. In all other cases – independent of the chosen retention period – we will assume that only bitstream preservation is required.

VI. New developments

The final chapter of this paper discusses two ongoing development projects at the ETH Library that will significantly extend the functionalities of the Research Collection in the coming months.

A. Integration with openBIS

The first of these projects is a collaboration project between the ETH Library and ETH Scientific IT Services. It addresses the need of bringing two currently separate tools together: openBIS as a tool for active research data management (Barillari et al., 2016) and the Research Collection as a tool for data publication. By connecting these two systems, we want to provide an integrated solution for ETH researchers that supports their workflows from active data management to publication and preservation.

Form a user perspective, this integration will provide researchers with a seamless workflow for publishing selected data from their openBIS instance via the Research Collection. The user starts in openBIS and selects the files they want to export to the Research Collection. openBIS then creates a ZIP container that includes some basic metadata about the exported data collection and another ZIP that contains the actual bitstreams. openBIS transfers the main ZIP container to the Research Collection via DSpace's SWORD API (Allinson et al., 2008). This API is a standard feature that comes with all DSpace installations. DSpace then creates a workflow item and presents it to the user. The user can add additional metadata, review the imported bitstreams and select access settings and an end-user licence. At this point, the Research Collection sends the permanent Handle URL of the item back to openBIS, in order to display the link in the user's publications collection. In the Research Collection, as with any other item, QA staff will review the submission and, if accepted, publish it in line with the access settings the user has chosen.

B. Solution for publishing large datasets

The second ongoing development project aims to provide a solution for publishing large datasets. This project was mainly driven by feedback from users who indicated that they need to publish files in the Research Collection that are much larger than what the repository can currently accommodate. At the moment, the maximum recommended file size is 10 GB per individual file and 50 GB as the maximum size for the total amount of files within one item.

When looking into possible technical solutions for this new requirement, we focused on leveraging existing tools within ETH Zurich, rather than setting up a completely new technical infrastructure. The chosen solution will integrate the Research Collection with a separate storage solution for large files based on ownCloud, which has already been in use at ETH Zurich under the brand name polybox.

The overall concept of this solution is that for large datasets, the Research Collection will provide a metadata record that is also used as a landing page for DOI resolution, and this metadata record then links to a download page on ownCloud. ownCloud will be used for data storage and for uploading and downloading the actual files. From a

technical point of view, ETH Zurich's IT Services have made this possible by extending the ownCloud infrastructure used for the polybox service with an additional server called libdrive (Fig. 6). While polybox is a service provided to individual users at ETH Zurich as a drop-box-like storage solution, libdrive is managed by the ETH Library and will be used for implementing the Research Collection large files workflow but also for other use cases within the library requiring transfer or storage of large files.

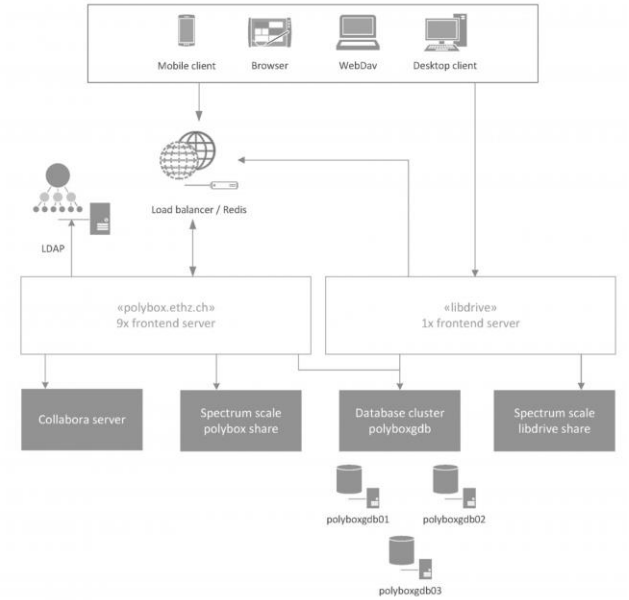


Fig. 6. ownCloud infrastructure including libdrive for publishing large datasets at ETH Zurich.

For all uploads – whether small or large datasets – the users will first go the Research Collection. In the upload form, they can either directly upload their small files to DSpace or request an access link for the upload of large files. Users with large files will then be asked to transfer their files via an ownCloud web client for files between 10 and 20 GB, via a local ownCloud client or WebDav for files approximately between 20 and 200 GB and via an offline USB device if the files are larger than 200 GB. We also expect that there might be some datasets with files that are too large to be downloaded via a browser or client. For these files, we plan to implement an email request form that enables end users to order these files to be sent to them on a USB device (Table 1).

TABLE I. UPLOAD AND DOWNLOAD WORKFLOWS FOR LARGE FILES

ile size	**Upload via**	**Download via**
<10 GB	Research Collection submission form	Browser
10–20 GB	ownCloud web client	Browser
20–200 GB (approx.)	ownCloud client or WebDAV	Browser of client
200 GB–1 TB (approx.)	Offline transfer via USB device	Offline transfer via USB device (request access via email form)

VII. WHAT'S NEXT? PLANS FOR THE FUTURE

For 2021, apart from finishing the two projects described in the previous chapter, one of the main goals for the Research Collection is to complete the application process for certification with the CoreTrustSeal (Dillo & de Leeuw, 2018). We believe that the certification process can help us detect gaps and weak spots in our policies and workflows and that the certification will increase the trust of our user community in the repository.

On the technical side, we are planning to improve the metadata fields used for geo-referencing datasets, so that users can more easily execute geolocation-based searches, and we will work on the integration of the Research Collection in Google Dataset Search.

References

Allinson, J., François, S., & Lewis, S. (2008). SWORD: Simple Web-service Offering Repository Deposit. *Ariadne*, 54, http://www.ariadne.ac.uk/issue54/allinson-et-al/

Barillari, C., Ottoz, D.S.M, Fuentes-Serna, J.M., Chandrasekhar, R., Rinn, B., & Rudolf, M. (2015). openBIS ELN-LIMS: an open-source database for academic laboratories. *Bioinformatics*, 32(4), 638–640. https://doi.org/10.1093/bioinformatics/btv606

DataCite Metadata Working Group (2019). *DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.3.* DataCite e.V. https://doi.org/10.14454/f2wp-s162

Dillo, I. & de Leeuw, L. (2018). CoreTrustSeal. *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare*, 71(1), 162-170. https://doi.org/10.31263/voebm.v71i1.1981

ETH Zurich (2007). *Guidelines for Research Integrity and Good Scientific Practice at the ETH Zurich.* https://rechtssammlung.sp.ethz.ch/Dokumente/414en.pdf

ETH Zurich (2018). *ETH Zurich's open-access policy dated 17 January 2018.* https://rechtssammlung.sp.ethz.ch/Dokumente/134en.pdf

ETH Zürich (2020). *Richtlinien für die wirtschaftliche Verwertung von Forschungsergebnissen an der ETH Zürich.* https://rechtssammlung.sp.ethz.ch/Dokumente/440.4.pdf

Hirschmann, B. (2018). Die Research Collection der ETH Zürich. *ABI Technik*, 38(3), 223–233. https://doi.org/10.1515/abitech-2018-3003

ICPSR (2021). *Data Confidentiality*. https://www.icpsr.umich.edu/web/pages/datamanagement/confidentiality/index.html

Lafferty-Hess, S., Rudder, J., Downey, M. Ives, S., Darragh, J., & Kati, R. (2020). Conceptualizing Data Curation Activities within Two Academic Libraries. *Journal of Librarianship and Scholarly Communication*, 8(1), eP2347. https://doi.org/10.7710/2162-3309.2347

Töwe, M. & Barillari, C. (2020). Who Does What? – Research Data Management at ETH Zurich. *Data Science Journal*, 19(1), 36. https://doi.org/10.5334/dsj-2020-036

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., 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, 160018. https://doi.org/10.1038/sdata.2016.18

Workflow for an Improved FAIR Environmental Data Publication in EnviDat

Ionuț Iosifescu Enescu
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
ionut.iosifescu@wsl.ch

Gian-Kasper Plattner
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
gian-kasper.plattner@wsl.ch

Lucia Espona Pernas
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
lucia.espona@wsl.ch

Dominik Haas-Artho
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
dominik.haas@wsl.ch

Rebecca Kurup Buchholz
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
rebecca.kurup@wsl.ch

David Hanimann
Programme EnviDat
Swiss Federal Research Institute WSL
Birmensdorf, Switzerland
david.hanimann@wsl.ch

Abstract—*The Swiss Federal Institute WSL strives to increasethe fraction of environmental data that is easily available for reuse. With the Environmental Data portal EnviDat, WSL facilitates the publication of FAIR (Findable, Accessible, Interoperable, Reusable) and high-quality environmental research datasets by providing: A) a formal data publication process for the data producers, B) a technical workflow for improving data-quality with automatic validation, interactive quality checks, and iterative improvement of (meta-)data quality in support of the formal publication process and C) a DataCRediT mechanism for declaration of data authorship roles.*

Keywords—EnviDat, FAIR, RDM, (meta-)data quality, DataCRediT, environmental data publication

I. INTRODUCTION

EnviDat is the institutional data portal of the Swiss Federal Research Institute WSL, dedicated to hosting and publishing environmental datasets from forest, landscape, biodiversity, natural hazards and snow and ice research. As graphically summarized in Fig.1, EnviDat offers a range of functionalities and services for publishing data, software, and documentation in support of best practices in Research Data Management (RDM) and Open Science (Iosifescu et al., 2018; Iosifescu et al., 2019). Through its capabilities to host and publish data sets, EnviDat provides unified and managed access to WSL's comprehensive reservoir of environmental research data, and thus actively contributes to the goal of increasing the fraction of environmental data that is easily accessible for reuse by researchers and the public.

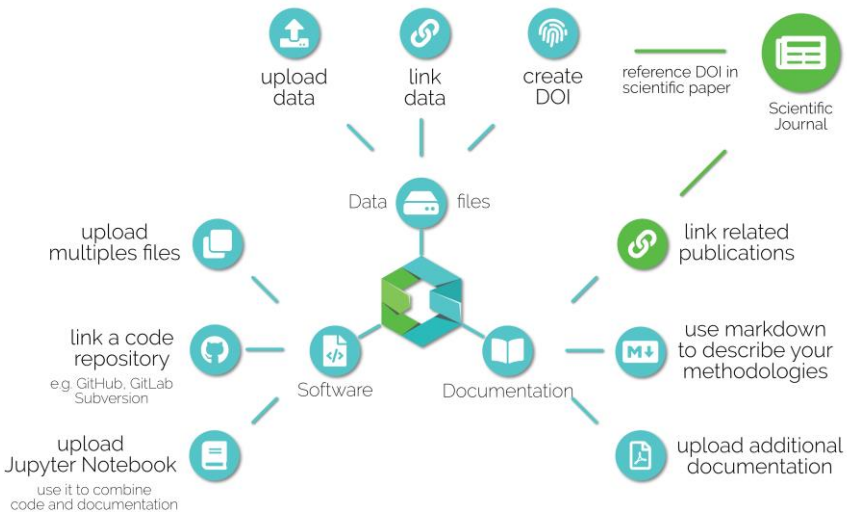


Fig. 1. EnviDat's core functionalities in a nutshell

II. ENVIDAT AS A FAIR PORTAL AND REPOSITORY

EnviDat actively implements the FAIR principles – Findability, Accessibility, Interoperability and Reusability – for scientific data management (Wilkinson et al. 2016). First, an essential requirement for a findable dataset is the assignment of a persistent identifier. The datasets published in EnviDat are assigned globally unique and persistent identifiers (PIDs). All open datasets are also assigned Digital Object Identifiers (DOIs). The minted DOIs are kept in a separate database that is managed independently from the EnviDat main database, in order to protect the persistence of the DOIs in case of technical failures. Moreover, datasets are described by comprehensive metadata records that explicitly include the persistent identifier(s) of the dataset, and present to users a formal citation for the corresponding dataset that includes the assigned DOI. Finally, the metadata of the EnviDat datasets are indexed and made searchable through www.envidat.ch, as well through DataCite Search, ESA's geoportal.org, NASA EarthData Global Change Master Directory (GCMD) and, in the near future, also through the Swiss public administration's central portal for open government data opendata.swiss.

Second, the EnviDat datasets are made accessible via their corresponding metadata landing pages on www.envidat.ch, which are linked to their assigned PIDs/DOIs through the HTTP(S) protocol. EnviDat uses the Open Knowledge Foundation (OKFn) Comprehensive Knowledge Archive Network (CKAN) as a backend metadata repository. CKAN includes a rich Application Programming Interface (API) which allows developers to write code that interacts with CKAN sites and their hosted datasets. The dataset resources are stored using established protocols and providing the HTTPS GET interface for a straightforward download of the data files. Furthermore, even if the metadata are open and accessible by default, the uploaded files and resources can be restricted. In case of restricted resources, interested data users can trigger an access request that will ask for the permission of the data owner, which the data owner must approve before users are allowed to download such restricted datasets. For these cases, EnviDat implemented passwordless authentication procedures, where a unique one-time login token is sent directly to a user's mailbox, thus completely eliminating the need for storing and securing user passwords. Finally, even if data files may be removed at the request of the depositor, the metadata will be kept and the DOIs will continue to point to "tombstone" landing pages containing a modified description that explains the withdrawal reasons. Valid reasons for withdrawal are: violations of WSL research integrity guidelines, proven copyright violation or plagiarism, or legal requirements. Therefore, any metadata registered in EnviDat will persist even if the data should no longer be available, thus avoiding broken links from scientific citations.

Third, the metadata records in EnviDat are organized according to a three-layer schema model (with core, optional and domain-specific research metadata), in order to meet current and future domain-relevant community standards and to ensure interoperability (Iosifescu et. al, 2018). At the core of the EnviDat metadata schema there exist a number of mandatory metadata fields: title, description, keywords, author(s), affiliation, license, publisher, publication year, contact information and the geometry of the spatial extent. Furthermore, a unique EnviDat PID and a DOI are automatically added by the system during the publication workflow as an integral part of the dataset's core metadata record. The EnviDat core metadata schema is designed to make maximal use of the latest DataCite metadata-schema (DataCite, 2019) for DOIs and exploits its ability to store spatial information about environmental measurements. In the EnviDat metadata schema there are also two optional metadata fields, namely "Related Publications" and "Related Datasets", designed to document associated scientific article(s) and other dataset(s) through qualified references/citations. These optional metadata fields improve the documentation of data provenance, since they record and link to additional information influencing the data of interest. The metadata captured with the EnviDat metadata schema is fully interoperable with (and exportable to) various standards such as: Dublin Core, the latest DataCite Metadata Schema (4.3), ISO 19139 or GCMD DIF 10.2 (current operational version). Yet, the interoperability of the data files is a constant concern. Even though EnviDat does not impose any restrictions regarding the format of the data itself, we advise, encourage and even support the users to publish their data in file formats that adhere to existing community standards. EnviDat also fosters new standards, one example of this effort being the Non-Binary Environmental Archive Data (NEAD). This format is being developed as a "generic and intuitive format that combines the self-documenting features of NetCDF with human readable and writeable features of CSV" and it is being specifically "designed for exchange and preservation of time series data in environmental data repositories" (Iosifescu et al., 2020).

Fourth and final, in order to support the reusability of the dataset, the data publishers need to select or define a license that documents the terms of use, thus defining rights, permissions and restrictions of use for the dataset. WSL makes its research data available to users in accordance with the WSL Data Policy, to the extent permitted by the relevant laws, ordinances and contracts with third parties. Any exemptions from the obligation to share research data with users after a clearly specified period must be substantiated and approved by the WSL Directorate. Consequently, EnviDat does not impose any restrictions on "free and open access conditions" conditions – depositors may freely choose to release their data under CC0 (Creative Commons "No Rights Reserved") or

equivalent license. In practice however, the most open licenses preferred by the dataset authors are ODC-ODbL (Open Data Commons Open Database License) or CC-BY-4.0 (Creative Commons Attribution 4.0 International), which both require an appropriate citation/attribution of the dataset.

III. IMPROVING FAIR

EnviDat takes the implementation of the FAIR principles seriously. Unfortunately, however, the FAIR principles do not address an important aspect, namely the quality of the published (meta-)data contents, because publishing a dataset by respecting the FAIR principles is not necessarily correlated to (meta-)data quality. EnviDat thus offers guidance and support to researchers throughout the entire data publication process and strives to achieve the highest quality for environmental research data with an improved FAIR publication workflow.

More precisely, EnviDat enables the publication of highquality environmental research datasets by providing (A) a formal data publication process for the data producers, (B) a technical workflow for improving data-quality in support of the formal publication process and (C) a DataCRediT mechanism for the declaration of data authorship roles.

A. Formal Data Publication Process

The formal data publication process has the role of making the researchers aware of their responsibilities and accountabilities when publishing a research dataset in EnviDat. This process, depicted graphically in Fig. 2, contains the following six main steps:

1. Login. An initial passwordless login in EnviDat registers the email of the researcher. Then the researcher is directed to inform their group leader. The group leader can either direct the researcher to the group's or unit's data manager (if available) or confirms the researchers request to publish data in EnviDat.
2. Receiving editing rights. If the group leader approves, a data manager or the EnviDat support team will grant the necessary rights for data publication and points them to the portal's guidelines and policies.
3. Creation of a "New Dataset" in EnviDat and registration of the necessary metadata according to the EnviDat metadata schema explained in the previous section.
4. Upload of the research data and further resources (e.g., images, software, supporting files etc.). The upload of large files is supported though the provision of individual FTP accounts or, for multi-TB datasets, the provision of individual object store access keys.
5. Publishing the dataset. After finalizing the metadata registration and uploading the research data and possibly other resources, the "Publish" button will become available. This step includes an important workflow for improving data-quality with interactive quality checks and iterative improvement of (meta-) data quality in support of the formal publication process, that will be explained and detailed in the next heading.
6. Curation of the research dataset. Since the researcher's responsibility does not end with the data publication, the data owners are encouraged to periodically revisit and improve the published dataset and the associated metadata. For example, if the published research data is linked to a scientific publication that is in review, then the researchers are asked to enter the final publication in the "Related Publications" field of the EnviDat metadata form. Also, in EnviDat the researchers have the possibility to perform corrections and additions to the published data by uploading new versions.

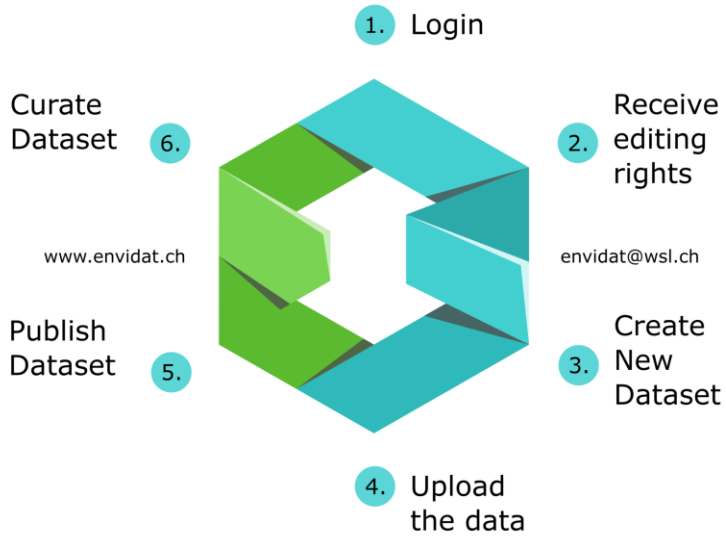


Fig. 2. Formal research data publication process in EnviDa

On one hand, EnviDat's formal publication process verifies that the group leaders are informed of the data publishing efforts by their supervised researchers. On the other hand, the researchers will get in touch with the EnviDat support team, which points them to important EnviDat guidelines and policies before receiving editing rights. For example, the researchers are made aware that the published metadata (but not necessarily the data files) will become public domain, therefore it can be re-used in any medium for any purpose and without prior permission. (although, in the EnviDat policies we request metadata harvesters, if technically feasible, to provide a link to the original EnviDat metadata record).

Furthermore, by reading the EnviDat guidelines and policies, researchers are also made aware that the validity, authenticity and quality of the content of submissions is their responsibility, hence they should only submit metadata and content items for which they have the necessary permissions and rights for distribution and publication. Copyright violations related to the submission of metadata and content items to EnviDat are the responsibility of the depositors.

Finally, the other steps of the process pertaining to how create a new EnviDat dataset record, how to upload the data, and how to publish and curate the dataset are detailed in the EnviDat's guidelines for data publication, with the most upto- date version available on www.envidat.ch.

