Expanding Perspectives on Open Science: Communities, Cultures and Diversity in Concepts and Practices
L. Chan and F. Loizides (Eds.)

doi:10.3233/978-1-61499-769-6-203

Living Digital Ecosystems for Data Preservation

An Austrian Use Case Towards the European Open Science Cloud

Raman Ganguly[a,1], Paolo Budroni[b] and Barbara Sánchez Solís[b]
[a] *University of Vienna, University Computer Center*
[b] *University of Vienna, University Library and Archive Services*

Abstract. This paper will address issues concerning the handling of complex data such as research data, multimedia content, e-learning content, and the use of repositories infrastructures. At the University of Vienna, an ecosystem for digital data preservation and research data management has already been established and will be subsequently be enlarged according to future needs and requirements. in the future. This living digital ecosystem is the foundation for research data management and was implemented from the beginning as a central service according to the FAIR principles as stated in the first HLEG-EOSC [1] report. With the help of ten years of professional experience, a model for digital data preservation was established to address the complexity of heterogeneous data. This was necessary because of different use cases assigned to the interdisciplinary data management team based at the Computer Centre and the Library. The source for the use cases are research projects, their different approach to research and their multifaceted requirements regarding the efficient re-use of data. The usage of this model might be considered as the foundation on which an ecosystem for digital data preservation can be built.

Keywords. visualization of data, repositories infrastructure, digital workflow, research data management, data life cycle

1. Introduction

A solid research data management system is the foundation of open science, open data and open access. Ten years ago, the University of Vienna inaugurated a project with the goal of creating a system which could house digital objects. With the idea of a simple repository to manage data, the project Phaidra (Permanent Hosting, Archiving and Indexing of Digital Resources and Assets) was born. From the beginning, openness was a key motivation and we invited every member of the University, including students, to use the repository. We also provided our technology to other universities and institutions, and so the Phaidra network was created. Today Phaidra is used at research institutions in five different countries.

As more users began to work with the repository, it became apparent that the system should be more flexible and more “agnostic”. For these reasons the management started a reengineering process and to rethink the whole setup. Back to the design phase, the management communicated with stakeholders and were confronted

[1] Corresponding Author, Raman Ganguly, University of Vienna, Universitaetsstraße 7 (NIG), 1010 Vienna, Austria, tel.: +43 1 4277 14189, e-mail: raman.ganguly@univie.ac.at,

with a broad range of research data and use cases. The goal became clear: to address as many stakeholder needs as possible. To meet this goal, it was first decided to refactor the technical structure of the repositories to a micro-services architecture, and second, models for data management were designed, which could be used for different use cases.

Furthermore, the management decided to start a nationwide project in 2014, including as many Austrian research institutions as possible. e-Infrastructures Austria[2] was a federally funded program for the coordinated expansion and continued development of data repositories across Austria, and was made possible by a grant from the Austrian Ministry of Science, Research and Commerce (BMWFW). The program enabled the safe archival and lasting availability of electronic publications, multimedia objects and other digital data from the research and teaching fields. Concurrently, topics relating to research data management and digital archiving workflows were being addressed. This project offered the ideal frame, to discuss and evaluate the present data preservation strategies with Austrian and international experts.

2. Models for data management

Using three different models as a guide, the management redesigned the repository infrastructure, an important starting point for the transition from a simple repository concept to a living digital ecosystem concept. Based on the suggestions of stakeholders, they took a close look at the research process regarding data. The data lifecycle became the focus of the first model.

The second model describes a workflow for the ingestion of entering data into an archiving system and making it available for re-use. When implementing data management from the start, future re-use is already included as the next step in the data lifecycle.

The third model was driven by the idea that no one system fits for all types of data. It suggests how data could be evaluated to determine which archiving system is ideal for storage.

2.1. Data lifecycle model

When publishing data, the data volume is usually small and appropriate archiving formats already exist. However, this is only the top of the iceberg – which as becomes evident when looking at data in the research process. The value of publications rests in their proper preservation, as stated in the PARSE Insight report: „Digital preservation of research data here means the careful storage of all research output in such a way that it remains accessible, usable and understandable over the long term." [2]

To get a closer look at this iceberg, the management Phaidra Management created a model based on the Data Publication Pyramid [3], and added data not directly included in publications, such as inconclusive and negative results. This worked from the point of view of the data and not the publications themselves.