B. Workflow for an impoved (meta-)data quality

The publishing of the dataset represents an important step in EnviDat. In order to improve the quality-assurance for data sets, we aim for introducing an approach that is similar to the peer-review process applied for scholarly articles, a process that is currently missing for research data publication. For this reason, the EnviDat team encourages nomination of EnviDat data managers by every data provider organization. Data managers can peer-review the (meta)data with regard to quality characteristics such as accuracy, completeness, reliability, relevance, and timeliness. EnviDat has extended this data publishing process step to be more than just a simple assignment of a DOI by chaining together researcher input, automatic validation, interactive quality checks, and iterative improvement of (meta-)data quality. During the quality assurance workflow, the request for a DOI is simply only one substep of the workflow which brings the dataset through an approval process with a double-checking principle. The publication workflow ends with the submission of the metadata-record to DataCite and the final publication of the metadata record in EnviDat

During this workflow, as depicted in Fig. 3, the dataset itself moves between different states, from "Unpublished" towards "Published", "Pending" and the "Approved" states.

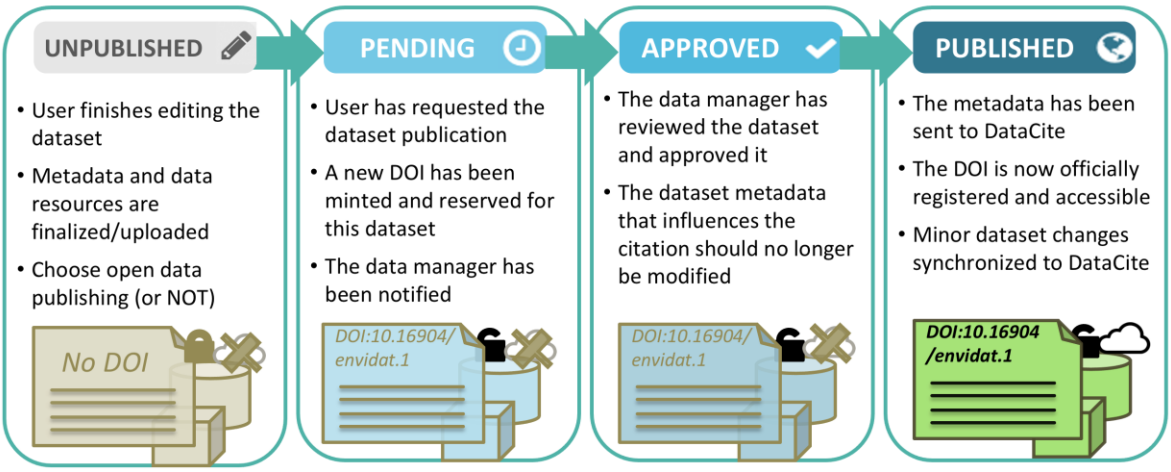


Fig. 3. EnviDat's quality assurance workflow

The workflow can be viewed as a decentralized peerreview and quality improvement process for safeguarding the quality of published environmental datasets. This workflow is being further developed and refined together with partner institutions within the ETH Domain on regular basis, with an especially strong cooperation regarding concepts and software existing between WSL and the Swiss Federal Research Institute for Aquatic Science and Technology Eawag (von Waldow and Iosifescu, 2020).

C. DataCRediT

The overall workflow for an improved FAIR data publication in EnviDat is further improved by increasing the transparency for the range of contributions that a dataset author's make to the published data. Therefore, the EnviDat metadata schema has fields designed to capture and document the individual author contributions to the publication of a particular dataset and related contents such as the software that was used to process or generate the data set. These fields are documented in The Data Authorship Contributor Roles Taxonomy – DataCRediT (WSL, 2018), a mechanism for data authorship specification inspired by and adapted from the Contributor Roles Taxonomy (CRediT) for scientific scholarly output developed by the Consortia Advancing Standards in Research Administration Information (CASRAI, 2018). DataCRediT currently covers six contributor roles: Collection, Validation, Curation, Software, Publication, and Supervision, as detailed in Fig. 4.

The taxonomy supports transparency of contributions to published research data sets by providing an improved system of attribution, credit, and accountability for scientific data publication, thus further encouraging the vigilant application of the FAIR data principles by the individual researchers.

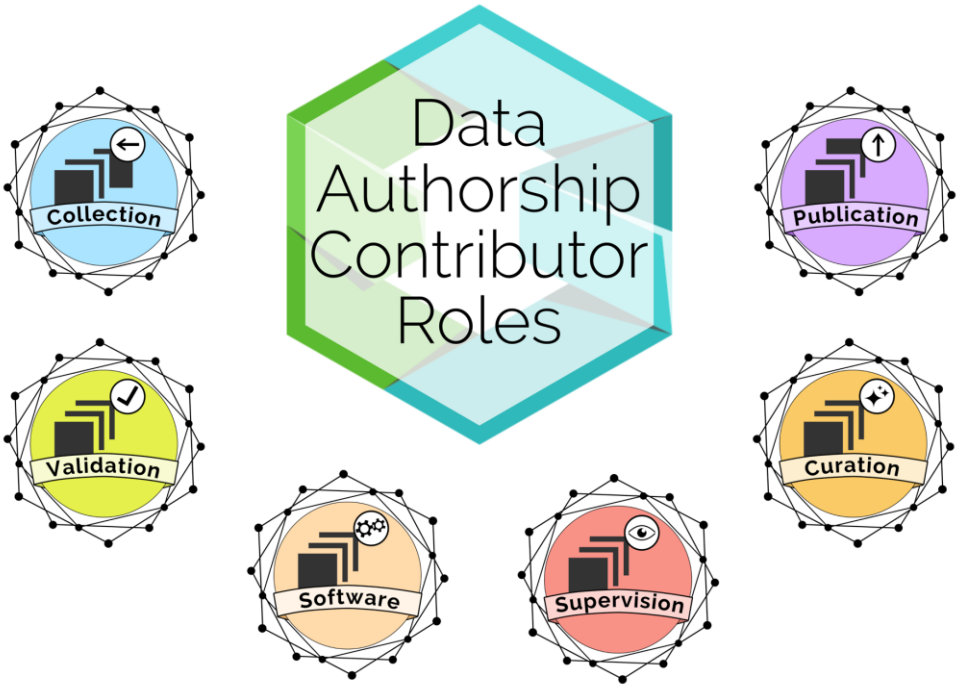


Fig. 4. The Data Authorship Contributor Roles Taxonomy (DataCRediT)

IV. CONCLUSION

The FAIR data publication workflow can be greatlyimproved by implementing basic quality assurance. Since publishing data comes with significant restrictions – like not being allowed to delete any parts of the already published data – it is important to ensure that the original data is of the highest possible quality from the beginning. This can be achieved with: A) a formal data publication process, B) iterative improvement of (meta-)data quality through an approval workflow with a double-checking principle, and C) an improved system of attribution, credit, and accountability forscientific data publication.

DEDICATION AND ACKNOWLEDGMENTS

The Environmental Data Portal EnviDat was initiated by Prof. Dr. Konrad Steffen, the former WSL director, who died in 2020 during field work in Greenland. We dedicate this article to him, to acknowledge and honor his crucial role for EnviDat. His vision was that EnviDat will facilitate the work of researchers by supporting them with the publication of their data, thus creating new collaboration opportunities within WSL, the ETH Domain and beyond.

REFERENCE

CASRAI (2018). CRediT – Contributor Roles Taxonomy. http://docs.casrai.org/CRediT – last accessed on February 15, 2021

DataCite Metadata Working Group (2019). DataCite Metadata Schema for the Publication and Citation of Research Data. Version 4.3. DataCite e.V. https://doi.org/10.14454/f2wp-s162

Iosifescu Enescu, I., Plattner, G. K., Bont, L., Fraefel, M., Meile, R., Kramer, T., Pernas, L. E., Haas-Artho, D., Hägeli, M. and Steffen, K. (2019). Open science, knowledge sharing and reproducibility as drivers for the adoption of FOSS4G in environmental research. In M. A. Brovelli & A. F. Marin (Eds.), International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: Vol. XLII-4/W14. FOSS4G 2019 – Academic Track (pp. 107-110).https://doi.org/10.5194/isprs-archives-XLII-4-W14-107-2019

Iosifescu Enescu, I., Plattner, G. K., Pernas, L. E., Haas- Artho, D., Bischof, S., Lehning, M., and Steffen, K. (2018). The EnviDat concept for an institutional environmental data portal. Data Science Journal, 17, 28 (17 pp.). https://doi.org/10.5334/dsj-2018-028

Iosifescu Enescu, I., Bavay, M. and Mankoff, K. (2020). Non-Binary Environmental Data Archive (NEAD) format. EnviDat. https://doi.org/10.16904/envidat.187

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

WSL (2018) DataCRediT - Contributor Roles Taxonomy for Data. https://www.wsl.ch/datacredit/#feat – last accessed on February 15, 2021

von Waldow, H. and Iosifescu Enescu, I. (2020). (Meta)Data Quality & Logistics: the FAIR Data Publication Workflow at Eawag and WSL. Presentation at the Swiss Research Data Day 2020 (online), Switzerland. https://mediaserver.unige.ch/play/137206 – last accessed on February 15, 2021

University of Lausanne's Open Science Strategy and Action Plan

Bagnoud Gérard
Information Resources and Archives Service (UNIRIS)
University of Lausanne
Lausanne, Switzerland
ORCID: 0000-0002-4403-7919

Crespo-Quesada Micaela
Department of Research, International Relations and Continuing Education
University of Lausanne
Lausanne, Switzerland
ORCID: 0000-0003-4129-1601

Jambé Carmen
Information Resources and Archives Service (UNIRIS)
University of Lausanne
Lausanne, Switzerland
ORCID: 0000-0001-5399-3948

Abstract*—Developed in 2019, University of Lausanne (UNIL)'s Open Science Strategy revolves around the free access to scientific publications (Open Access) and the opening of research data (Open research Data) in order to rise up to the challenge of disseminating knowledge. It aligns with several other recent institutional and cantonal policies and strategies. In order to implement it, UNIL has defined an action plan based on 5 priority areas: Governance; Organization; Infrastructures; Training and advice; New culture and communication. Each of these 5 axes is broken down into concrete measures to be carried out by different stakeholders. Completed and ongoing projects include: the creation of a new Open Science web portal; the elaboration of an UNIL Data Management Plan template with the tool DMPonline; the development of a new directive on Open Access and on Research data; the institutional repository for publications (SERVAL); the creation of a custom wizard for Open Access (Papago); the provision of an Open Access editing and publishing portal; the development of support and training for researchers.*

Keywords — Open Science, Open Access, Data Curation, Data Management Plan, FAIR Principles, Research Data Lifecycle, Research Data Management, Open Research Data.

I. The University of Lausanne in today's research landscape

The University of Lausanne (UNIL) Rectorate's 2017-2021 plan of intent [1] states that:

"UNIL researchers and teachers devote a significant part of their time to their research activities and the funds invested in them are considerable. However, the visibility of the results of this research cannot be taken for granted. It depends essentially on the motivation of their authors to publicize them, beyond their traditional publication in scientific journals, books or conference proceedings. Today's research is open, participatory and transdisciplinary. […]

The Rectorate of the University of Lausanne intends to adopt a very clear promotion policy in favour of openness, both for publications (Open Access) and for research data (Open Data). This policy of openness must be carried out in collaboration with editorial partners, UNIL researchers, [...] as well as national partners [...], the political world, the research community or the Consortium of Swiss University Libraries."

The UNIL's Open Science (OS) strategy links to UNIL's 2019 digital strategy [2] as well as the digital strategy of the *Canton de Vaud* [3] drawn up by the *Conseil d'État* (executive body). This OS strategy is also intended to respond to the 2017-2022 state legislative program [4], which aims to promote open and participatory science, as well as the University of Lausanne's strategic plan [5] adopted by the *Grand Conseil* in May 2019.

The new paradigm of Open Science is transforming the research environment and the way researchers do and share science. In an era of digitization, citizen science and "fake news", Open Science offers an opportunity for a verifiable, reproducible, closer to the citizens, and overall fairer Science.

II. *Open Science:* an open and free Science

A. Accessible and high-quality scientific knowledge

Open Access to scientific knowledge and research results has the potential to improve the quality of science by making it more transparent, more integrated, more responsive to societal challenges, more inclusive and more accessible to new users.

The Amsterdam Call for Action on Open Science [6], based on the reflections of many experts gathered in 2016 by the Dutch Presidency of the EU, defines Open Science as

"Open Science is about the way researchers work, collaborate, interact, share resources and disseminate results. A systemic change towards open science is driven by new technologies and data, the increasing demand in society to address the societal challenges of our times and the readiness of citizens to participate in research."

Several factors must be considered in order to successfully open up science: integration at all levels and in all aspects of current practices; taking into account the different disciplinary sensitivities to ensure transparency throughout the system; setting up administrative and financial support to minimise the administrative impact on researchers, etc.

B. *Binding national and international rules*

On a global scale, and particularly on a European scale, it can be observed that transparent research data management – Open Research Data (ORD) – has become a priority for both states and public funding bodies. They require the preparation of Data Management Plans (DMPs), as well as free access to scientific publications and its underlying data. On the other hand, an increasing number of scientific publishers now have data policies which request access to data, metadata, codes, materials, methods and protocols associated with both qualitative and quantitative research results.

In Switzerland, the Swiss National Science Foundation (SNSF) has been committed to opening up science since 2006[27]. Beneficiaries of SNSF grants are required to submit a DMP with all funding applications since 2017, as well as to open all scientific works resulting from the projects the SNSF funds [7]. Furthermore, swissuniversities aims for 100% open access for all scholarly publications from 2024 onwards, in line with its national Open Access strategy [8].

Since November 2015, UNIL has been a signatory of the LERU Statement on Open Access to Research Publications [9], which aims to promote opening publications, archiving and the availability of scientific data. UNIL also signed in 2018 the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (Berlin Declaration) [10] and the San Francisco Declaration on Research Evaluation (DORA) [11], which questions the widespread use of bibliometric rankings as metrics for the evaluation of research and researchers.

C. *The benefits of Open Science*

Many benefits can be expected when science becomes Open:

- **Transparency and visibility**: Open science is synonymous with honest, accountable, transparent, reproducible, valid and good research. The visibility of researchers and universities is increased as open data and publications are more downloaded, read and shared.
- **Impact and new discoveries**: through its greater visibility, Open Science enables higher impact: many studies show that the citation rate increases when data and publications are open. The circulation of knowledge is also improved, thus fostering innovation and the development of new knowledge.
- **Democratization of knowledge**: access to knowledge is a universal right. Open Science reduces the gaps between states, institutions and citizens. It defends free and open access to knowledge and opposes any discrimination based on financial criteria.
- **Public funding = public good**: most of the research carried out at UNIL is financed by public funds and, consequently, by citizens. The data, publications and research results thus obtained are a public good and must therefore be accessible quickly and freely for the benefit of society.

Open Science is also a way to restore trust between citizens and the science they fund while strengthening its integrity.

D. *The two priorities of Open Science for UNIL: Open Access and Open research Data*

OS is an umbrella term that includes a wide variety of initiatives and movements. UNIL intends to focus on the challenges of disseminating knowledge by developing its Open Science approach mainly around Open Access to scientific publications and Open research Data, while integrating components from the other pillars of Open Science.

III. Open Science Action Plan 2019-2021

As part of its reflections, its surveys of researchers and faculty and the involvement of stakeholders within faculties and services, UNIL has defined a plan of action in 5 priority areas ("Fig. 1.") which are broken down into specific objectives, concrete measures to be carried out and expected deliverables:

[27] http://www.snf.ch/fr/pointrecherche/dossiers/open-science/Pages/default.aspx

1. **Governance**: to develop an Open Science policy, strategy, processes and guidelines to support UNIL's vision.
2. **Organization**: to set up administrative and support structures that ensure a participatory and inclusive approach to researchers.
3. **Infrastructures and tools**: to provide the technical means to manage, store, secure, share and archive scientific information.
4. **Training and advice**: to support, accompany and empower researchers in the management of their projects
5. **A new culture and communication**: to raise awareness in the community and the public about the challenges and opportunities of OS.

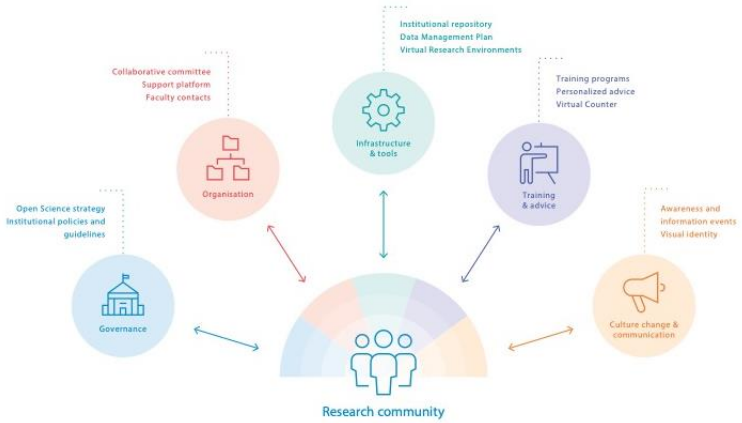


Fig. 1. UNIL's Open Access Strategy and Action plan 2019 – 2021. Five priority axes to support UNIL researchers

IV. Open Access - open access to scientific publications

A. An institutional approach that guarantees academic freedom

The results of a survey conducted in 2017 [12] show a good predisposition of the UNIL academic community towards Open Access, highlighting a desire to democratize knowledge and a concern for financial considerations.

Given the richness and variety of UNIL's disciplinary fields, a unique approach to Open Access that favours one path over another could never succeed. UNIL has the academic freedom of its researchers at heart and thus wants to develop a mixed and pragmatic approach where the golden and green roads coexist and complement each other. Researchers will firstly choose the journal or editor best suited to their case based on scientific criteria, and they will then be able to choose which path to follow to make their publication freely accessible.

B. Challenges of Open Access for UNIL

This mixed strategy requires the improvement of SERVAL (**SERV**eur **A**cadémique **L**ausannois), which is the institutional repository at UNIL and CHUV. Over the past two years, SERVAL has undergone a substantial optimisation to become a tool focused on the needs of researchers and the current challenges of Open Access publication: ease of use, internationalization of Lausanne-based research, visibility of scientific works, citations of UNIL researchers, etc.

As for the publication of monographs, the path is yet to be paved. UNIL will develop its policy in partnership with the research community and stakeholders, including publishers, historical partners in the promotion of scientific research.

Solutions acceptable to all parties will still have to be found, considering the requirements of funding bodies, the national strategy, the needs of researchers and the institutional challenges of a public university, which must reach beyond cantonal and national borders through the quality of its research and teaching.

UNIL therefore gives priority to:

- Enabling its researchers to **focus on their research** and not in their administrative or other obligations;
- **Supporting** its researchers and **communicating** the opportunities offered by Open Access;
- **Training of** the community in new complementary modes of scientific publication (bibliodiversity);
- The **development of** technical **infrastructures** and the provision of **tools for publishing** journals in Open Access;

- The consideration, in the context of the researchers' **evaluations**, of visibility and access to scientific results efforts (SERVAL and Open Access).

This reasonable and thoughtful approach should allow us to meet the challenges of OA and scientific communication landscape of the beginning of the 21st century.

C. Examples of projects and initiatives launched in the field of Open Access

We have launched and completed several Open Access initiatives within each of the five axes of our Open Science strategy (see above):

1. Governance

- Directive 4.6 on Open Access and the use of our institutional repository, SERVAL [13]. The Directive entered into force on 1 July 2020 and a dedicated webpage[28] was created to explain in a clear language what is expected from our researchers, especially for our English-speaking community, since there is no official text in English.

2. Organization

- UNIL is a highly decentralised organisation. Disciplinary specificities are abundant and federating all faculties under a purely central organism would prove inefficient. Therefore, central initiatives stemming from the Department of Research, International Relations and Continuing Education are consulted and vetted at each faculty through their research consultants. Research consultants represent an invaluable two-sided communication channel through which institutional initiatives can be distilled to each faculty and through which disciplinary specificities can be filtered.

3. Infrastructure

- A vast project was launched to integrate researcher-centred features in our institutional repository: SERVAL. Giving ownership of the records to researchers was a first step to empower them to take possession of their profiles and to make sure the information is correct. Furthermore, the user interface and workflow were drastically simplified and the visibility of SERVAL's records was greatly improved, with approximately a 230% increase in downloads from January 2018 to December 2020.
- A publication platform project was launched in order to provide internally-edited journals with the infrastructure needed (based on Open Journal Systems open source software[29]) to streamline publication and increase visibility of their journals, EdiPub. For the time being, three journals started to collaborate in our platform, and an inter-university initiative to extend the scope and create a community around it is being discussed.
- In response to the seemingly infinite number of possible cases with respect to Open Access obligations, rights and financial possibilities, we developed, in collaboration with the University of Fribourg, Papago[30], a personal Open Access assistant capable of giving personalised advice on Open Access rights, obligations and financial opportunities based on a few simple questions. The code to Papago is openly available on GitHub[31] and it can be personalised to include each Swiss higher education institution logo and specific institutional information.

4. Training

- Several schemes, checklist and fact sheets were developed in order to guide and simplify the ever-growing requirements and options concerning Open Access. Some examples are "The roads to Open Access" [14] and "How to make your research available to everyone" [15].

[28] https://www.unil.ch/openscience/en/home/menuinst/open-access/open-access-a-lunil/directive-open-access-unil-serval.html
[29] https://pkp.sfu.ca/ojs/
[30] https://www.unil.ch/openscience/en/home/menuinst/open-access/papago---your-open-access-personal-assistant.html
[31] https://github.com/micaelacq/Papago

- A self-paced course was prepared and launched on Moodle[32] in order to introduce the subject of Open Access and its legal implications. A module is also devoted to SERVAL and its use. More than eighty members of UNIL have already enrolled to the course.
- Several short video tutorials[33] are available in French and English to quickly demonstrate how some frequent operations are performed in SERVAL.
- A simplified legal guide [16] was prepared to introduce researchers into copyright and how it applies to scientific publications and Open Access publishing.

5. New research culture and communication

- For the past two years, we prepared and shared an Open Science advent calendar[34] showing snippets of Open Access or Open Research Data practices, both general and specific for UNIL. The 2020 advent calendar can be found in our website[35].
- We developed a Jeopardy!-style game to introduce the subject of Open Access at one of UNIL's summer schools. The Open Access session, rather than purely passive and academic, was introduced with a very short theoretical introduction, followed by a quiz round where students discovered by themselves the key concepts of OA publishing. The session was very much appreciated and students seemed to have interiorised the concepts thanks to the hands-on experience. The game will be openly shared so that others can build upon it.
- A large and diverse number of events have been organised at UNIL around the theme of Open Access and Open Science, including the first national conference in Open Access in 2018[36]. Our latest event was the Open Science Tour of UNIL[37] which, given the covid circumstances, was held completely online.

V. Open research Data - transparent and reasonable data management

D. An open and responsible institutional approach

UNIL's research data strategy is defined within a binding international and national framework. It is also based on the needs and expectations of its community as identified in a survey conducted in 2015 [17].

In this context, UNIL advocates honest and responsible research. This approach aims to manage research data in a transparent and open manner, within the limits of the law and scientific requirements in terms of ethics, professional conduct and compliance with standards for the protection of individuals and intellectual property.

E. Challenges of Open research Data for UNIL

Research data derived from scholarly work is a public good whose management – in the short, medium and long term – raises many scientific, ethical, deontological, legal, technical, economic and societal issues. Proper data management is essential and crucial in many respects: it ensures compliance with legal and regulatory frameworks as well as with the requirements of scientific funders and publishers. It also guarantees the authenticity, integrity, reliability and usability of data as well as facilitating its reproducibility, sharing and reuse. Finally, it makes research results more visible [21] and contributes to their quality.

These challenges and their complexity require a high number of skills that must first be identified and then reinforced to assist researchers. Additionally, disciplinary specificities must be carefully considered, as well as a number of obligations that arise from today's research environment.

To meet these multiple challenges, UNIL focuses its interventions and support on the following areas:

- **Awareness** and **communication of** this "new" scientific culture;
- The support and **training** of its researchers in the face of this evolution;
- The **development** of technical **infrastructure**;

[32] https://moodle.unil.ch/course/view.php?id=14160
[33] https://www.unil.ch/openscience/en/home/menuinst/open-access/serval/how-to-use-serval.html
[34] https://www.unil.ch/openscience/home/menuguid/evenements/decembre-2020---calendrier-de-lavent-open-science-2020.html
[35] https://www.unil.ch/openscience/en/home/menuguid/evenements/decembre-2020---calendrier-de-lavent-open-science-2020.html
[36] https://www.unil.ch/openscience/en/home/menuguid/evenements/octobre-2018---conference-nationale-open-access.html
[37] https://www.unil.ch/openscience/en/home/menuguid/evenements/novembre-2020---le-tour-open-science-de-lunil.html

- The establishment of a **participatory organization** and **governance** capable of effectively meeting the needs of its community.

This multivariate approach must make it possible to respond to the challenges and issues of OrD so that the management *of* research data becomes a responsible management of data *for* research.

F. Examples of projects and initiatives launched in the field of Open research Data

We have launched and completed several Research Data Management (RDM) and OrD initiatives within each of the five axes of our Open Science strategy:

1. Governance

- Directive 4.5 on Research Data Processing and Management [18]. Adopted on June 11, 2019 (with retroactive effect to January 1, 2019), this Directive regulates the processing, storage, archiving and long-term preservation of research data. It also a) defines the rules for the management of research data; b) sets out the roles and responsibilities of the various stakeholders; c) defines the pricing principles for the use of the resources of the Computing and Research Support Division (DCSR) of IT Service.

2. Organization

- The management of research data raises many issues and requires multiple professional skills. The OrD UNIL team works closely in a participatory manner with all the actors involved, both internal partners – Central Services (e.g. Dep. of Research, IT Service, Ethics team, legal Service) and Faculties (Research consultants and discipline specific research groups) – and external partners (e.g. BiUM[38], DaSCH[39], dhCenter[40], FORS[41]). The goal of this OrD network is to support UNIL's researchers in managing research data throughout their lifecycle [19] with best practices and implement FAIR Data Principles during research.

3. Infrastructure

- Since 2019, the UNIL IT Service has integrated a new Division, the DCSR[42] whose mission is to provide UNIL with computing and storage resources, as well as expertise that is transversal to its faculties and institutes. This field of expertise includes high-performance computing (HPC) support, (sensitive or non-sensitive) data storage as well as web and database development mandates.
- To help its community of researchers write their DMPs, UNIRIS has developed a generic template in the tool DMPonline[43] which provides help and examples to facilitate the DMP's redaction. In order to help researchers in using the tool, an interactive tutorial has been developed[44]. In the future, different discipline-specific templates can be developed in collaboration with the faculties.
- UNIL, in collaboration with the University of Zurich, is actively participating in the SWISSUbase[45] project led by FORS, which aims to develop a general, non-commercial, open and sustainable data repository to comply with the *FAIR Data Principles* [20].

4. Training

- In partnership with the Graduate Campus[46], different workshops on good practices in RDM – RDM's basis; DMPs; Data Organisation; Data Storage & Security; Data Sharing (OrD); Data Archiving, etc. – are offered to PhDs and post-doctoral students during the academic year. In order to broaden the audience and reach all researchers (e.g. seniors), a course catalogue online platform[47] has been created.

---

[38] https://www.bium.ch/
[39] https://dasch.swiss/
[40] https://dhcenter-unil-epfl.com/
[41] https://forscenter.ch/
[42] https://unil.ch/ci/dcsr
[43] https://dmp.unil.ch
[44] https://www.unil.ch/openscience/home/menuinst/formations/dmponline-unil-en.html
[45] https://info.swissubase.ch/
[46] https://www.unil.ch/graduatecampus/fr/home.html
[47] https://conference.unil.ch/cours/openscience

- In the near future, the OrD website will host online self-study tutorials (like the one on DMPonline) following best practices in research data management and topics – DMP; Organize your Data; Storage & security; Preservation & sharing – provided on the main web page "How to manage your data?"[48].

5. New research culture and communication

- In collaboration with the OA Team, we co-organise various events and OS awareness campaigns (e.g. OS advent calendar, OS Tour) for the research community.
- For events related to RDM and OrD, we appreciate organising them in collaboration with the OrD UNIL team. We also like to include UNIL's researchers so that they can share issues they encounter in RDM or OrD with their colleagues.

IV. Conclusion

As mentioned before, Open Science offers many opportunities and challenges, not only for the research community but for society in general. In the years to come, UNIL will continue to promote OS and other components of this broad "spectrum of openness" (e.g. Citizen Science, Open educational resources, Open methodologies, etc.) can be integrated into the strategy developed and actions carried out.

A. Abbreviations and Acronyms

Bibliothèque universitaire de médecine (BiUM); Centre hospitalier universitaire vaudois (CHUV); Data and Service Center for the Humanities (DaSCH); Data Management Plan (DMP); Digital studies interface for the arts, humanities and social sciences (dhCenter); Division Calcul et Soutien à la Recherche (DCSR); Findable, Accessible, Interoperable, Accessible (FAIR); Swiss Competence Center for Social Sciences (FORS); High-Performance Computing (HPC); Open Access (OA); Open research Data (OrD); University of Lausanne (UNIL); Service des ressources informationnelles et archives de l'Université de Lausanne (UNIRIS); Serveur Académique Lausannois (SERVAL); Swiss National Science Foundation (SNSF).

References

[1] Université de Lausanne. (2017). *Plan d'intentions de l'Université de Lausanne 2017-2021*. Consulted at https://www.unil.ch/central/files/live/sites/central/files/docs/plan-intentions-unil-1721.pdf

[2] Université de Lausanne. (2019). *Stratégie numérique de l'UNIL et plan de réalisation*. Consulted at https://www.unil.ch/numerique/files/live/sites/numerique/files/Documents/Strategie_numerique_v20191014.pdf

[3] Conseil d'État du Canton de Vaud. (2018). *Stratégie numérique du Canton de Vaud*. Consulted at https://www.vd.ch/fileadmin/user_upload/accueil/ConferencePresse/StrategieNumVD2018.pdf

[4] Conseil d'État du Canton de Vaud. (2017). *Programme de législature 2017-2022*. Consulted at https://www.vd.ch/fileadmin/user_upload/organisation/ce/fichiers_pdf/prog-leg_2017-2022-final-numerique.pdf

[5] Canton de Vaud. (2017). *Exposé des motifs et projet de décret sur le plan stratégique pluriannuel 2017-2022 de l'Université de Lausanne*. Consulted at https://www.unil.ch/central/files/live/sites/central/files/docs/plan-strat-unil-1722.pdf

[6] European Union. (2016). *Amsterdam Call for Action on Open Science*. Consulted at file:///Users/cjambe/Downloads/amsterdam-call-for-action-on-open-science-1.pdf

[7] Swiss National Science Foundation. (2015). *Funding regulations. Article 47: Publication and accessibility of research results*. Consulted at http://www.snf.ch/en/funding/documents-downloads/Pages/regulations-funding-regulations.aspx#br_a_47

[8] Swissuniversities. (2017). *Swiss National Strategy on Open Access*. Consulted at https://www.swissuniversities.ch/fileadmin/swissuniversities/Dokumente/Hochschulpolitik/Open_Access/Open_Access_strategy_final_e.pdf

[9] League of European Universities. (2012). *LERU Statement on Open Access to Research Publications*. Consulted at https://www.leru.org/files/LERU-Statement-on-Open-Access-to-Research-Publications-Full-paper.pdf

[48] https://www.unil.ch/openscience/en/home/menuinst/open-research-data/gerer-ses-donnees-de-recherche.html

[10] Max Planck Gesellschaft. (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. Consulted at https://openaccess.mpg.de/67605/berlin_declaration_engl.pdf

[11] DORA. (2013). *San Francisco Declaration on Research Assessment*. Consulted at https://sfdora.org/read/

[12] Crespo-Quesada, M. and Bussy, F. (2018). *L'Open Access à l'UNIL, Sondage Open Access UNIL 2017 - Rapport et Vision*. Consulted at https://serval.unil.ch/fr/notice/serval:BIB_6CF16A34F7E0

[13] Université de Lausanne. (2020). *Directive de la Direction 4.6 Dépôt et diffusion des publications scientifiques dans le serveur institutionnel de l'Université de Lausanne, SERVAL*. Consulted at https://www.unil.ch/central/files/live/sites/central/files/textes-leg/4-rech/dir4-6-serval.pdf

[14] Crespo-Quesada, M. (2020). *The roads to Open Access*. Consulted at https://serval.unil.ch/en/notice/serval:BIB_FB6B8BA8A4E3

[15] Crespo-Quesada, M. (2018). *How to make your research available to everyone*. Consulted at https://serval.unil.ch/en/notice/serval:BIB_C88C41000F23

[16] Crespo-Quesada, M. and Takeuchi, T. (2019). *Open Access and Copyright*. Consulted at https://serval.unil.ch/fr/notice/serval:BIB_505C9CCA1E2B

[17] Jambé, C. (2015). *La gestion des données de recherche à l'Université de Lausanne : enjeux transdisciplinaires*. Consulted at http://doc.rero.ch/record/258023

[18] Université de Lausanne. (2019). *Directive 4.5 Traitement et gestion des données de recherche*. Consulted at https://www.unil.ch/central/files/live/sites/central/files/textes-leg/4-rech/dir4-5-donnees-rech2.pdf

[19] UK Data Service. (2012). *Research data lifecycle*. Consulted at https://www.ukdataservice.ac.uk/manage-data/lifecycle.aspx

[20] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). *The FAIR Guiding Principles for scientific data management and stewardship*. Consulted at https://doi.org/10.1038/sdata.2016.18

[21] Piwowar, H. A. and Vision, T. J. (2013). *Data reuse and the open data citation advantage*. Consulted at https://doi.org/10.7717/peerj.175

Data integration in systems genetics and aging research

Alexis Rapin
Laboratory of Integrative Systems Physiology
EPFL
Lausanne, Switzerland
alexis.rapin@epfl.ch,
https://orcid.org/0000-0003-3448-5459

Maroun Bou Sleiman
Laboratory of Integrative Systems Physiology
EPFL
Lausanne, Switzerland
maroun.bousleiman@epfl.ch
https://orcid.org/0000-0002-1375-7577

Johan Auwerx
Laboratory of Integrative Systems Physiology
EPFL
Lausanne, Switzerland
johan.auwerx@epfl.ch
https://orcid.org/0000-0002-5065-5393

Abstract*—*Human life expectancy has dramatically improved over the course of the last century. Although this reflects a global improvement in sanitation and medical care, this also implies that more people suffer from diseases that typically manifest later in life, like Alzheimer and atherosclerosis. Increasing **healthspan** by delaying or reverting the development of these age-related diseases has therefore become an urgent challenge in biomedical research. Research in this field is complicated by the multi-factorial nature of age-related diseases. They are rooted in complex physiological mechanisms impacted by heritable, environment and life-style factors that can be unique to each individual. Although technological advances in high-throughput biomolecular assays have enabled researchers to investigate individual physiology at the molecular level, integrating information about its different components, and accounting for individual variations remains a challenge. We are using a large collection of "omics" and phenotype data derived from the **BXD mouse genetic diversity panel** to explore how good data management practices, as fostered by the **FAIR principles**, paired with an **explainable artificial intelligence** framework, can provide solutions to decipher the complex roots of age-related diseases. These developments will help to propose innovative approaches to extend healthspan in the aging global population.

Keywords—data integration, systems genetics, metabolism, aging, XAI, ML, graphs, omics.

I. Introduction

Age-related diseases are associated to a number of heritable, environmental and life-style factors that impact physiology. In this context, the **systems genetics** community investigates the links between genetics, metabolism and individuals' traits to discover underlying molecular mechanisms, which could be utilized to design therapies and treatments. In the context of aging, the aim would be to reverse its effect and delay the development of its associated disorders. Systems genetics leverages the high-throughput capacity of "omics" technologies coupled with the sample availability and controlled experimental conditions offered by model organisms to assess biomolecular mechanisms in cells and tissues. While generating enormous amount of biomolecular data, the research community generally focuses on assessing the links between pairs of biological layers, such as associations between genetics and phenotypes, or between genetics and gene expression. Although this approach allows to reveal associations between individual factors, it hardly addresses interactions between more than two factors and patterns involving multiple tissues and multiple layers of physiological regulation. In contrast, there is increasing evidence that essential mechanisms underlying complex disorders can only be unveiled by considering multiple layers of biological observations together [1]. Despite this, only few attempts to integrate and analyze diverse biomedical data as a whole have been attempted. Indeed, both the construction of integrated **knowledge bases**, and the subsequent application of analysis methods, are technically challenging. Here, we summarize the main challenges for data integration in biomedicine, highlight trends and describe our current effort to overcome these challenges.

II. Age-related diseases

A. The leading causes of premature death

As the global population grows, and life expectancy increases, so does the number of people at risk for developing age-related diseases. Indeed, advances in sanitation, medicine and food security have contributed to considerably reduce child mortality and expand lifespan globally along the course of the last century. Subsequently, improving the health of the elderly and extending the so-called healthspan has emerged as a new challenge in medicine, as the leading causes of premature death moved from infections to cardiovascular diseases and cancer [2].

B. *Multiple risk factors*

Age-related diseases represent a large spectrum of disorders, including neurodegenerative, cardiovascular and musculoskeletal diseases as well as cancer. The development of these disorders is typically multi-factorial. Along with age, important risk factors include genetics, diet, life-style, smoke and environmental exposures, as well as one's history of diseases and medication. In addition, intricate factors, such as the accumulation of epigenetic changes throughout life, referred to as the epigenetic "clock", the microbiome, and interactions between factors, are also important (Fig.1) [3]–[6]. As the unique combination of risk factors certainly differs from one patient to another, individual variations need to be taken into account both in research and the clinic. For example, studies in mouse have shown that the effect of preventive interventions aiming at extending lifespan, such as dietary restriction, can vary from beneficial to detrimental depending on one's genetic makeup [7]. Similarly, the composition of the intestinal microbiome has been shown to affect how food is absorbed and metabolized [8], [9].

At the cellular level, aging is characterized by the loss of intracellular proteostasis, mitochondrial homeostasis, and epigentetic alterations (Fig. 1) [10]. Identifying bio-molecular pathways that can be exploited to slow down, delay or reverse these biological aging processes is therefore critical to address the leading health threats of today and tomorrow. This requires the investigation of physiological mechanisms at the molecular level across organs and tissues, while taking inter-individual variations into account.

III. PRECISION MEDICINE

A. *Research approach*

Precision – or personalized – medicine addresses complex diseases by adapting therapeutic approaches to the individual characteristics of the patient, and in particular to the genetics. This approach has been unlocked by the development of high-throughput biomolecular assays, or "omics", technologies, which allow to asses individuals physiology at the molecular level by drawing biomolecular profiles of tissues. From the research perspective, investigating the physiological mechanisms underlying complex conditions demands to profile numerous tissues, if not single cells, from a large diversity of subjects across multiple experimental conditions [11]. Such comprehensive research cannot be easily carried out in humans, mainly because the access to samples, and the control over important factors such as genetics, diet or environmental exposures, is limited. Indeed, although epidemiological studies can provide insights into the role of many factors that can reasonably be measured in human settings, such as genetics, clinical phenotypes and environmental factors, true experimentation to decipher of the underlying biomolecular mechanisms requires the use of experimental models. Relevant models range from cell lines and nematodes to larger organisms like mammals, depending on the question at hand.

B. *Mouse genetic diversity panels*

In order to study complex systems, researcher's strategy is to control as many variables as can be while measuring as many of those that cannot be controlled and simultaneously inducing controlled variations of one or multiple variables. With this regards, mouse genetic diversity panels are a model of choice to assess the links between genetic variations and physiological traits associated with complex conditions, and are seen as the experimental counterpart of **precision medicine** [12]. Indeed, these panels are composed of genetically diverse inbred strains of mice and are designed to provide a stable and reproducible genetic diversity across cohorts. This model therefore allows to introduce defined genetic variations while simultaneously controlling environmental conditions and diet while providing access to a large variety of biological samples and enabling the measurement of a variety of phenotypes.

C. *Translational research*

Although most of our knowledge in fundamental biology comes from model organisms, there are undeniable differences between human and mouse physiology. The research process in precision medicine is therefore a continuous cycle driven by epidemiological observations leading to the design of experiments on model organisms, and which results further demand validation in human settings [13]. Eventually, this translational process can lead to the prioritization and design of further human studies.

IV. SYSTEMS GENETICS

A. *Associative studies*

Linking genetic variations to phenotypes or to biomolecular profiles of tissues has been classically carried on in so called systems genetics studies using associative approaches, such as Genome-Wide Association Studies (GWAS) and Quantitative Trait Locus (QTL) mapping (Fig. 1)

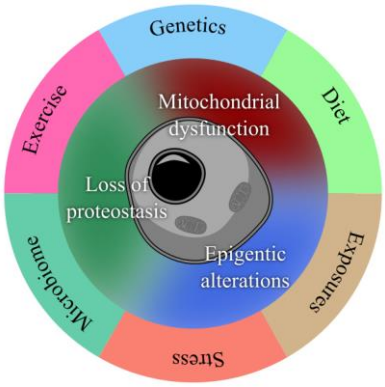


Fig. 1Cellular aging is associated to mitochondrial dysfunction, the loss of proteostasis and epigenetic alteration and impacted by heritable, life-style and environmental factors

2A). These approaches screen "omics" data for any association with phenotypes, gene variations or any observation on other "omics" layers [14]. Although these approaches unveiled numerous insights into the genetic roots of complex diseases, they may not exploit the full potential of the multi-modal "omics" datasets that can be collected. Indeed, such associative approaches are limited to assessing the links between pairs of biological layers, such as between genetics and phenotypes, or between gene expression and phenotypes (Fig. 2A). They are therefore bound to catch the "low-hanging fruits": single factors clearly associated with phenotypes (e.g. the genetic variant A is associated with a high body mass). This can miss important patterns of interactions across multiple layers of biological observations [15]: what if the genetic variant A was associated with a high body mass, but only if gene B is highly expressed in the liver while bacteria X is harbored in the gut?