To get a closer look at this iceberg, the management Phaidra Management created a model based on the Data Publication Pyramid [4], and added data not directly

[2] e-infrastructures Austria: https://e-infrastructures.at/en

included in publications, such as inconclusive and negative results. The illustration [figure 1] worked from the point of view of the data and not the publications themselves.

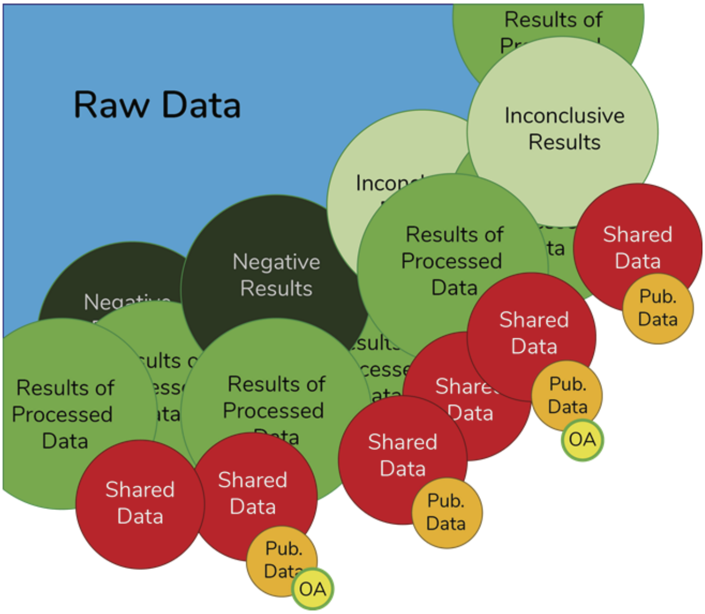

Figure 1. Data life cycle model.

2.2. *Digital workflow model*

The workflow model is the central model for the ideal of the digital ecosystem and is based on the OAIS environment model from the CCSDS [5]. In this simple model, archives are in the centre, surrounded by the producer, consumer and management. The digital workflow model describes the environment for data management more specifically than the OAIS and defines the points at which data will be transferred from one party to another. The involved parties are the data producer, the archiving manager and the data re-user.

According to the model and to the terms of use of the digital archive, which covers all phases of the life cycle of the data, the data producer is the party who creates and owns the data. It is the data producer's role to define in which quality, how long and in which way and in which context, the data and the related metadata can be re-used. Much clarification is necessary and a data management plan is a useful key instrument for the data producer in answering these three key questions. Data management plans are like a project plan for data and like any other living document should be kept up to date throughout the entire project. They are also a useful tool for data management and data inventory in preservation planning. All that is required is machine-readable output from the data management tool.

Information from the data producer is essential for data management. Data management maintains data quality over a specified time and ensures that only authorized users can access the data. During the ingest process, the data and the responsibility for it are transferred from the producer to the manager. The next transfer of responsibility occurs when the data are delivered to the re-user. For the data re-user, the allowed methods of re-use must be clear, so license agreements must be provided to the re-user and accompany the data.

The FAIR principles [6] (Findable, Accessible, Interoperable, Re-usable) principle should act as a guiding principle for the data re-user. This principle should be adhered to starting, at the latest, at the ingest phase. In this phase, data conversion and enrichment occur. In Phaidra this is possible.

The illustration [figure 2] symbolizes a common legal space for the data. It should be a space where there are common terms of use and data can move without legal barriers from one system to another. Clarification of ownership and license agreements at the ingest process help to create a kind of "Schengen Area" for the populations of data being preserved and managed in this area. Policies, governance, rules of engagement and terms of use for services and data management policies on an institutional level complete the clarification of data usage.