B. Integrative analyses

Age-related diseases such as atherosclerosis cannot be explained by a single factor, like genetics alone, and in fact result from combinations of factors. Therefore, a large potential for discoveries and associated therapeutic opportunities is seen in integrative approaches that assess multiple factors together (Fig. 2B) [1]. In addition, patients Electronic Health Records (EHR), as well as observations from model organisms (e.g. biomolecular "omics" data) generally have a sparse nature. For example, EHR data collection depends on patients condition and specific needs, thus different sets of data, measured in different sequences at different time points, are generally not available for each patient. Similarly, all studies have a different design, which prioritizes the investigation of certain tissues with different methods according to the scientific question at hand and available resources. In this context, observations across biological layers and studies may complement each-other: information that may be present, but could not be measured, in one biological layer, may be available from another mechanistically connected layer. For instance, genes (DNA), gene expression (RNA) and proteins are linked by the central dogma of cell biology: genes are transcribed into mRNA, which is translated into proteins. Proteins in turn form the backbone of metabolic pathways by acting on other biomolecules in complex cascades of biochemical reactions. Despite these links, observations from one layer are not sufficient to predict the state of other layers because of the many mechanisms of regulation that exist within and between these layers. However, integrating multiple related layers of biological data can offer a more informative picture than if considering each layer separately because they may complement each other while interactions between them could be also taken into account. Last but not least, integrative studies may highlight which type of observation, and which tissue, may carry the most relevant information with regard to a particular disease. This could help prioritizing experiments and guiding experimental designs to maximize the "return on investment" of generally expensive data collection processes.

Overall, and despite being collected under controlled genetic and environmental conditions, "omics" datasets derived from animal models of genetic diversity are complex, noisy and incomplete [14]. However, they still remain more comprehensive and coherent than their counterparts in human settings. While the need to overcome the limitations of pairwise associative approaches like GWAS is increasingly recognized, considering multiple layers of biological observations in integrative approaches remains challenging and to date only a few attempts were made in the field [1]. This is due to two main challenges: 1) the combination of heterogeneous and sparse data into coherent knowledge bases. And 2) deriving actionable insights out of complex patterns.

V. Open science and data integration

A. *Metadata are the glue that links datasets*

Conceptually, the potential for the discovery of complex patterns grows with the heterogeneity and diversity of the analyzed data. But in order to keep results generalizable, the number of observations, or sample size, must grow as the diversity of considered observations increases (Fig. 2C). Unfortunately, study budgets rarely allow to simultaneously collect large amount of diverse data from large cohorts. Instead, studies generally face a trade-off between the sample size and the number of observations and tissues collected, and prioritize these according to their scientific objectives. Combining data from multiple studies is therefore the solution to reach sample sizes that may not be achievable in a single study, and unlocking the investigation of new scientific questions. However, the construction of a knowledge base from multiple independent datasets greatly depends on the interoperability of these datasets and essentially relies on the existence of sufficiently rich, detailed and standardized metadata (Fig. 2D). While vertical integration (i.e. within a study) of datasets may be facilitated by the use of common nomenclatures and annotations within a same study, horizontal integration (i.e. across studies) is typically more challenging, as practices can differ between research groups.

B. *The FAIR principles*

The FAIR principles were first formulated in 2016, as a mean to enable new discoveries by facilitating data integration and reuse [16]. These principles promote data management practices that enable the integration of compatible, or complementary, datasets. Indeed, Inter-operability and Reuse require that 1) data and metadata should be recorded using standard formats, notations and vocabularies, so that independent researchers could understand them and link information across datasets with as little ambiguity as possible. And 2), that the datasets should be documented with metadata that are rich and detailed enough for independent researchers to understand their exact provenance. In the case of integration, the description of the samples, the experimental design and the methods behind the data must allow any investigator to appreciate whether two datasets could be compared or merged together, and under which conditions.

C. *Data models*

General metadata schema, such as schema.org (https://schema.org) or Dublin Core (https://dublincore.org) provide a generic tool to describe datasets in a standard manner, yet fall short in describing the complex context of experimental procedures behind most biomedical datasets. In biomedicine, domain-specific metadata schema such as the Investigation Study Assay (ISA) model provides an appropriate framework to link “omics” data through a database that describes their often intricate relationship of origin, measurement technology, sequencing runs, and experimental conditions [17]. Besides metadata schema, standard notations, controlled vocabularies and ontologies are essential to provide descriptions that can be searched and compared in an automated manner. Indeed, as the amount of generated data grows, so does the need to automate the process of metadata searching and matching.

D. *Driving forces*

Public data repositories are instrumental in promoting good practices that facilitate data sharing and integration. For instance, domain-specific databases such as the European Nucleotide Archive (ENA) enforces the use of metadata models like ISA while generic repositories such as Zenodo promote more general-purpose standards like schema.org. Publishers and funding bodies increasingly demand that datasets associated to publications and projects are shared publicly on appropriate data sharing platforms. These strong driving forces in the research ecosystem facilitates the construction of knowledge bases across datasets and studies to enable larger integrative studies. However, data integration across independent “omics” studies remains challenging due to the inevitable differences and complexity of the experimental procedures.

E. *Reproducibility*

Last but not least, documenting data processing is also critical to understand how processed data should be handled and interpreted as well as if and how they could be integrated. Although documenting data processing code has been facilitated by the now ubiquitous version control systems (e.g. Git), ensuring the actual reproducibility of data processing workflows remains a technical challenge to most biomedical researchers today. Indeed, reproducing workflows often demand, on top of the data processing code, specific sets of software dependencies (i.e. the computing environment) as well as an understanding of the links between processing steps, data sources and results (i.e. a knowledge graph of the workflow) [18]. This is typically addressed using virtualization technologies such as Docker (https://www.docker.com [19]) and workflow orchestration systems like the Common Workflow Language (CWL) [20]. Although building and using data processing systems that enable the full reproducibility of workflows can be perceived as an unessential overhead in today’s context of competitive and time-pressured research, off-the-shelves complete data science technology stacks like Renku (https://renkulab.io [21]) are emerging. This will reduce

the barriers to the adoption of technologies that enable the reproducibility of data processing, and facilitate data integration in the future.

VI. Explainable artificial intelligence

A. *Machine learning unlocks integrative analysis*

In biomedicine, machine learning (ML) is used or investigated in a variety of applications, from patient diagnosis and prognosis to the design of new drugs and the prediction of their effects [1]. It is a tool of choice to identify and use complex patterns across multi-modal data that are otherwise non-obvious to the human researcher and hard to assess with more classic statistical tools. While predictive algorithms have so far dominated this scene, there is a growing interest for methods including a strong explainability aspect. Indeed, in the context of aging research and systems genetics, which study the links between biomolecular factors and health-related traits, predicting phenotypes is typically of interest for disease-interception applications that require to anticipate the development of disorders in healthy individuals and applying preventive interventions to delay this development and expand healthspan. However, this is of little value for the identification of possible treatment targets, without an understanding of the key factors that drive the prediction. Interpretable ML methods that can highlight key predictive features of phenotypes that are relevant to diseases across multi-modal datasets are an emerging alternative to overcome the limitation of pairwise associative methods. An early example of such approach has been used to predict tissue-specific protein functions based on a network of protein-protein interactions built across a variety of tissues.

The interpretable nature of the ML algorithm could then have been used to highlight the specific features important for the prediction of a function [22]. In the context of systems genetics, a similar method could be used to highlight, in a network of “omics” observations constructed across tissues, features relevant for the prediction of a trait of interest. A network could be based on gene co-expression, external knowledge bases such as publicly available annotations regarding gene-protein encoding (e.g. Ensembl [23]), protein-protein interactions (e.g. IntAct, MINT, MatrixDB [24]–[26]) and metabolic pathways (e.g. GO, KEGG [27]–[31]). Although building such networks across heterogeneous, noisy and sparse datasets requires significant efforts, such eXplainable Artificial Intelligence (XAI) framework could be a powerful tool to assist researchers in making discoveries.

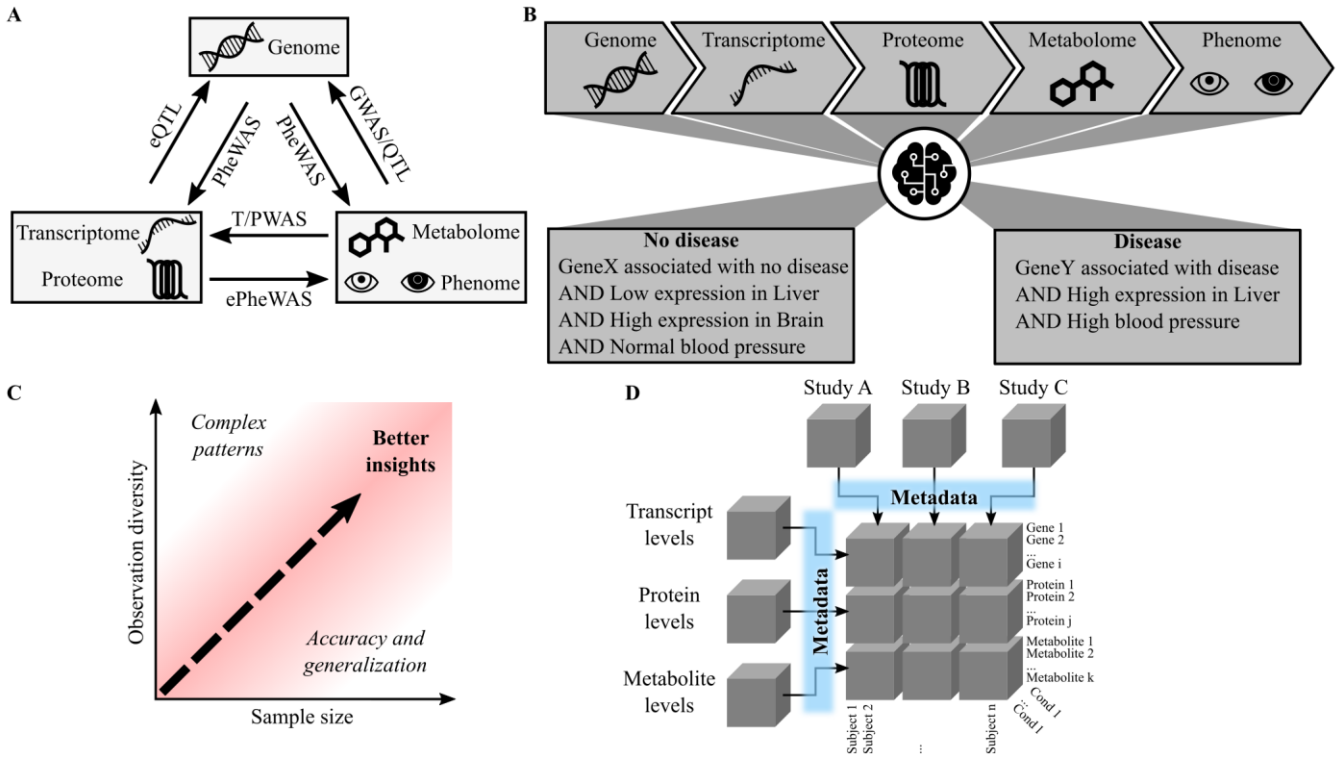


Fig. 2. **A)** Overview of "pairwise" associative studies in systems genetics. Genome Wide Association Studies (GWAS) and Quantitative Trait Locus (QTL) analyses search for associations between a phenotype and genetic variants. Phenome Wide Association Studies (PheWAS) look for associations with a gene variant within a collection of phenotypes or intermediate phenotypes. Expression Quantitative Trait Locus (eQTL) look for associations between an intermediate phenotype and genetic variants. Expression-based PheWAS (ePheWAS) search for associations with an intermediate phenotype within a collection of phenotypes. Transcriptome/Proteome-Wide Association Studies (T/PWAS) look for associations between a phenotype and variations in gene expression/protein levels. Adapted from Li *et al.* 2018 [14]. **B)** Integrative approaches assessing multiple layers of biological data together may capture complex patterns relevant to diseases, that could not be captured by "pairwise" associative approaches. Adapted from Li *et al.* 2018 [14] and Zitnik *et al.* 2019 [1]. **C)** The integration of diverse observations allows to capture more complex patterns, yet demand larger sample sizes in order to conserve the accuracy and generalization of insights. **D)** Vertical and horizontal data integration both require detailed and rich metadata. Metadata act as a glue that can link datasets across types of measurements (e.g. transcript, proteins or metabolite levels) or across studies.

VII. Use-case – Multi-omics integration in systems genetics

A. *The BXD family*

In a unique use-case, we intend to assemble a large knowledge base from a collection of "omics" and phenotype datasets collected on the BXD mouse genetic diversity panel and to use it to investigate integrative, exploratory approaches [32]. These datasets include genetics as well as gene expression, protein, lipids and metabolite levels measured across multiple tissues, as well as the composition of the gut microbiome and data from a large array of phenotyping tests targeting metabolic activity (blood pressure, body fat and lean mass, cardiac activity, glucose tolerance, etc.). This collection of datasets has been generated internally and its majority is publicly available on domain-specific repositories. Some of these data have been previously associated to discoveries reported across peer-reviewed publications [33]–[40]. These datasets are therefore well documented and our group has an excellent understanding of its complications and limitations, which is critical to determine under which conditions they can be integrated.

B. *Building a knowledge base*

Our first goal is to consolidate a knowledge base through an extraction, transformation and load (ETL) process in order to standardize data and metadata notations across datasets and link measurements across strains, tissues, experimental conditions and assays technologies (i.e. vertical integration). The use of domain-specific ontologies and standards, such as the Mammalian Phenotype ontology (MP) (http://purl.obolibrary.org/obo/mp.owl), the Vertebrate Trait ontology (VT) (http://purl.obolibrary.org/obo/vt), the Mouse Adult Gross Anatomy (MA) (http://purl.obolibrary.org/obo/ma.owl) and the Ontology of Biomedical Investigations (OBI) (http://purl.obolibrary.org/obo/obi [41]) will also facilitate future integration with independent studies carried out internally or by other research groups (i.e. horizontal integration).

C. *Machine learning on graphs*

This knowledge base will be used as a test bed for the development of integrative data analysis approaches based on ML. This will primarily focus on approaches based on graphs, as these are tools of choice to describe heterogeneous biological observations together with their links (such as interactions between proteins, mechanistic

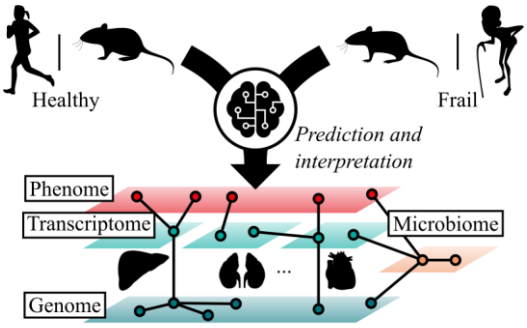


Fig. 1. Integrative systems genetics approach based on graphs. Biological data layers including diverse types of observations (e.g. genetic variations, gene transcription, phenotypes or microbiome composition) across multiple tissues (e.g. liver, kidney, heart) are integrated into a graph. Graph features that are key to predict a trait that is relevant for a disease (e.g. the body mass) are extracted using the interpretability of a graph-based ML algorithm.

links between genes, transcripts and proteins and metabolic pathways). In particular, we envision novel applications for graph Convolutional Neural Networks (CNN) [42]–[45]: Traits relevant to diseases could be predicted based on a network of connected "omics" observations across tissues. The inherent interpretability of graph convolutional neural networks could then highlight key predictive features of this network, which would help discovering complex biological mechanisms and potential therapeutic targets (Fig. 3).

VIII. CONCLUSION

There is a critical need to develop new strategies for preventing, delaying or reversing the course of age-related diseases in the growing and aging global population. Age-related diseases have complex multi-factorial roots which demand to take individual physiological characteristics into consideration both for research and treatment. Understanding the metabolic mechanisms underlying the aging process helps developing interventions to compensate its effects. In particular, integrative approaches that combine biological observations across multiple tissues promise to generate valuable insights into these complex biomolecular mechanisms. Although "omics" technologies and the use of model organisms enable the detailed investigation of tissues and cells physiology, identifying complex patterns and regulatory systems across tissues and biomolecular layers remains challenging. Indeed, it requires integrative approaches that can combine a large amount - and a diversity - of "omics" observations across multiple tissues and varying conditions. Integrative analyses need to combine multiple datasets of different types (i.e. genomic, proteomic, metabolomic) that could be generated within a same, or within multiple independent studies. Such integration requires a deep understanding of each dataset's characteristics and specificities. This demands rich, detailed and harmonized documentation and metadata. Although promoted by increasingly adopted open science standards, the necessary level of details is rarely accessible for publicly available datasets and such approach therefore remain marginal in biomedicine.

In order to provide a first use-case in the field of systems genetics, we are assembling a large knowledge base of heterogeneous "omics" datasets derived from the BXD mouse genetic diversity panel. This will enable to test the application of XAI methods for assisting researchers in the discovery of complex biological mechanisms relevant for age-related diseases. This study will allow to investigate the links between genetics, metabolism, tissues and phenotypes. It may enable the identification of novel therapeutic targets against complex disorders and set the ground for further integrative approaches in biomedicine.

ACKNOWLEDGMENT

We thank Benjamin Ricaud (EPFL) for his advises and guidance regarding graph-based ML, Christine Choirat (Swiss Data Science Center - SDSC), Eric Bouillet (SDSC) and Emma Jablonski (SDSC) for their advise and support regarding reproducible data science. We thank the SDSC for funding this work as part of its collaborative data science program (project SysGen).

Glossary

- **BXD mouse genetic diversity panel**: A set of ~200 strains of recombinant inbred mice derived from C57BL/6 and DBA/2 parents. Thanks to patterns of genetic recombinations that are unique to each inbred strain, this family of mice allows to resolve the effect of 6 million DNA variants on heritable traits.
- **FAIR principles**: Data management guidelines formulated in the 2016 paper "The FAIR Guiding Principles for scientific data management and stewardship" by Wilkinson et al. aiming at promoting Findability, Accessibility, Interoperability, and Reuse of digital objects. The FAIR principles constitute a key stone in open science as they clearly identify the essential elements needed for data reuse by the community.
- **Healthspan**: The period of life in which a person is in healthy condition.
- **Explainable Artificial Intelligence (XAI) and interpretable machine learning (ML)**: ML approaches focusing on models which underlying logic can be understood by the user. Explainable ML models are seen as white boxes. In contrast, models which logic cannot be understood by the user because it is based on too high levels of abstraction are considered black boxes.
- **Knowledge base**: In this text, the term "knowledge base" is used in its most general sense to describe any form of organized information around a dataset, indistinctively of its form or complexity (i.e. whether as a simple text table or as a complex relational or graph database). This includes metadata, metadata description and possible links within and between these elements.
- **Precision medicine**: An approach to medicine that takes patient individual characteristics into account for the design of personalized treatments. While this concept is not new and has been applied in the past (e.g. blood transfusion needs to be adapted to patient's blood type), the terms "precision medicine" (interchangeable with the term "personalized medicine") refer to emerging approaches that account for complex characteristics, or combinations of them, found in genetics, life-style and a patient's environment.
- **Systems genetics**: A research approach to understand complex traits. Systems genetics investigates the links between genetics variations, intermediate molecular phenotypes (i.e. gene expression, metabolites levels, etc.) and traits.

References

[1] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman, "Machine Learning for Integrating Data in Biology and Medicine: Principles, Practice, and Opportunities," *Inf. Fusion*, vol. 50, pp. 71–91, Oct. 2019, doi: 10.1016/j.inffus.2018.09.012.

[2] T. Vos *et al.*, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, Oct. 2020, doi: 10.1016/S0140-6736(20)30925-9.

[3] S. Horvath and K. Raj, "DNA methylation-based biomarkers and the epigenetic clock theory of ageing," *Nat. Rev. Genet.*, vol. 19, no. 6, Art. no. 6, Jun. 2018, doi: 10.1038/s41576-018-0004-3.

[4] C. A. Boix, B. T. James, Y. P. Park, W. Meuleman, and M. Kellis, "Regulatory genomic circuitry of human disease loci by integrative epigenomics," *Nature*, vol. 590, no. 7845, Art. no. 7845, Feb. 2021, doi: 10.1038/s41586-020-03145-z.

[5] M. Yu, W. D. Hazelton, G. E. Luebeck, and W. M. Grady, "Epigenetic Aging: More Than Just a Clock When It Comes to Cancer," *Cancer Res.*, vol. 80, no. 3, pp. 367–374, Feb. 2020, doi: 10.1158/0008-5472.CAN-19-0924.

[6] V. D. Badal *et al.*, "The Gut Microbiome, Aging, and Longevity: A Systematic Review," *Nutrients*, vol. 12, no. 12, Dec. 2020, doi: 10.3390/nu12123759.

[7] C.-Y. Liao, B. A. Rikke, T. E. Johnson, V. Diaz, and J. F. Nelson, "Genetic variation in the murine lifespan response to dietary restriction: from life extension to life shortening," *Aging Cell*, vol. 9, no. 1, pp. 92–95, 2010, doi: https://doi.org/10.1111/j.1474-9726.2009.00533.x.

[8] D. Zeevi *et al.*, "Personalized Nutrition by Prediction of Glycemic Responses," *Cell*, vol. 163, no. 5, pp. 1079–1094, Nov. 2015, doi: 10.1016/j.cell.2015.11.001.

[9] A. A. Kolodziejczyk, D. Zheng, and E. Elinav, "Diet–microbiota interactions and personalized nutrition," *Nat. Rev. Microbiol.*, vol. 17, no. 12, Art. no. 12, Dec. 2019, doi: 10.1038/s41579-019-0256-8.

[10] S. Kaushik and A. M. Cuervo, "Proteostasis and aging," *Nat. Med.*, vol. 21, no. 12, Art. no. 12, Dec. 2015, doi: 10.1038/nm.4001.

[11] N. Almanzar *et al.*, “A single-cell transcriptomic atlas characterizes ageing tissues in the mouse,” *Nature*, vol. 583, no. 7817, Art. no. 7817, Jul. 2020, doi: 10.1038/s41586-020-2496-1.
[12] H. Li and J. Auwerx, “Mouse Systems Genetics as a Prelude to Precision Medicine,” *Trends Genet.*, vol. 36, no. 4, pp. 259–272, Apr. 2020, doi: 10.1016/j.tig.2020.01.004.
[13] J. H. Nadeau and J. Auwerx, “The virtuous cycle of human genetics and mouse models in drug discovery,” *Nat. Rev. Drug Discov.*, Jan. 2019, doi: 10.1038/s41573-018-0009-9.
[14] H. Li *et al.*, “An Integrated Systems Genetics and Omics Toolkit to Probe Gene Function,” *Cell Syst.*, vol. 6, no. 1, pp. 90-102.e4, Jan. 2018, doi: 10.1016/j.cels.2017.10.016.
[15] M. Claussnitzer *et al.*, “A brief history of human disease genetics,” *Nature*, vol. 577, no. 7789, pp. 179–189, Jan. 2020, doi: 10.1038/s41586-019-1879-7.
[16] M. D. Wilkinson *et al.*, “The FAIR Guiding Principles for scientific data management and stewardship,” *Sci. Data*, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.
[17] S.-A. Sansone *et al.*, “Toward interoperable bioscience data,” *Nat. Genet.*, vol. 44, no. 2, Art. no. 2, Feb. 2012, doi: 10.1038/ng.1054.
[18] B. K. Beaulieu-Jones and C. S. Greene, “Reproducibility of computational workflows is automated using continuous analysis,” *Nat. Biotechnol.*, vol. 35, no. 4, Art. no. 4, Apr. 2017, doi: 10.1038/nbt.3780.
[19] C. Boettiger, “An introduction to Docker for reproducible research,” *ACM SIGOPS Oper. Syst. Rev.*, vol. 49, no. 1, pp. 71–79, Jan. 2015, doi: 10.1145/2723872.2723882.
[20] “Common Workflow Language, v1.0.” figshare, Jul. 08, 2016, doi: 10.6084/m9.figshare.3115156.v2.
[21] *SwissDataScienceCenter/renku*. Swiss Data Science Center, 2021.
[22] M. Zitnik and J. Leskovec, “Predicting multicellular function through multi-layer tissue networks,” *Bioinformatics*, vol. 33, no. 14, pp. i190–i198, Jul. 2017, doi: 10.1093/bioinformatics/btx252.
[23] A. D. Yates *et al.*, “Ensembl 2020,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D682–D688, Jan. 2020, doi: 10.1093/nar/gkz966.
[24] S. Orchard *et al.*, “The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases,” *Nucleic Acids Res.*, vol. 42, no. D1, pp. D358–D363, Jan. 2014, doi: 10.1093/nar/gkt1115.
[25] L. Licata *et al.*, “MINT, the molecular interaction database: 2012 update,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D857-861, Jan. 2012, doi: 10.1093/nar/gkr930.
[26] O. Clerc *et al.*, “MatrixDB: integration of new data with a focus on glycosaminoglycan interactions,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D376–D381, Jan. 2019, doi: 10.1093/nar/gky1035.
[27] Gene Ontology Consortium, “The Gene Ontology resource: enriching a GOld mine,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D325–D334, Jan. 2021, doi: 10.1093/nar/gkaa1113.
[28] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology,” *Nat. Genet.*, vol. 25, no. 1, Art. no. 1, May 2000, doi: 10.1038/75556.
[29] M. Kanehisa and S. Goto, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000, doi: 10.1093/nar/28.1.27.
[30] “Toward understanding the origin and evolution of cellular organisms - Kanehisa - 2019 - Protein Science - Wiley Online Library.” https://onlinelibrary.wiley.com/doi/full/10.1002/pro.3715 (accessed Feb. 10, 2021).
[31] M. Kanehisa, M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe, “KEGG: integrating viruses and cellular organisms,” *Nucleic Acids Res.*, vol. 49, no. D1, pp. D545–D551, Jan. 2021, doi: 10.1093/nar/gkaa970.
[32] P. A. Andreux *et al.*, “Systems Genetics of Metabolism: The Use of the BXD Murine Reference Panel for Multiscalar Integration of Traits,” *Cell*, vol. 150, no. 6, pp. 1287–1299, Sep. 2012, doi: 10.1016/j.cell.2012.08.012.
[33] E. G. Williams *et al.*, “Systems proteomics of liver mitochondria function,” *Science*, vol. 352, no. 6291, p. aad0189, Jun. 2016, doi: 10.1126/science.aad0189.
[34] Y. Wu *et al.*, “Multilayered Genetic and Omics Dissection of Mitochondrial Activity in a Mouse Reference Population,” *Cell*, vol. 158, no. 6, pp. 1415–1430, Sep. 2014, doi: 10.1016/j.cell.2014.07.039.
[35] E. G. Williams *et al.*, “Quantifying and Localizing the Mitochondrial Proteome Across Five Tissues in A Mouse Population,” *Mol. Cell. Proteomics*, vol. 17, no. 9, pp. 1766–1777, Sep. 2018, doi: 10.1074/mcp.RA118.000554.
[36] K. Gariani *et al.*, “Eliciting the mitochondrial unfolded protein response by nicotinamide adenine dinucleotide repletion reverses fatty liver disease in mice,” *Hepatol. Baltim. Md*, vol. 63, no. 4, pp. 1190–1204, Apr. 2016, doi: 10.1002/hep.28245.
[37] E. G. Williams *et al.*, “An Evolutionarily Conserved Role for the Aryl Hydrocarbon Receptor in the Regulation of Movement,” *PLOS Genet.*, vol. 10, no. 9, p. e1004673, Sep. 2014, doi: 10.1371/journal.pgen.1004673.

[38] D. Ryu *et al.*, “NAD+ repletion improves muscle function in muscular dystrophy and counters global PARylation,” *Sci. Transl. Med.*, vol. 8, no. 361, pp. 361ra139-361ra139, Oct. 2016, doi: 10.1126/scitranslmed.aaf5504.
[39] P. Jha *et al.*, “Genetic Regulation of Plasma Lipid Species and Their Association with Metabolic Phenotypes,” *Cell Syst.*, vol. 6, no. 6, pp. 709-721.e6, Jun. 2018, doi: 10.1016/j.cels.2018.05.009.
[40] P. Jha *et al.*, “Systems Analyses Reveal Physiological Roles and Genetic Regulators of Liver Lipid Species,” *Cell Syst.*, vol. 6, no. 6, pp. 722-733.e6, Jun. 2018, doi: 10.1016/j.cels.2018.05.016.
[41] “The Ontology for Biomedical Investigations.” https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0154556 (accessed Feb. 10, 2021).
[42] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, “Geometric Deep Learning: Going beyond Euclidean data,” *IEEE Signal Process. Mag.*, vol. 34, no. 4, pp. 18–42, Jul. 2017, doi: 10.1109/MSP.2017.2693418.
[43] J.-B. Cordonnier and A. Loukas, “Extrapolating paths with graph neural networks,” *ArXiv190307518 Cs Stat*, Mar. 2019, Accessed: Jul. 10, 2020. [Online]. Available: http://arxiv.org/abs/1903.07518.
[44] F. Dutil, J. P. Cohen, M. Weiss, G. Derevyanko, and Y. Bengio, “Towards Gene Expression Convolutions using Gene Interaction Graphs,” *ArXiv180606975 Cs Q-Bio Stat*, Jun. 2018, Accessed: Jul. 10, 2020. [Online]. Available: http://arxiv.org/abs/1806.06975.
[45] S. Rhee, S. Seo, and S. Kim, “Hybrid Approach of Relation Network and Localized Graph Convolutional Filtering for Breast Cancer Subtype Classification,” pp. 3527–3534, 2018.

FAIR4Health: Improving Health Research in EU through FAIR Data

Vasiliki Foufi
Division of Medical Information Sciences
Geneva University Hospitals
University of Geneva
Geneva, Switzerland
vasiliki.foufi@unige.ch

Jean-Philippe Goldman
Division of Medical Information Sciences
Geneva University Hospitals
University of Geneva
Geneva, Switzerland
jean-philippe.goldman@unige.ch

Jessica Rochat
Division of Medical Information Sciences
Geneva University Hospitals
University of Geneva
Geneva, Switzerland
jessica.rochat@hcuge.ch

Carlos Luis Parra Calderón
Andalusian Health Service (SAS)
Seville, Spain
carlos.parra.sspa@juntadeandalucia.es

Christophe Gaudet-Blavignac
Division of Medical Information Sciences
Geneva University Hospitals
University of Geneva
Geneva, Switzerland
christophe.gaudet-blavignac@hcuge.ch

Christian Lovis
Division of Medical Information Sciences
Geneva University Hospitals
University of Geneva
Geneva, Switzerland
christian.lovis@hcuge.ch

***Abstract*—*This paper presents an overview of the FAIR4Health European project (2019-2021) which aims at encouraging the reuse of research data generated by publicly funded research projects. The project is coordinated by the Virgen del Rocío University Hospital, Andalusian Health Service (SAS) and the consortium consists of 17 partners from 11 EU and non-EU countries. A technological platform and tools are being developed for data FAIRification and data mining tasks. To test the feasibility of the technological solutions on real data, two pathfinder case studies will be performed.*

Keywords*—*FAIR data, interoperability, data sharing, data mining

I. Introduction

Sharing meaningful data is an important challenge in the field of personalized medicine. Besides legal and ethical aspects, there are technical challenges to be met so as to handle massive amounts of multimodal and heterogeneous distributed data, and also semantic challenges in order to build an interoperable framework. The H2020 FAIR4Health (F4H) project (2019-2021) aims at facilitating and encouraging the European Union (EU) Health Research community to FAIRify, share and reuse datasets derived from publicly funded research initiatives. A user-centered FAIR4Health platform and F4H agents are being developed to enable the translation from raw (meta) data to FAIR (meta) data. To validate the F4H platform and demonstrate the feasibility of FAIRifying medical research data, two pathfinder case studies prototypes will be performed on real medical data. The F4H consortium consists of 17 partners from both public and private entities. (*FAIR4Health - Consortium*, n.d.)

II. The FAIR4Health Project

A. Objectives

The F4H EU project, coordinated by the Virgen del Rocío University Hospital, Andalusian Health Service (SAS), consists of:

- 6 health research organizations;
- 2 universities, experts in data management;
- 4 academic partners with strong background on medical informatics;
- 5 business actors.

The ultimate goal of the project is to facilitate and encourage the EU Health Research community to FAIRify, i.e. make findable, accessible, interoperable and reusable, share and reuse their datasets derived from publicly funded research initiatives. The specific objectives are:

1. To design and implement an effective outreach strategy at EU level.

2. To produce a set of guidelines to set the foundations for a FAIR data certification roadmap.

3. To develop and validate an intuitive, user-centered F4H platform and F4H agents.

4. To demonstrate the potential impact in health research and health outcomes.(*FAIR4Health - Project*, n.d.-a)

B. Challenges and technological solutions

To meet these objectives, various aspects should be taken into account: legal and ethical aspects, technical aspects as well as semantic aspects. To address these challenges, various technological solutions are being developed in the framework of the F4H project.

First of all, FAIRification tools (*FAIR4Health - Newsletter*, n.d.) are being built to enable users transform the raw data into FAIR datasets. These tools are composed of:

- a repository created based on the Health Level Seven International (HL7) Fast Healthcare Interoperability Resources (FHIR) standard; (*Overview - FHIR v4.0.1*, n.d.)
- a Health Digital Terminology for concept translation and mappings (SNOMED CT (*SNOMED Home Page*, n.d.), LOINC (*Home*, n.d.), ICD (*Classification of Diseases (ICD)*, n.d.), etc.);
- a Data Curation & Validation tool to connect health data sources and migrate data into the HL7 FHIR Repository;
- a Data Privacy tool for privacy challenges on sensitive health data via de-identification and pseudonymization techniques.

The technological solution that will support this project will be based on two main entities: The F4H platform and F4H agents. The F4H agents, which will be located at the data owner's premises, will enable the FAIRification of local datasets. At the end of this process, datasets will be normalized, curated and mapped to domain vocabularies and ontologies, acting like Data FAIRports. The agents will also host instances of the Privacy-Preserving Distributed Data Mining (PPDDM) services so they could be run locally without the need of hosting these datasets outside the owner's premises. (*FAIR4Health - Newsletter*, n.d.)

Figure 1 displays the main architecture of the F4H platform and the interaction among the F4H community

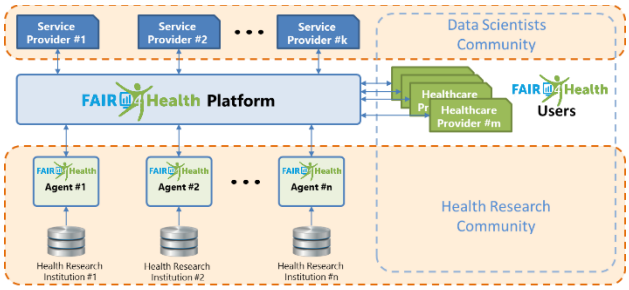


Fig. 1. FAIR4Health Open Community. (FAIR4Health - Project, n.d.-b)

As shown in the figure, the health researchers will be the main contributors in the FAIRification tasks. In turn, they will be able to browse and access to FAIRified datasets from their peers, always in compliance with the data owners' policies in terms of licenses, an approval from the local ethics committee, the regulatory framework, etc.

C. Use Cases

To validate the F4H platform and demonstrate the feasibility of FAIRifying medical research data, two pathfinder case studies prototypes will be performed on real medical data, both retrospective and prospective.

1. The first use case will focus on the identification of multimorbidity patterns and polypharmacy correlation on the risk of mortality in elderly patients via a multicentric observational study on datasets from 5 different European cohorts.

2. In the second use case, an early prediction service for 30-days readmission risk in patients with Chronic Obstructive Pulmonary Disease (COPD) will be developed. For this goal, both a retrospective and a prospective observational study will be carried out. (*FAIR4Health - Newsletter*, n.d.)

III. Conclusion

The F4H project goes beyond the health research domain and addresses the beneficial impact that the FAIR data strategy may have in health outcomes as well. There is a triple win behind this community:

1. Health researchers could access large datasets and accelerate the discovery of knowledge while avoiding bias due to local datasets;

2. eHealth services providers could develop and exploit innovative services in the EU Digital Single Market;

3. Healthcare providers could have access to this eHealth services portfolio to improve their quality of care. (*FAIR4Health - Project*, n.d.-a)

Interested parties who would like to benefit from the FAIRification workflow and tools developed by FAIR4Health, and the knowledge gained, can join the FAIR4Health community via the website https://www.fair4health.eu/en/membership.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 824666.

References

Classification of Diseases (ICD). (n.d.). Retrieved May 5, 2021, from https://www.who.int/standards/classifications/classification-of-diseases

FAIR4Health—Consortium. (n.d.). Retrieved February 11, 2021, from https://www.fair4health.eu/en/partners

FAIR4Health—Newsletter. (n.d.). Retrieved February 11, 2021, from https://www.fair4health.eu/en/newsletter

FAIR4Health—Project. (n.d.-a). Retrieved February 11, 2021, from https://www.fair4health.eu/en/project#

FAIR4Health—Project. (n.d.-b). Retrieved February 11, 2021, from https://www.fair4health.eu/en/project#

Home. (n.d.). LOINC. Retrieved May 5, 2021, from https://loinc.org/

Overview—FHIR v4.0.1. (n.d.). Retrieved May 5, 2021, from https://www.hl7.org/fhir/overview.html

SNOMED Home page. (n.d.). SNOMED. Retrieved May 5, 2021, from /

Challenges for Putting FAIR into Practice

Peter Wittenburg
Max Planck Computing and Data Facility
Max Planck Society
Munich, Germany
peter.wittenburg@mpcdf.mpg.de

***Abstract*—*The understanding is growing that the emerging integrated and interoperable data domain (I2D2) will have far reaching consequences for society, research and economy comparable to the changes caused by the Internet, for example. Such large infrastructures have a global dimension and require global agreements which in general are simple pre-competitive standards. Despite the FAIR Principles and the concept of FAIR Digital Objects (FDO), it seems that we are still far away from agreements on convergent specifications if we for example look at practices in the data labs. In this paper we describe the evolution of the FDO concept and point to the crucial role of global, unique, persistent and resolvable identifiers as basis for FDO.*

Keywords*—*Data Management, FAIR data, Digital Objects

I. Introduction

In 2018 Wittenburg and Strawn wrote a paper with the title "Common Patterns in Revolutionary Infrastructures and Data" [Wittenburg 2018] in which they found common patterns in the evolution of a few large infrastructures such as electrification, Internet and the Web. As figure 1 indicates, an early vision is taken up by an increasing number of people that explore the landscape of possible solutions. This creolisation phase leads to many different suggestions, testbeds and implementations increasingly lacking coherence and creating interoperability challenges. As consequence, the wishes to improve convergence get into the focus of developments leading to some attractors which are then evaluated and discussed at various levels such as by technologists, economists and politicians. Finally, after some time of debates there are decisions about convergence such as 50 Hz AC for electricity transmission, TCP/IP for Internet message exchange and HTTP/HTML for the Web information exchange. More examples can be mentioned from history such as for railroads and telephone systems. These mostly simple core standards lead to convergence which people could then rely on and build applications which led to enormous exploration waves. These simple standards reduce complexity and are pre-competitive, i.e. they have the chance to be globally accepted.

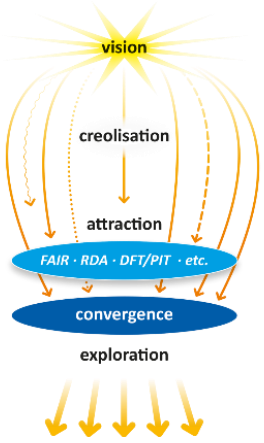


Figure 1 describes the typical pattern found in the evolution of large infrastructures – from an early vision to an utterly dynamic exploration phase

Wittenburg and Strawn applied this pattern now to the area of data since there is no doubt that we will need a globally Integrated and Interoperable Data Domain (ISD2) to make use of the value of the increasing volumes of data that will become available for the benefits of society, research and economy. They concluded that with the publication of the FAIR principles [Wilkinson 2016] and the specification of the FAIR Digital Objects [DFT 2016, Paris 2019] implementing the FAIR principles enormous steps have been made towards a convergence for building a global I2D2. Of course, from their study it was obvious that it often takes decades between launching a vision and coming to agreements on convergence. One reason for the delays is to find in the possible impact and political relevance of such infrastructures, i.e. different stakeholders want and need to have a saying in this process.

The assumption was that with the FAIR Principles and the FAIR Digital Objects basic elements of a future I2D2 are available and would accelerate the discussion about convergence finding. However, this agreement cannot yet be seen for FDOs, although there is broad agreement on the FAIR Principles. However, principles are not blueprints for building infrastructures and different interpretations of the FAIR principles already emerged.

In chapter II I will relate the hopes on fast agreement finding with the reality in many data labs. In chapter III I will elaborate on the development of the concept of Digital Objects which was recently extended to FAIR Digital Objects (FDO). In chapter IV I will briefly explain the crucial role of identification and in chapter V draw some conclusions.

II. Data Practices

It is well-known that about 80% of the time in data projects is spent on efforts related to what is called data wrangling and these inefficiencies seem to be in the same order in all sectors [Wittenburg 2018]. Data wrangling includes all steps that are necessary to be carried out before one can start the analytics. From industry it is known that about 60% of data projects fail which to a large extent is devoted to underestimating the costs of data wrangling. Given my own experience in a Max Planck Institute which from its beginning was data-driven and also dependent on data from other institutes I would claim that in the research domain the failure rate is comparable. The high costs associated with data intensive research have as consequence that many smaller departments and individual researchers are widely excluded from data science, since a technical support staff and student assistents that carry out all the data wrangling are needed. In many institutes PhDs need to carry out this basic work as part of their thesis work.

Recently, we had the chance to analyse about 70 research reports in great detail and we could identify a number of paradoxes which illustrate some additional challenges [Jeffery 2021]: (1) Researchers have already heard about the FAIR principles and support their basic messages. However, they like to shift making digital objects FAIR to the end of the project, since this allows them to continue in the labs with what they are used to and to not suffer from disruptions. We know however, that this does not work well and that the costs for delayed curation processes are by factors higher than for immediate actions [Beagrie 2008]. (2) Many researchers support the open science principle, however, only make commitments about the data that is being associated with publications. Since more than 90% of the data being created is being reused in the processes in the data labs, this has the consequence that much data will not be part of open sharing. We should note here, that of course data is already being exchanged, but often without being FAIR, i.e. people exchange files without submitting identifiers and metadata. (3) The statement of G. Strawn that "standards are good for science, but not for scientists" was confirmed [Strawn n.d.]. In general, researchers are not really interested in standards, but in tools that allow them to address their challenges at that very moment. Therefore, they like tools and of course are happy when tools support some useful standard. Research organisations have a different attitude towards standards, since they will at the end reduce inefficiencies and thus lead to more results for the investments made. (4) The number of excellent tools which is being used across all the labs is increasing, but this does currently not increase interoperability between the data silos. There are so many smart young researchers and developers who just take any new technology available and apply them. Some of these tools incorporate some workflow steps and increase local efficiency. (5) Discipline experts believe that their methods and processes are unique. Comparing processes, however, across disciplines indicates that there are many recurring patterns that could be subject of automation by prefabricated workflows. (6) In modern data science different researchers and institutions are involved in the processes due to different types of expertises. In such situtations responsibility for making data FAIR is not clarified and thus shifted between the actors. (7) DMPs are now seen by funders as a tool to improve data practices. But in general researcher see it as bureaucratic act which may help at the very beginning of a project to clarify basic needs. Experience shows that to a large extent DMPs are not useful. This may change when they would be pro-actively supported by data stewards.

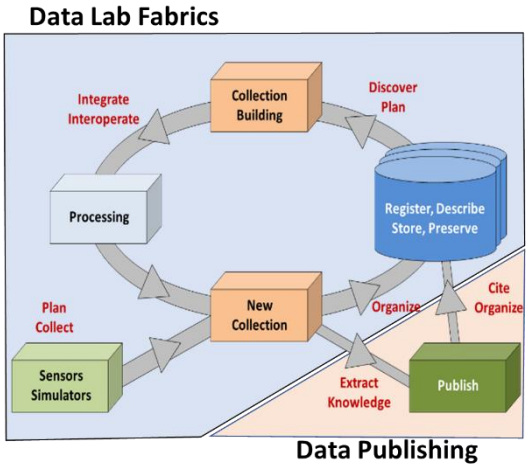


Figure 2 points to the typical data processing cycles in data labs where researchers organize new and existing data stored in repositories to collections to carry out some processing. Often these cycles are repeated until results can be produced that show evidence that will lead to a final publications often with large delavs.

Summarising, I can state that the recent study widely confirms the results which we gained in 2014 in the RDA EU project [Stehouwer 2016] which led to building the RDA Data Fabric Interest Group (DFIG)[49]. The DFIG had as goal to not analyse and optimise the final step of data publishing, but to look at the processes in the data labs for two reasons: (1) The publishers and librarians are already highly active to optimise the data publishing processes as indicated in the low-right corner of the diagram. (2) If we ever want to realise Open Science (or FAIRness) by Design [BRDI 2018], i.e. from the beginning of projects, we need to optimise data practices in the labs as indicated in the centre of the diagram which is of course a much more complex task. But it will be the only way to make data practices more FAIR and efficient.

After long discussions the DFIG came to the conclusion that the concept of Digital Objects (DO) which later became FAIR Digital Objects (FDO) will be the most promissing way in the long-term to change practices based on smoothly integrated data

[49] https://www.rd-alliance.org/group/data-fabric-ig.html

infrastructures. Therefore in the next chapter I will focus on the evolution of the concepts of DO and FDO.

III. (FAIR) Digital Objects

A. *Digital Objects*

When R. Kahn started the Internet and invented TCP/IP it was obvious that the messages being exchanged at the TCP/IP level are meaningless. Meaningful messages that will be exchanged between two centres are chopped into small pieces in the sending centre, transmitted and aggregated on the receiving side again. Already when the Internet

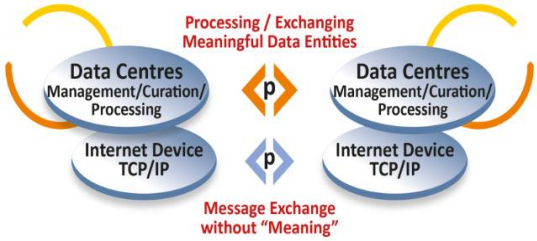


Figure 3 indicates the two layers of information exchange using the Internet. At the TCP/IP layer in general chunked and thus meaningless messages are exchanged. At a higher-level, meaningful messages must be exchanged such as files, emails. messages, etc.

was born it was obvious to the designers that there need to be unifying mechanisms at the application level. We have seen that in the early days for example FTP was designed to transmit files and SMTP as protocol to exchange emails. Roughly a decade later T. Berners-Lee invented the pair HTML and HTTP as mechanism to exchange "web-pages" on top of TCP/IP. Web-pages are written in HTML and HTTP understands HTML encoded information streams. The Web is using URNs as identifiers which come in two forms: URNs as used in some national libraries and URLs which are generally used for all kinds of information.

Roughly at the same time R. Kahn and his team developed the Handle System[50] to have a means for global, unique, persistent and resolvable identifiers (PID). Handles are independent of any technology like an ISBN is, while URLs always encode semantics such as ownership and location and are dependent on the HTTP protocol. The publishers realised the fundamental difference between URNs and Handles at an early phase and soon defined DOIs[51] which are basically Handles with a prefix 10 combined with a specific business model. Instead of web-pages the Handle resolver returns structured data which can be interpreted by machines.

At the same time (around 2000), first labs dealing with large amounts of data started using Handles. My own team decided to follow that path, i.e. in our repository with about 80 TB of organised data in 2010, all stored data and metadata items have assigned a unique Handle. Other repositories with even more data took the same step and in 2014 the Max Planck Society, for example, decided to run a persistent Handle services for all its institutes and researchers. Motivated by the uptake of the Handle System, in 2005 R. Kahn and R. Wilensky revised their early paper from 1995 on digital objects [Kahn 1995, Kahn 2006]. It was the first time that the term Digital Object was coined to indicate the items that were being exchanged via the Internet protocols. A Digital Object can contain bit-sequences of any type: data, metadata, software, assertions, etc. DOs therefore are the most abstract definition of the content that can be transferred. Due to the great success of the Web which was soon be used for all kinds of applications, the notion of Digital Objects was widely forgotton although it exists per definition in the term Digital Object Identifier (DOI).

[50] https://www.handle.net/
[51] https://www.doi.org/

In 2013 when the research Data Alliance[52] was set up one of the first groups that was established was the Data Foundation and Terminology working group co-chaired by the author[53]. Based on many use cases from various disciplines we ended up in defining the Core Data Model [Berg-Cross 2016] which is an extended version of the DO model as introduced by Kahn & Wilensky. The RDA DFT Digital Object model states the following: (1) Each DO has a structured bit-sequence encoding its content. (2) Each DO is assigned a PID and associated with metadata (can be of different types from type, descriptive, provenance to rights and transactions). (3) Each metadata description is a DO in itself. (4) DOs can be aggregated to Digital Collections which are also DOs.

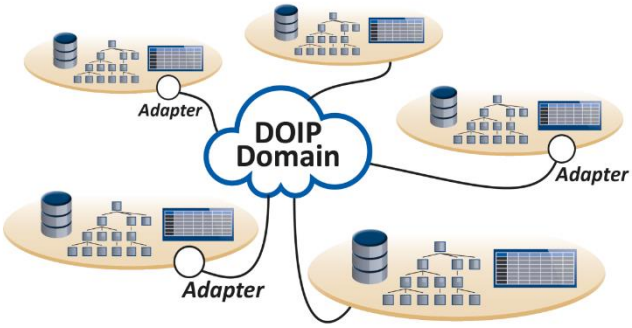


Figure 4 describes the emerging landscape of data repositories and other data providers/consumers all using different technologies and data organizations. A domain integrated by the Digital Object Interface Protocol would reduce complexity substantially, opening the path to an integrated data domain.

Two other RDA groups are of relevance it defining the core of DOs. The PID record can contain a number of Kernel attributes[54] which are returned during the resolution step, and which need to be typed and registered to make the resolution result fully machine actionable. The RDA Kernel group defined a set of attributes which are often being used and they were registered in an open Data Type Registry instance. The RDA Data Type Registry group defined a model for defining and registering Kernel types[55]. Note that the DO definition does not make any specifications about the nature of metadata.

A. Extension to FAIR DO

In 2019 experts from RDA Data Fabric, RDA GEDE[56] and GOFAIR[57] started interacting about the FAIRness of DOs [Schultes 2018]. FAIR implies machine actionability of data and metadata. As indicated above, the definition of DOs makes recommendations about typing PID attributes, but does not strictly require their definition and registration which needs to be changed. In addition, it was found that a linear registry of types might not be sufficient at the end to cover the complexity of the domain of digital object types. Therefore, an ontological approach was suggested. One aspect is clearly underspecified: metadata standards are in the hand of the research communities and most of them spent much effort during the last decades to specify their metadata schemas and concepts. Most of these specifications do not meet the criterium of machine actionability, but communities will hesitate to adapt their standards quickly. At the Paris workshop the FDO Framework was specified which should be the basis of all FDO discussions[58].

Summarising, we can state that the difference between DOs and FDOs are as follows: (1) The DO concept does not strictly request typing of all PID Kernel attributes, while the FDO concept does. It is recommended to use RDF compliant type specifications. (2) Over time the current linear type registry should be extended by a more complex ontology. (3) The DO model does not make strong requirements about the metadata provided by communities, while the FDO concept requires machine actionability. However, it will take time to meet this criterium since communities need to be convinced to do adaptations. Therefore, the DOs/FDOs including their specifications of PID Kernel Types and about Data Type Registries are widely FAIR compliant. Yet the DO/FDO domain is lacking a systemic implementation appraoch.

[52] https://rd-alliance.org/
[53] https://www.rd-alliance.org/dft-work-group
[54] https://www.rd-alliance.org/groups/pid-kernel-information-profile-management-wg
[55] https://www.rd-alliance.org/groups/data-type-registries-wg.html
[56] https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda
[57] https://www.go-fair.org/
[58] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/FDOF

B. Integrative (F)DO Domain

One of the big sources of inefficiencies in the data domain is the variety of technologies and organisations schemes of data and metadata in the thousands of existing and emerging repository systems. In figure 4 repositories with cloud, file and database systems are indicated and there are different variants of them. Much more problematic are the differences in data organisation, i.e. while data is often stored in files for example, different types of metadata (descriptive, scientific, rights, transactions, etc.) are stored in a variety of different containers and mostly without machine actionable relations between the different entities. This is the reason why in daily practices still mostly only files are being exchanged and the metadata is lost or forgotten on the way, which implies that reuse is often associated with time consuming wrangling.

Figure 5 indicates the referencing structures used for the domain of books. The ISBN numbers point to some metadata representing the "book as a work". Separated from this are local catalogues that refer to shelf locations where the "printed books" can be found.

Digital Objects as well as FAIR Digital Objects offer now a completely different opportunity since they can be used as common glue to achieve interoperability. In the world of FDOs the Digital Object Interface Protocol (DOIP)[59] acts as an interoperable gateway like TCP/IP acted as a gateway between the different network types that had been developed beforehand. What does DOIP solve? It reduces complexity from N*N to N*1 as TCP/IP did for networking since one protocol is now sufficient to interconnect the thousands of data repositories and clients. It puts responsibility for the integration to each repository (or other DO service) and is relying on persistent identifiers which may turn out as one of the few salient corner stones in an utterly dynamic scenario. In addition, it opens the path for service providers to stepwise make their data organisation FDO compliant implying that the connector will then be trivial. For any large data infrastructure such as EOSC and NFDI, FDOs therefore offer big chances to create the Integrated and Intereoperable Data Domain (I2D2) which we are dreaming of. Of course, DOIP is not a protocol addressing, for example, all the challenges related with semantic cross-walking, it is just the stable basis for the complexity of the future global data domain.

IV. Note on Persistence

V. Cerf is warning for the Dark Digital Age[60] we can enter when we will not take appropriate measures. There is much talk about making data more persistent, however, it is mostly not spelled out whether we speak about strategies for 10 years, 100 years, or even longer time periods. In this paper I do not want to elaborate on this aspect implying long-term data curation which will not be trivial to achieve. Here I want to focus on an aspect which is often overlooked: the stability and long-term persistence of the relations between digital objects which we are creating manually or increasingly often automatically.

Research infrastructure projects such as DISSCO [Koureas 2019] in the area of biodiversity indicate the challenges we will have: they deal with 1.5 billion specimens – now as digital twins – from about 500 different natural history museums. Each digital specimen is part of different contexts such as classifications, relationships to other specimens, is associated with different kinds of observations such as photos, gene sequence data etc. Therefore, the number of relationships for each of these specimens is estimated to be around 30 times as big, i.e. we speak about roughly 50 billion relationships which are important to understand the details of specimens. It would be a disaster when these relationships would be lost, since they incorporate the accumulated research knowledge about nature in the digital age where we cannot go back to paper anymore.

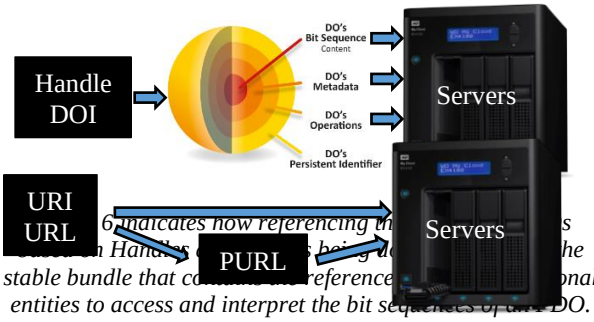
The aspect of long-term stability of relations is widely ignored in our discussions. Librarians and publishers have addressed this issue when preparing themselves for the digital age. They first created the ISBN numbering system which makes a difference between the "book as a work" and the "printed book on a book shelf" (Figure 5). The domain of the "books as works" is nicely separated from the book on shelves, since the first has to be persistent while the latter is ephemeral.

[59] https://www.dona.net/specsandsoftware
[60] https://www.theguardian.com/media-network/2015/may/29/googles-vint-cerf-prevent-digital-dark-age

With Fair Digital Objects we apply an equivalent 2-step strategy. Global, unique and persistent identifiers that resolve associated Kernel Attributes to machine actionable metadata point to all essential information components of an FDO (Figure 6). The attributes will then contain path and other crucial information that is needed to access and interpret FDOs content. This principle is also applied by the publishers which request to associate a DOI issued by the service providers organised in the International DOI Foundation to any electronic publication. The rational behind this is that the publishers will take care that the DOIs based on the Handle System will be persistent. It should be noted here that "persistence" is not only a characteristic achieved by technological choices, but mainly a community responsibility and effort.

To support data labs where the usage of DOIs is not the prefered choice currently more than 3000 Handle services[61] have been established worldwide. It is up to the community behind those services to guarantee persistence. In several countries national data providers and in some large research organisations data centres have taken responsibility to run such services[62]. Handles/DOIs have the capability to include references to an unlimited number of copies of the various components, however, requesting an effort from the remote repositories to change update information in case of changes. This will only work out if the processes in repositories will be automated.

Updating values of attributes associated with PIDs is a highly sensitive operation, since wrong code or operations could destroy all crucial information. Therefore, in the case of the Handle System such operations are highly protected. Each record has a clear owner and for management operations a public key infrastructure (PKI) system is being used to protect records against unallowed access.



6 indicates how referencing th… s … he stable bundle that co… reference… onal entities to access and interpret the bit se… DO.

Figure 7 indicates how referencing in the Web is being done. In general, references occur as URLs encoding ephemeral semantics such as location. PURLs compensate for this deficit.

In the case of URIs, which in general are URLs, the mechanism is different. URLs point directly to locations at specific servers, i.e. they include location information that will change over time and is therefore hardly persistent. To cope with this deficite the Web community invented the PURL system, for example, which allows an indirection step (Figure 7). As in the case of Handles/DOIs repositories are responsible to submit updated information.

However, a PURL cannot cope with multiple copies which is increasingly important in our digital domain for several known reasons and with the association of types and other attributes. The Web community has addressed this gap by suggesting to use Signpost[63] (Figure 8). Ideally a URL points to a PURL which then redirects to a structured HTML landing page which contains standardised and thus machine actionable attributes, and therefore can direct machines to other references. With this mechanism basically the same goals are being achieved as by using standardised kernel attributes in the Handle/DOI case. The major differences can be described as follows: (1) Handle resolution offers immediately structured information while in the Signpost case another indirection step is required

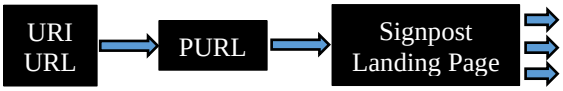


Figure 8 indicates another deficit in the Web-protocol stack which is the lack of a standardized and machine actionable information where different information about a digital object can be found. Signposted landing pages are meant to overcome this gap.

[61] It should be noted that the differences between DOIs and other types of Handles can be found mainly in the differences between the fee structure, the degree of flexibility, for example, in assigning kernel attributes and performance considerations.
[62] https://www.pidconsortium.net/
[63] https://signposting.org/

and HTML needs to be parsed. (2) The directly implemented security mechanisms for Handles/DOIs are stronger compared to those for web-pages.

V. Conclusion

In this paper we started describing our hopes that the FAIR Principles and the FAIR Digital Objects are the needed attractors to come to an Integrated and Interoperable Data Domain (I2D2) which is highly needed to come to a new stable situation for digital objects and not take the risk to enter a Digital Dark Age. I still believe that this is the way to go, however, we still seem to be far away from fast agreements on a core for the global I2D2.

I described then the situation in the data labs where the mass of data is being managed of which only a small amount will be associated with publications (<10%). Not so much has been changed yet with respect to data management efficiency and interoperability despite that researchers are provided by increasingly powerful tools to carry out their research. In general, researchers are not interested in standards that may have a long-term relevance. Due to the pressure on them to show relevant research results they will focus on solutions that are available on short term. Impulses for innovation, if they include the risk of disruptive phases, need to come from other stakeholders.

I then explained why we believe that FDOs can be the solution we are looking for to implement FAIR data and that major components such as, for example, the Digital Object Interface Protocol and the Data Type Registry are ready to be used. Some communities have started to make use of the FDO concept. However, we need to admit that (a) stricter specifications and (b) reference implementations with a reasonable size are yet missing. Only these latter will convince policy makers that the suggested approach will work and scale. Many other voices especially from an IT side can be heard that are not convinced about the FDO approach. Some colleagues believe that a stepwise improvement of the existing service landscape will remove the critical roadblocks on the way towards an I2D2, overlooking that services populate an infrastructure but do not define an integrative core. Others argue that we should leave leadership to big cloud companies observing the huge amounts of investments they currently do, overlooking that cloud systems do not solve the FAIR challenges and that big companies are not interested in achieving an open data domain. Again, others argue that the Web does all we need without knowing even the details as suggested by, for example, Signpost and overlooking speed and security aspects.

Also, we see that the FDO user community is currently fragmented. Various groups and initiatives are trying out the FDO concepts relying on Handles/DOIs for stable referencing but yet do not have an effective global forum to exchange experience, to increase power to achieve necessary changes in the service landscape and to work on additional specifications for standardisation. The RDA Data Fabric IG is focusing on FDOs and is an excellent platform for exchange, however, it lacks the power to set standards and push developments. In this respect the FDO community needs to take urgent action.

Acknowledgment

I need to thank all those who contributed until now to the specification of Digital Objects and here in particular R. Kahn for his early papers and inspirations, a variety of RDA groups, the GEDE group with many excellent experts from research infrastructure projects and those who contributed to the Paris and other workshops[64].

References

Beagrie, N. (2008). Keeping Research Data Safe - JISC Research Data Digital Preservation Costs Study. APA-Conference Budapest.

Berg-Cross, G., Ritz, R., Wittenburg, P. (2016). DFT Core DFT Core Terms and Model. http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318

Committee on Toward an Open Science Enterprise - Board on Research Data and Information (2018). OPEN SCIENCE BY DESIGN - Realizing a Vision for 21st Century Research. https://www.nap.edu/read/25116/chapter/1

Jeffery, K., Wittenburg, P., Lannom, L., et.al. (2021). Not Ready for Convergence in Data Infrastructures. Data Intelligence. Vol.3:1. https://doi.org/10.1162/dint_a_00084

[64] https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop

Kahn, R., Wilensky, R. (1995). A Framework for Distributed Digital Object Services. http://www.cnri.reston.va.us/k-w.html

Kahn, R., Wilensky, R. (2006). A Framework for Distributed Digital Object Services. https://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf

Koureas, D., (2019). DISSCO – Distributed System of Scientific Collections. https://github.com/GEDE-RDA-Europe/GEDE/blob/master/FAIR%20Digital%20Objects/Paris-FDO-workshop/GEDE_Paris_Session%202_%20Koureas.pptx

Herzcog, E., Mons, B., Lannom, L., et.al. (2019). Report Paris Meeting on Moving Forward to Data Infrastructure Convergence. https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop

Schultes, E., Wittenburg, P., (2018). FAIR Principles and Digital Objects: Accelerating Convergence on a Data Infrastructure. In DAMDID Conference on Data Analytics and Management in Data Intensive Science, 2018.

Strawn. G., personal communication (see also Jeffery 2021)

Stehouwer, H., Wittenburg, P., (2016). RDA Europe : Data Practices Analysis. http://hdl.handle.net/11304/6e1424cc-8927-11e4-ac7e-860aa0063d1f

Wilkinson, M.D., Dumontier, M., et. al. (2016). The FAIR Guiding Principles for Scientific Data Management and Stewardship. Scientific Data, Vol. 3, Article Number 160018

Wittenburg, P., Strawn, G. (2018). Common Patterns in Revolutionary Infrastructures and Data. http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0

Open Research Data and Innovative Scholarly Writing: OPERAS highlights

Elisa Nury
Digital Humanities +
SIB Swiss Institute of Bioinformatics
Lausanne, Switzerland
elisa.nury@sib.swiss

Claire Clivaz
Digital Humanities +
SIB Swiss Institute of Bioinformatics
Lausanne, Switzerland
claire.clivaz@sib.swiss

Marta Błaszczyńska
Institute of Literary Research of the Polish Academy of Sciences
Warsaw, Poland
marta.blaszczynska@ibl.waw.pl

Michael Kaiser
Max Weber Foundation
Bonn, Germany
Kaiser@MaxWeberStiftung.de

Agata Morka
European Coordinating Office
Open Access Books
Berlin, Germany
agata@openbookpublishers.com

Valérie Schafer
C²DH
University of Luxembourg
Luxembourg, Luxembourg
valerie.schafer@uni.lu

Jadranka Stojanovski
Department of Information Sciences
University of Zadar
Zadar, Croatia
Centre for Scientific Information
Ruđer Bošković Institute
Zagreb, Croatia
jadranka.stojanovski@irb.hr

Erzsébet Tóth-Czifra
Open Science
DARIAH-EU
Berlin, Germany
erzsebet.toth-czifra@dariah.eu

Abstract*—*We present here highlights from an enquiry on the innovations in scholarly writing in the Humanities and Social Sciences in the H2020 project OPERAS-P. This article explores the theme of Open Research Data and its role in the emergence of new models of scholarly writing. We examine more closely the obstacles and fostering conditions to the publication of research data, both from a social and a technical perspective.

Keywords—Open Research Data, Scholarly writing, Academic publishing, Innovation, Social Sciences, Humanities

I. INTRODUCTION

Since the last decade, new models of scholarly writing have emerged alongside the practice of sharing Open Research Data. The transformation has manifested itself in different ways in Social Sciences and Humanities (SSH) and STEM disciplines according to their respective epistemic culture. The SSH have focused more on encoding standards such as the Text Encoding Initiative (TEI) or the integration of multimedia. In STEM, publishing descriptions and providing links to datasets and databases has become a prominent topic, in journals like *Earth System Science Data*[65], or on platforms featuring datasets with observations like *ScienceMatters* founded in 2016[66]. Certain fields such as Neurosciences, Astronomy or the Life Sciences have already been engaged in data sharing and open science practices for several years, while the transition also impacted SSH, that goes progressively in the direction to sharing Open Data (Vanholsbeeck et al., 2015).

In this paper, we present first results demonstrating that the production and publication of research data may deeply transform the creation of knowledge SSH. These results have been produced in the H2020 project OPERAS-P (Open Scholarly Communication in the European Research Area for Social Sciences and Humanities – Preparation). OPERAS-P wishes to help prepare "a long-term, evidence-based strategy for the development of [OPERAS] infrastructure and its services", one of the four purposes of the project[67]. OPERAS is an emerging European research infrastructure[68] that aims to address the scholarly communication needs of European SSH researchers with the

[65] See https://www.gbif.org/data-papers.
[66] https://www.sciencematters.io/. About data and publication process, see Parsons and Fox (2013).
[67] See https://www.operas.unito.it/projects/operas-p/. The H2020 project OPERAS-P is the preparatory phase of the building of OPERAS.
[68] Since the end of 2019, OPERAS has been founded as an AISBL, *Association internationale belge de droit public*, and counts today 56 organizations from 17 countries; website: https://operas-eu.org.

appropriate infrastructure. It supports the successful implementation of the emerging global, European, and national data policies, along with the European Research Infrastructures (ERICs) in SSH[69].

Within the OPERAS-P framework, an enquiry on the transformation of scholarly writing is led by the Institute of Literary Research of the Polish Academy of Sciences (IBL PAN), in collaboration with five other institutions[70]. In the context of this paper our research questions are: What are the necessary conditions for research data to become new scholarly models of writing? What is propelling us in that direction? What is preventing one from accommodating the novel means of scientific writing to one's own practices? After conducting an overview of the existing literature on innovative scholarly writing practices, the team has led about forty qualitative interviews with diverse stakeholders: scholars, editors, publishers, librarians. The interviews have been transcribed and translated into English when needed. We present here the first highlights of this enquiry.

II. Overview of the Current Landscape

The publication of research data comes within a larger context. One aspect is the adoption of the digital medium for both reading and writing, an important milestone in the history of written communication (Vandendorpe, 2013). However, there is a gap between scholarship creation practices, which have adopted the digital medium, or even create complex digital scholarly objects as outcomes, and publication mechanisms which are still very close to the digital surrogates following paper-based publishing paradigm (Tóth-Czifra, 2019). Still, new possibilities offered by digital medium have encouraged the growth of innovative models of writing and publishing. New trends are emerging, such as demands for publishing multimodal content alongside the text, or a growing interest in research data among SSH scholars as well as funders who now often require different forms of data sharing and Data Management Plans (DMPs) from applicants. Changes in conducting research, managing data and reporting on the research results are influenced by the pressure of "publish or perish", the biases and flaws in the peer review process (Lee et al., 2013), or an "insupportable economic model" (Fitzpatrick, 2011).

Publishing research data provides researchers with the benefits of reusing data, being able to reproduce results, and assessing the value of a publication (Pettifer et al. 2011). In SSH, reproducibility is a contested notion, but sharing research data facilitates scholarly transparency to understand where the source ends and where the interpretation starts. Tracking the provenance of data and sources as well as the layers of interpretation that have been added to it is central to SSH research[71]. Scholars in the field of the Life Sciences also pushed the data paper as a new form of scholarly publication following academic standards which describes datasets and the circumstances of their collection, and provide a link to their repository (Chavan & Penev, 2011). The following years have seen the infrastructure being developed for the purpose of publishing datasets and data papers and preserving them for the long term: the creation of open data repositories, such as Nakala in France, DANS in Netherlands or DARIAH-DE Repository in Germany, and data journals published by prestigious academic publishers[72].

The SSH have followed the same direction shortly thereafter. In 2015 the *Journal of Open Humanities Data* started publishing data papers, and the first volume of Brill *Research Data Journal for SSH* came out in 2016. Both journals are Open Access and have collaborations with dedicated data repositories, for instance DANS and *Dataverse Network of Harvard University*, and more general repositories like Figshare and Zenodo. More recently, De Gruyter, with the C^{2}DH at the University of Luxembourg, launched the *Journal of Digital History*[73], an innovative publication platform for "multi-layered" articles that include data, methodology and a narrative layer. Therefore, the data paper is becoming an established form of scholarly writing. The purpose of a data paper is, among other things, to give credit for the effort required to prepare, curate, and contextualize data with the proper metadata. As such, data and its description indeed form a new model of scholarly writing in SSH.

---

[69] DARIAH (Digital Research Infrastructure for the Arts and Humanities), https://www.dariah.eu/; CESSDA (Consortium of European Social Science Data Archives), https://www.cessda.eu/; CLARIN (Common Language Resources and Technology Infrastructure), https://www.clarin.eu/; SHARE (Survey of Health, Ageing and Retirement in Europe), http://www.share-project.org.

[70] DARIAH-EU with SIB (CH) as partner, the Max Weber Foundation (DE), Open Book Publishers, the University of Luxembourg and the University of Zadar (Croatia).

[71] See for example the *Research Data Journal for the Humanities and Social Sciences*; the *Journal of Open Humanities Data Dataverse*; the *Journal of Open Archaeology Data*.

[72] Brill has its own data repository with Figshare: https://web.archive.org/web/20201020000158/https://brill.figshare.com/. Elsevier and Nature have both launched journals to publish research data and data papers in 2014: see http://web.archive.org/web/20200325153944/https://www.elsevier.com/authors/author-resources/research-data, and http://web.archive.org/web/20200325154111/https://www.nature.com/sdata/about. De Gruyter also integrates a widget to publish code alongside articles since 2018: http://web.archive.org/web/20200325154206/https://www.eurekalert.org/pub_releases/2018-07/co-dgp071118.php

[73] See SIBDARIAH03 and 04; all interviews are referred in the final bibliography. See also: https://journalofdigitalhistory.org/en/about

SSH data often depend on artefacts owned by various Cultural Heritage institutions that impose their own policies and copyrights restrictions, or on qualitative interviews that may contain sensitive personal information, which affect the culture of data publication (Tasovac et al., 2020). It should also be highlighted that scholars receive no credit for data publication as such, a point related to national or/and institutional open science policy. Researchers need to present the data into a journal article format. The reasons for this are complex, but one explanation has to do with the lack of scholarly information management systems that are inclusive with digital scholarly objects rather than publication texts. However, the emergence of search engines like Google Datasets[74], discovery systems like the OpenAIRE Research Graph[75], and TRIPLE[76] platform for SSH data discovery and reuse will improve the situation.

III. Conditions, Obstacles, and Fostering Elements

What are the opinions of SSH researchers regarding the publication of research data? During our interviews, the respondents expressed a variety of opinions when asked "What is your opinion about publishing the entire material from a given study in SSH: whole interviews, annotated texts, research protocols, data collected in the research process etc.?" There were doubts related to the time-consuming aspect of academic life that already requires researchers to read, write and peer review articles:

> *"It doesn't make any sense. I already don't have time to read all the articles I want to read. I understand it intellectually, but given the time I have, I don't think I would take the time to get into an underground area below the article." (UniLux01, 2020)*
>
> *"In History, we are already happy if one person takes time to read what we write! Who is going to read research notes on a subject for which the final monograph or publication will already be read by too few scholars?" (SIBDARIAH02, 2020)*
>
> *"I'm very much in favour of there actually being digital data repositories that allow as much data as possible to be accessed by people who are interested. I think that [...] the accessibility part of the data should be increased online [...]. The problem is that the research data is only relevant to a very small portion of the readers. That is to say that, in fact, it's like footnotes, footnotes are very important for the epistemological and ethical guarantee of the work." (SIBDARIAH08, 2020)*

As summarized by the last quotation, research data are very important as a guarantee of valid research processes, but they may be used only by a minority of scholars, which means that it represents a large investment in time for what might seem like little return. However, the availability of research data is crucial for the accuracy and reliability of peer review. On the other hand, interviewees also stressed the importance of transparency, with caveats about privacy, copyrights, and reuse of data:

> *"Yes, I would. It is even necessary, or it is becoming more and more mandatory in certain cases. Today, transparency is very important." (SIBDARIAH01, 2020)*
>
> *"I'm really in favour of that. [...] I think it depends on the field a lot but in my field having to publish alongside your manuscript which should really be your reflections on the data that you've collected. Publishing the data and publishing what you did with that data so publishing some form of code that you used to get from data to conclusions, and to create visualizations, and tables, and stuff like that I think that would be very beneficial […]. Also, it would make the whole process much more transparent and it would not eliminate, it would reduce the margin for foul play." (UNIZD01, 2020)*
>
> *"One of the things that we come up against is that, culturally, people expect transparency. That becomes dangerous because then you can violate things like privacy. [...] But if I put that stuff out there, scholar X is going to take that data and write that next book that I'm not going to right now. Because the incentives of scholarship are what they are, you still have to be careful about what full publication would look like." (SIBDARIAH07, 2020)*

In some cases, the publication of research data was a necessity in the context of reporting mathematical and statistical experimentations where a traditional article was not sufficient: the sources needed to be made available as the software used and the raw data from the experiment. Ideally, the research data would be published with the same standards of rigor as traditional academic publications, however the peer review of data would raise an enormous

[74] https://datasetsearch.research.google.com/
[75] https://www.openaire.eu/
[76] https://operas.hypotheses.org/projects/triple

challenge in terms of the workload it would impose on reviewers that may already be short on time. Even researchers who agree that research data should be peer-reviewed admit that realistically, we will never be able to do that at scale:

> *"And really the labour involved in evaluating these things just goes through the roof. And I just don't think people are going to have time to do that kind of evaluation for every piece of digital scholarship that emerges in the next few years. So, I think there's a looming crisis for the labor of peer review." (DAE03, 2020)*

The interview samples above highlight an interest in the need to publish research data in SSH, although there are limitations and obstacles. The question for us was to better understand those obstacles and identify the conditions that could foster the publication of data. From the preliminary analysis of our interviews, as well as the scientific literature on the subject, we identified two broad topics, a social one and a technical one. The social one covers research challenges of the relationship between the various stakeholders involved in research data management – Galleries, Libraires, Archives, and Museums (GLAM) professionals, data processors such as Digital Humanities labs, research teams, repository managers, data stewards and publishers – and the current misalignments between data sharing policies and academic rewarding criteria. The second area encompasses technical problems of data curation and storage.

A. Research Challenges

One of the biggest obstacles to innovative scholarly publications, such as research data, is the "reward structure", that is, how research is assessed and credited within academic institutions (Moore & Adema, 2020). In fact, scholarly publications are not only about disseminating knowledge, but they also play an important role in assessing academic success, in evaluating and promoting researchers. However, the currency of academic credit is not money, it is "reputation" (Andrews, 2017), largely measured by a series of analytics: the number of publications, the number of citations received by those publications, the impact factor of journals where they are published, the prestige of the publisher, or the publication type (in certain cases SSH scholars receive more credit for a book than an article, for example).

As a result, there is a strong incentive for researchers to publish traditional scholarship in prestigious venues for their field of study, to receive the credit, they need to get funding, a stable position, and to advance their career. In this context, and within the time constraints of research projects, this situation creates a tension for scholars to balance the need for traditional publications and desire for innovative practices, as highlighted during our interviews: publishing data is time-consuming, which is a disadvantage (SIBDARIAH10); the work often has to be done twice, once for preparing and depositing the digital output, and once for a more traditional publication (SIBDARIAH01). These problems are not new and have been repeatedly highlighted in the field of Digital Humanities: "Digital humanists find, time and time again, that they are expected to perform twice the labor of traditional scholars; once for the work itself and once again for its evaluation" (Eve, 2020; see also Baillot, 2016 and Fitzpatrick, 2011).

Closely related to the concept of academic credit is "data hugging" [77], the opposite of data sharing. Scholars are often reluctant to release data, as they must cope with a culture of perfection, and they dare only present data of utmost perfection. Since good publications bring credit, this is a valid fear, especially considering that SSH scholars may take years to gather their data, to fully analyze them, interpret and write monographs. On the other hand, the emerging practice of data publications carry the potential to immediately claim early attribution and credit the authors which deconstruct the dynamics that fuel the current data hugging phenomena. The publication of research data would also improve the recognition that collecting data is already doing valid academic research (Truan, 2019).

One solution to encourage data sharing would be for scholars to get acknowledgement early in the process, and not after, for instance, the final monograph is published. In STEM, there is a culture of sharing preprints, and it can be adopted for SSH articles as well, but for monographs we need more innovative writing models: for instance, web books where chapters can be published consecutively (SIBDARIAH01)[78]. Promoting a data citation culture would bring research data into the spotlight: scholars cited for their data would receive academic credit, which would in turn be considered by funders and by committees for promotions. But to say it in short with Tóth-Czifra, "the information systems measuring the (re)use and impact of digital tools and scholarly data are still in their infancy" (2020).

However, it is often not the reservations of researchers that are the main obstacle to developing data sharing practices. The nature of SSH data, already discussed in the previous section, has implications in the areas of copyright and privacy. If most of the research data used in a research project consists of existing artefacts, such as objects owned by a third party (artworks, texts, audiovisual materials), the opportunities for publishing the dataset will largely be

[77] The term has been coined by Dr. Hans Rosling in a quite famous talk about the "Data-Hugging Disorder" given in May 2009 (see Frydman, 2009).

[78] Web books are books presented in the format of a website, which are regularly updated (Fauchié, 2016).

constrained. While such information should be explicitly provided (Angelaki et al., 2020), sometimes it is even difficult to verify the ownership of an object or to check on what license it was originally shared. Furthermore, social scientists conducting surveys or interviews will have to obtain an explicit consent from respondents to reshare the data (raw answers, transcriptions or recordings). In many cases time-consuming – and sometimes complex – anonymization and pseudonymization procedures will be necessary.

This also raises the question of times and temporalities. More and more funders are requiring provision of a DMP from the very beginning of the project, but this raises complementary questions about the maintenance of data as well as derived research data and scientific outputs at the end of the project. Willingness is often not sufficient when researchers at the same time must deal with long-term needs of data preservation, mid-term funding and the paradoxically short life cycles of the data, formats, devices, tools and platforms. As demonstrated in Barats et al. (2020), there are "a number of different temporalities […] and multi-stakeholder issues that require collective reflection to clearly identify the actors and locations that are best adapted to implement and support the challenges of data sustainability."

In summary, the challenges described above therefore require for the academic community to rethink research assessment practices, to change the metrics of academic credit, and to take into account the time and amount of labor necessary for the publication of quality research data. There are initiatives in that sense, for instance the SF Declaration on Research assessment (DORA), that has been signed by many institutions in Europe, or the HuMetricsHSS initiative[79]. This claims for new peer review practices, as developed in Digital Humanities for evaluating scholarship (Baillot, 2016). Peer reviewed data papers can be part of the solution, but we need to develop criteria and procedures that certify the quality of research data.

B. Data Storage and Curation

On the side of the more technical problems, there is a need for infrastructures to access and store data, along with the relevant metadata that is necessary to interpret and reuse data. A first concern is about access to data. As noted by Rieder & Hofmann (2020), "the concept of observability starts from the recognition of a growing information asymmetry between platform companies, a few data brokers, and everyone else. The resulting data monopoly deprives society of a crucial resource for producing knowledge about itself." Some datasets may be stored behind a paywall and thus only accessible for researchers with funding. This may increase the gap between those who can afford paying to access data and those who cannot.

One challenge that came up during the OPERAS-P interviews is also the fragmented nature of data repositories (Mostern et al., 2016) and the need for a single point for discovery, such as a European wide search engine, e.g., Isidore. As noted by (Gregory et al., 2020), "[b]efore data can be reused, they must first be discovered", and data finding can be hampered by the technical infrastructure (researchers use Google with mixed success) and is also dependent on the researchers' social context. One may add the fragmented nature of data and sources that may be divided up in several repositories and the issue of hosting of complex corpora. In addition, one researcher dealing, for example, with born digital heritage may share derived data, some metadata and permalinks to web archives that are preserved in national institutions, but will not be able to share more, because authors' rights apply. This also highlights the need for interoperability, but the large number of metadata standards can make it complicated, as there is a variety of standards in SSH: general standards (DublinCore), standards for text (TEI), images (IIIF), archival materials (EAD), cultural heritage (CIDOC-CRM), and so on.

Another concern is how to link various outputs of a project, the data, the articles, the code, the source materials etc. The common practice now is to use persistent identifiers such as DOI. This also has implications for publishers and libraries: publishers will have to deal with projects that have multiple outputs (SIBDARIAH07) – how do those outputs hold together as a unified, complex entity? How do librarians' catalogue and provide access to a publication made of multiple parts? Are PIDs sufficient? Can they be used to keep track of citations for the data? There seems to be no accurate information management system in place for that[80]. Moreover, while certain writing tools allow for a greater integration of data into the scholarly text, often only the minority of researchers use it. One of the interviewees feels that in their field:

> *"the relation between data and writing is still a bit conflictual because people write in Word and there's no way to integrate nicely your statistics or your lines of code and to have good synchronization between the data and the text you're writing or to provide interaction between the text and the reader." (IBL08, 2020)*

FAIR data and current trends in Open Science underline the possibilities and opportunities of use and reuse of data but this also raises other challenges. First, there is a need for new models of peer review, as scholars who are

[79] https://humetricshss.org/. See also https://www.dariah.eu/activities/working-groups/impact-factors-and-success-criteria/.
[80] See the *Journal of Open Humanities Data*: https://web.archive.org/web/20201027160132/https://openhumanitiesdata.metajnl.com/about/.

really able to evaluate these data and review them may be rare. One may also push for an interdisciplinary peer review, mixing several levels of digital, engineering, and scientific skills. Innovative scholarly outputs including datasets may be challenging for the readers who are more used to traditional publications. In addition, some interviewees underlined their lack of time to read all papers related to their field of research, therefore being very skeptical about their availability to go deeper in the reading and discovery of data. Regarding reuse, the point is not just about sharing data, but also about contextualizing them to allow a genuine reuse. Finally, there is also a challenge of maintenance - and eventually repair. This is often a part which is forgotten in the process and may create plenty of data lakes that are unexploitable because they lack transparency, contextualization, updates, etc. Morselli & Edmond (2020) note that work is lost due to resource and technical challenges, but they also illustrate how the sustainability of the results of digital research projects can be thought of as a process instead of an end product that involves more than ensuring a long-term hosting data infrastructure.

IV. Conclusion

A complex environment is at stake, consisting of data brokers, engineers, researchers, publishers, funders, several kinds of data, as well as several legislative environments in an internationalized world. There is a need for incentives at all levels and for understanding that this investment has a cost (may it be in terms of funding, maintenance, engineering, time, etc.), but this may cost less than losing vast amounts of data and research. Capitalizing on the existing infrastructures may in a mid-term perspective create a strong reward. Parallel to these systemic changes we also need a cultural shift to view the publishing of data as a valuable scholarly output.

Conceptual models may be needed to help design intelligent and efficient solutions. One example is the data scope concept by Hoekstra et al., which, for example, suggests “classifying data” to group them “to reduce complexity”. This adds a level of abstraction to the data” (2018). How do we shrink the gap between those who are able and those who are not to share data, and how might we direct a whole generation of researchers to this transition? Are there new skills that all researchers should develop, and will these new skills create new research profiles and kinds of support? Data stewardship (Mons, 2018) is developing and may in the future be more and more pertinent, becoming a new fundamental position.

Acknowledgment

Our deep gratitude goes to Vera Chiquet (DaSCH, University of Basel), who was employed in June and July 2020 at SIB to help us to gather data from three interviews done in the German-speaking part of Switzerland. We also warmly thank all the people who agreed to be interviewed in this project.

References

[1] Andrews, T. (2017, video). After the Spring: Digital forms of scholarship and the publication ecosystem [Conference Keynote] (26.09.17). *20th International Conference on Electronic Publishing Positioning and Power in Academic Publishing: Players, Agents and Agendas. 7–9 June 2016 in Göttingen, Germany, Göttingen*. https://www.youtube.com/watch?v=44FaPFjYxMk

[2] Angelaki, G., Badzmierowska, K., Brown, D., Chiquet, V., Colla, J., Finlay-McAlester, J., Grabowska, K. et al. “How to Facilitate Cooperation between Humanities Researchers and Cultural Heritage Institutions. Guidelines.” *Digital Humanities Centre at the Institute of Literary Research of the Polish Academy of Sciences, 10 March 2019*. https://doi.org/10.5281/zenodo.2587481.

[3] Baillot, A. (2016, conference communication). A certification model for digital scholarly editions: Towards peer review-based data journals in the humanities. Conference communication. *Digital Scholarly Editing: Theory, Practice, Methods*, Université d'Anvers, Oct 2016, Anvers, Belgium, https://halshs.archives-ouvertes.fr/halshs-01392880

[4] Barats, C., Fickers, A., & Schafer, V. (2020), Fading Away... The challenge of sustainability in digital studies. *Digital Humanities Quarterly* 14 (03), http://www.digitalhumanities.org/dhq/vol/14/3/000484/000484.html

[5] Chavan, V., & Penev, L. (2011). The data paper: A mechanism to incentivize data publishing in biodiversity science. *BMC Bioinformatics*, *12*(S15), S2. https://doi.org/10.1186/1471-2105-12-S15-S2

[6] DAE03: Tóth-Czifra, E. (2020). Transcript Interview DAE03 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/10.34847/nkl.3e9bjm89

[7] Eve, M. P. (2020). Violins in the Subway: Scarcity Correlations, Evaluative Cultures, and Disciplinary Authority in the Digital Humanities. In: Edmond, Jennifer (ed.) *Digital Technology and the Practices of Humanities Research*. Cambridge: Open Book Publishers, 105-22.

[8] Fauchié, A. (2016, blog post). Le livre web, une autre forme du livre numérique (24.10.16). *Quaternum.Net*, https://www.quaternum.net/2016/10/24/le-livre-web-une-autre-forme-du-livre-numerique/

[9] Fitzpatrick, K. (2011). *Planned Obsolescence: Publishing, Technology, and the Future of the Academy*. NYU Press; JSTOR. www.jstor.org/stable/j.ctt9qg9mh

[10] Frydman, G. (2009), “e-Patients Demand: Put An End To Data-Hugging Disorder” (04.06.09), *Society for Participatory Medicine. Transforming the Culture of Patient Care Blog.*

https://participatorymedicine.org/epatients/2009/06/e-patients-do-not-suffer-from-database-hugging-disorder.html

[11] Gregory, K. M., Cousijn, H., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Understanding data search as a socio-technical practice. *Journal of Information Science*, *46*(4), 459–475. https://doi.org/10.1177/0165551519837182

[12] Hoekstra, F. G., Koolen, M., Burgers, J. W. J., van Faassen, M., Nijenhuis, I. J. A., & Derks, S. (2018, report). Inleiding Data Scopes. *Huygens ING*. https://doi.org/DOI: 10.13140/RG.2.2.23812.01928

[13] IBL08: Błaszczyńska, M. (2020). Transcript Interview IBL08 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/10.34847/nkl.e87827w3

[14] Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

[15] Mons, B. (2018). *Data Stewardship for Open Science. Implementing FAIR Principles.* Routledge Press.

[16] Moore, S., & Adema, J. (2020, blog post). COPIM Experimental Publishing Workshop—Part 1: Inhibitions Towards Experimental Book Publishing (19.10.20). *COPIM*. https://copim.pubpub.org/pub/experimental-publishing-workshop-part-1/release/2

[17] Morselli, F., & Edmond, J. (2020). Sustainability of digital humanities projects as a publication and documentation challenge. *Journal of Documentation*. https://doi.org/10.1108/JD-12-2019-0232

[18] Mostern, R., and Arksey, M. Don't Just Build It, They Probably Won't Come: Data Sharing and the Social Life of Data in the Historical Quantitative Social Sciences. *International Journal of Humanities and Arts Computing* 10 (2016), 205–24. https://doi.org/10.3366/ijhac.2016.0170

[19] Parsons and Fox (2013), Is data publication the right metaphor? *Data Science Journal* 12, https://www.jstage.jst.go.jp/article/dsj/12/0/12_WDS-042/_pdf/-char/en

[20] Pettifer, S., McDermott, P., Marsh, J., Thorne, D., Villeger, A., & Attwood, T. K. (2011). Ceci n'est pas un hamburger: Modelling and representing the scholarly article. *Learned Publishing*, *24*(3), 207–220. https://doi.org/10.1087/20110309

[21] Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1535

[22] SIBDARIAH01: Nury, E. (2020). Transcript Interview SIBDARIAH 01 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/11280/d4aafc66

[23] SIBDARIAH02: Nury, E. (2020). Transcript Interview SIBDARIAH 02 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/11280/133a7853

[24] SIBDARIAH03 and 04: Nury, E. (2020). Transcript Interview SIBDARIAH 03 and 04 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/11280/d2905a1b

[25] SIBDARIAH07: Nury, E. (2020). Transcript Interview SIBDARIAH 07 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/11280/a7f68f73

[26] SIBDARIAH08: Nury, E. (2020). Transcript Interview SIBDARIAH 08 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/10.34847/nkl.64b9ratx

[27] SIBDARIAH10: Chiquet, V. (2020). Summary Interview SIBDARIAH 10 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/11280/ac0ffef5

[28] Tasovac, T., Chambers, S., & Tóth-Czifra, E. (2020, position paper). *Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper*. https://hal.archives-ouvertes.fr/hal-02961317

[29] Tóth-Czifra, E. (2019, blog post). Laying the Pavement Where People Actually Walk: Thoughts on Our Chances of Bringing Scholarship Back to the Heart of Scholarly (23.10.19). DARIAH Open. https://dariahopen.hypotheses.org/645

[30] Tóth-Czifra, E. (2020, blog post). 10 practical tips to fight against the culture of non-citation in the humanities (03.03.2020). *DARIAH OPEN blog*. https://dariahopen.hypotheses.org/747

[31] Truan, N. (2019, blog post). "How to make the most of your publications in the humanities?". (22.01.19), *Ici et là*. https://icietla.hypotheses.org/994

[32] UniLux01: Schafer, V. (2020). Transcript Interview UniLux01 (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/10.34847/nkl.66b98590

[33] UNIZD01: Zauder, K. (2020). Transcript Interview UNIZD01. (H2020 OPERAS-P). *[Text] Nakala*. https://doi.org/10.34847/nkl.6aaf8jl9

[34] Vandendorpe, C. (2013). Reading on Screen: The New Media Sphere. In S. Schreibman & R. Siemens (Eds.), *A Companion to Digital Literary Studies*. Wiley Online Library. https://doi.org/10.1002/9781405177504.ch10

[35] Vanholsbeeck, M., Engels, T., & Istenič Starčič, A. (2015). Guidelines for Data Sharing and Data Citation in Social Sciences and Humanities Journals Perspectives and Insights from the Cost Action Enressh. *Archives et Bibliothèques de Belgique - Archief- En Bibliotheekwezen in België* 106, 83–92.

Data Life-Cycle Management's Massive Open Online Course on Research Data Management

Silas Krug
Information Sciences Dept.
Geneva School of Business Administration
Carouge, Switzerland

Basma Makhlouf Shabou
Information Sciences Dept.
Geneva School of Business Administration
Carouge, Switzerland
ORCID 0000-0003-0980-0517

Abstract*—*In this paper, we will first have a quick look at what is a MOOC (Massive Open Online Course) and e-learning in general. Then we will present the five main modules which structure the course, who is responsible of which one, and what are the three optional specialized modules. We will then talk about the established partnerships for the project, with Swiss MOOC Service and DevPro. Then we will have a look at the content types which you will find in this course, and finally list the future project's milestones.

I. Introduction

In its mission to be the reference for Research Data Management (RDM) in Switzerland, DLCM (Data Life-Cycle Management) must provide accurate learning tools. The European Commission appeals to include e-learning in HEI's (Higher Education Institution) training strategies, which is why one of the DLCM's deliverables is to create MOOC (Massive Open Online Course) on RDM. This MOOC will have the following objectives :

- Provide accessible online materials to researchers,
- Support HEI units that assist researchers,
- Assist swiss and external researchers,
- Offer a complementary resource to OLOS users,
- Promote DLCM and OLOS services and expert network.

The MOOC belongs to family of e-learning methods. Several definitions have been given to the e-learning, such as the one from the Commission of the European communities in 2001, « *[the e-learning is the] use of new multimedia technologies and the Internet to improve the quality of learning by facilitating access to resources and services as well as remote exchanges and collaboration* », the one from the Walloon Telecommunications Agency in 2008, « *E-learning (e-learning) : online learning centred on the development of skills by the learner and structured through interactions with the tutor and peers* », and the one from the Joint Information Systems Committee (JISC) in 2017, « *learning facilitated and supported through the use of technology to support learning as part of a 'blended' approach (a combination of traditional and e-learning approaches), to learning that is delivered entirely online. Whatever the technology, however, learning is vital element* ».

There are several varieties of e-learning, notably depending on the teaching timing which can be simultaneous or not. The MOOC is one of these sorts and here is what the words composing the acronym imply :

- Massive : without pre-requisites, the number of enrolments is potentially higher than for universities or schools.
- Open : because registration is open to all, without any special conditions (registration at a university, linked to a level of study or professional status, etc.)
- Online : the whole course is online, courses, homework and exercises, but also registration and exams.

Also, two kind of MOOC can be distinguished :

- xMOOC is designed as a classic course with a teacher-defined progression.
- cMOOC is decentralized. The material is made available to each learner who chooses his or her own course.

DLCM's MOOC will be somewhere between both concepts, as the contents will be freely available at any time, but its five general modules will be involved in an evaluation process, with a defined time schedule, leading to a certification.

II. DLCM'S MOOC'S GENERAL MODULES

The general modules are the following :

A. *GM1 – Research Data Governance*

This first module introduces RDM, giving its context in the institutions' governance. It clarifies RDM's role and impact on the institutions' policies and the work done inside.

B. *GM2 – Active Research Data Management*

The second module concentrates on the life-cycle of the research data in its active part, meaning its management while the actual research is conducted. It starts on several details regarding the capture, collection and creation of the data. It then explains how it is structured, described and as a final step, analyzed and visualized.

C. *GM3 – Research Data Sharing*

RDM's problematic is highly related to another one : the Open Science, and its philosophy. Open research data is at the crossroad of both problematics and has to consider ethical and legal aspects in order to be led correctly. The data sharing is also deeply related to practical considerations such as the interoperability factors ; these are also studied in this module.

D. *GM4 – Research Data Preservation*

This fourth module basically talks about the last step of the Active Research Data Management, but it requires to be studied separately. Its content will review basics of digital archiving, such as the Open Archival Information System standard (ISO 14721). This module presents the swiss official solution OLOS.

E. *GM5 – Data Management Plans*

Considering the fact that the usually most familiar aspect of RDM for researchers is the Data Management Plan (DMP), for the simple reason they are regularly appealed to produce one to obtain research funding, this specific module is mostly about practical considerations around DMP's creation and use in researchers daily activities. It is focused on DMP addressed to the Swiss National Science Foundation (SNSF), but it will be universally useful for any kind of DMP.

The lead of all these modules have been distributed between very competent personalities in each of the respectively concerned domain. Dr. Basma Makhlouf-Shabou, co-director of the Data Life-Cycle Management project and responsible of the Master of Information Studies at the Geneva School of Business Administration, has major knowledge in information sciences and knows in particular how to manage an institutional governance, since she was project manager of the *Information and organizational governance policy : from design to implementation* project, from 2014 to 2016. Considering these abilities she took the responsibility of the module 1 on research data governance. In the DLCM project, she is also in charge of the Coordination Desk management, e-mail address that receives requests of help from researchers. Current statistics show that the most usual demands are DMP corrections, which is why she also took the lead of the module 5. The lead of the second general module is shared by Dr. Basma Makhlouf-Shabou and Eliane Blumer, coordinator of the EPFL's Research Data Team. Dr. Basma Makhlouf-Shabou is teaching research methodology in her IS Master, and thus is very familiar with the active research data management. Eliane Blumer is also in regular contact with the EPFL's researchers and, in particular, have a deep active life cycle. The third module is taken in charge by Dr. René Schneider, responsible of Information Studies at the Geneva School of Business Management. As a specialist of Linked Open Data, the Open Science and research data sharing largely belongs to his field of competences. Finally, Dr. Pierre-Yves Burgi, director of the DLCM project, and deputy IS director at the University of Geneva is responsible of the fourth module, on Research Data Preservation. Belonging to the development team of OLOS, the DLCM's solution for research data preservation, he is naturally well indicated to lead this part of the course.

III. DLCM'S MOOC'S SPECIALIZED MODULES

There will be four optional modules which can be studied as complement, depending on students' interests. Three of them have exclusive contents, but the fourth will be a compilation of what has been told about life-cycle facilities in each general module. The three exclusive modules are the following :

A. *SM1 – Research Data types and paradigms*

In this module, different types of data, depending on the nature of the research they are resulting from (quantitative or qualitative) and on their file formats (text, picture, sound, mixed, etc.), are studied. The main goal of this specialized module is to make discover the difference of management induced by these data types.

B. *SM2 – RDM fields of study*

This module will review of RDM depending on different academical fields of study. Some fields can have important specificities such as high caution on data sensitivity in medical fields, massive quantity of data in nuclear research or few actual data creation in literary studies.

C. *SM3 – Tools, technologies and services*

This final specialized module will review existing solutions and their features, open source or not, whose use covers the entire data life-cycle or only a specific phase.

IV. CONTENTS' CREATION

This MOOC is created in close contact with Swiss MOOC Service. Like DLCM, Swiss MOOC is a P-5 project of swissuniversities, what makes the collaboration between the two projects obvious. Within the framework of the MOOC project, Swiss MOOC provides in particular four services which are:

- An introduction to the arcane of shooting MOOC videos for the DLCM team as well as for some external contributors.
- The availability of their studio, in the EPFL premises, for the shooting of the videos.
- The post-editing of the videos shot on location.
- And the addition of subtitles to the videos, customizable in four languages.

For practical considerations, and also because the HEG, in a sense thanks to the pandemic crisis, now possess its own video studio, some of the footage will be directly shot there. Some video editing will also not be delegated to Swiss MOOC but done internally at the HEG with the software Camtasia. Finally, the MOOC's hosting will be done in collaboration with Cyberlearn, responsible of the HES-SO continuing formation, which is the official partner for the DLCM trainings' organization.

The didactic contents of the different modules will take various forms. Naturally, the emphasis will be placed on videos, which will consist of interviews and monologues, shot in the Swiss MOOC or HEG film studios, or in iconic locations. Other contents will have a classic PowerPoint format created on the base of a generic template, with or without voice cover to present them, and finally a few simple texts will complete the set. A short quiz is available for each lesson to ensure the learners test their comprehension of the presented content.

Although if the general structure of the modules will vary, each one will offer for between one and two hours of study (depending on the student's level), a bibliography of all the used sources, a follow-up certificate, as well as overall summaries in English, German, French and Italian. An advanced certification with fully assessment by OLOS experts is also available, delivered in collaboration between OLOS, the HEG-Geneva and DevPro. With regard to this last aspect, DLCM's aims at producing deliverables that are accessible to the entire Swiss population, with people who doesn't necessarily speak the project's language. Therefore, all national languages will be represented.

V. CONCLUSION

Today, three further milestones are planned for the MOOC : modules' certification for the end of March 2022, the introduction of a community management before summer 2022, and the completion of the specialized modules for February 2022. The subscription to the course implies that it will be possible to obtain an official certification by succeeding to scheduled exams. While the MOOC will permit to obtain an attestation for free, it will have a cost to subscribe to the advanced certification.

Built in collaboration with very specialists of RDM and MOOC's creation, with an ambitious plan and various didactic formats and experts from all places, DLCM's MOOC join a set of other trainings in the domain of RDM. Indeed, DLCM also organize half-day interactive trainings in collaboration with DevPro, and provides a Coordination Desk (dlcm@hes-so.ch), where any question on RDM can be asked, and receive an answer in the three following days.

This MOOC will probably become a major element of the training in the domain of RDM at the swiss level. No doubt that it will also be a very original experiment and an opportunity to open RDM training to a broader public, very familiar with the most advanced way of studying.

References

AGENCE WALLONNE DES TELECOMMUNICATIONS (AWT). (2008). Qu'est-ce que l'e-learning ? *Site de l'Agence du Numérique* http://www.awt.be/web/edu/index.aspx?page=edu,fr,gui,080,010

ALLEN, I. E., SEAMAN, J. (2008). *Staying the Course : Online Education in the United States, 2008*. Needham : Sloan Consortium, November. https://www.onlinelearningsurvey.com/reports/staying-the-course.pdf

BENRAOUANE, S. A. (2011). *Guide pratique du e-learning : Stratégie, pédagogie et conception avec le logiciel Moodle*. Paris : Dunod.

COMMISSION OF THE EUROPEAN COMMUNITIES. (2001). *Communication from the Commission to the Council and the European Parliament : The eLearning Action Plan – Designing tomorrow's education*. Bruxelle, 28 March. http://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52001DC0172

PLOURDE, M. (2013). MOOC : every letter is negotiable. *Flickr*. 4 April 2013. https://www.flickr.com/photos/mathplourde/8620174342/in/photolist-e6NW3q-dSyZ94-e8JDvy-gTdSqd-gskoKF-eejpte

POMEROL, J-C., EPELBOIN, Y., THOURY, C. (2014). *Les MOOC : conception, usages et modèles économiques*. Paris : Dunod.

ROSCORLA, T. (2012). Massively Open Online Courses Are 'Here to Stay'. *Converge*. 18 July. http://www.centerdigitaled.com/policy/MOOCs-Here-to-Stay.html

ROYAL HOLLOWAY. (2017). Defining E-learning. Royal Holloway Website. 13 October. https://www.royalholloway.ac.uk/e-learning/aboutelearning/what-is-e-learning.aspx

THOT CURSUS. (2016). Plates-formes de e-learning et e-formation – 2016. *Thot Cursus – Formation et culture numérique*. 21 August. http://cursus.edu/institutions-formations-ressources/formation/13486/plates-formes-learning-formation-2016/#.WddDPmc6w8A