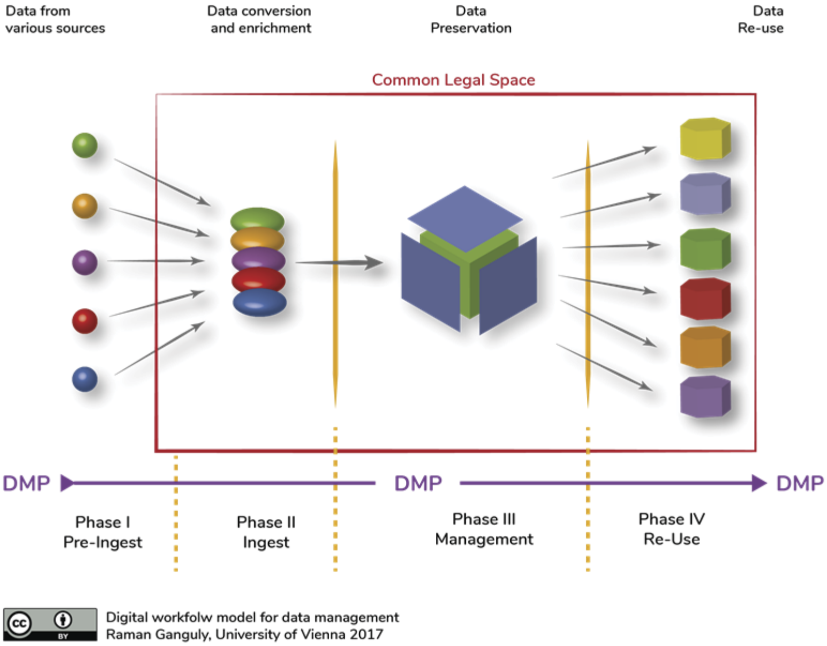
Data from various sources
Data conversion and enrichment
Data Preservation
Data Re-use
Common Legal Space
DMP
DMP
DMP
Phase I Pre-Ingest
Phase II Ingest
Phase III Management
Phase IV Re-Use
Digital workfolw model for data management
Raman Ganguly, University of Vienna 2017

Figure 2. Digital workflow model.

2.3. Evaluation of data

The third model takes a closer look at data management itself and the decision of where to store data. Due to the heterogeneity of research data, one repository or archiving system cannot fit every for all kinds of data. A data manager must decide where to store data in order to *a) maintain quality and b) make data available for re-use*. The attributes of model three can be used to evaluate the data and the archiving system: amount of data, duration of archiving, and complexity of the data format. The attributes of the data should be written in the data management plan, which can then be compared with the features of the archiving system.

The amount of data is easy to measure by counting the files and the file size. Of relevance here is to determine whether there are many small files or only a few large files. This is a major factor when choosing an appropriate storage system. Archiving data is costly and not all data must necessarily be preserved for the long term. For some data, preservation for three to ten years may suffice (e.g for some kind of educational resources), but this should also be carefully planned and executed. The complexity of the data format should be examined from the perspective of data preservation and re-use. Audio and video files are more complex than document files. Databases and software (plus the related contextual and provenance metadata) have special needs in the re-use phase. As the illustrations [figure 3] shows both, the facts of data and the repositories can be added to a grid and compared.

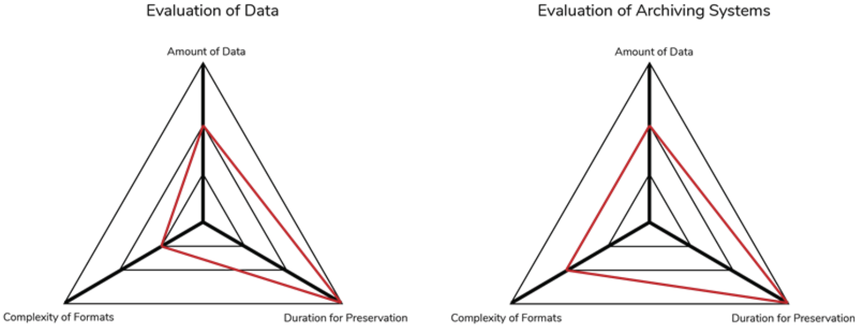

Figure 3. Evaluation of data and repositories.

3. From the repository to the digital ecosystem

The European Science Agenda [6] identifies three layers for data-driven science: data, services and governance. E-Infrastructures cover them all, and is the foundation for data preservation, since data are managed and curated at the infrastructure level. With infrastructure as a foundation, and taking into consideration the three layers proposed by the EU-Commission, services for ingest and re-use are built. This brings value to the infrastructure. Services should be easy to use and appropriate for the use cases of the data producers. Finally, governance is the framework which through appropriate and published policies provides an institutional format for data preservation.

The illustration [figure 4] provides an overview of the strategies used to build a digital ecosystem. Further discussion regarding the infrastructure and service layers will be provided in the following chapters. The governance layer is relevant for designing ecosystems, but is not the main driver and will therefore not be discussed in this paper.

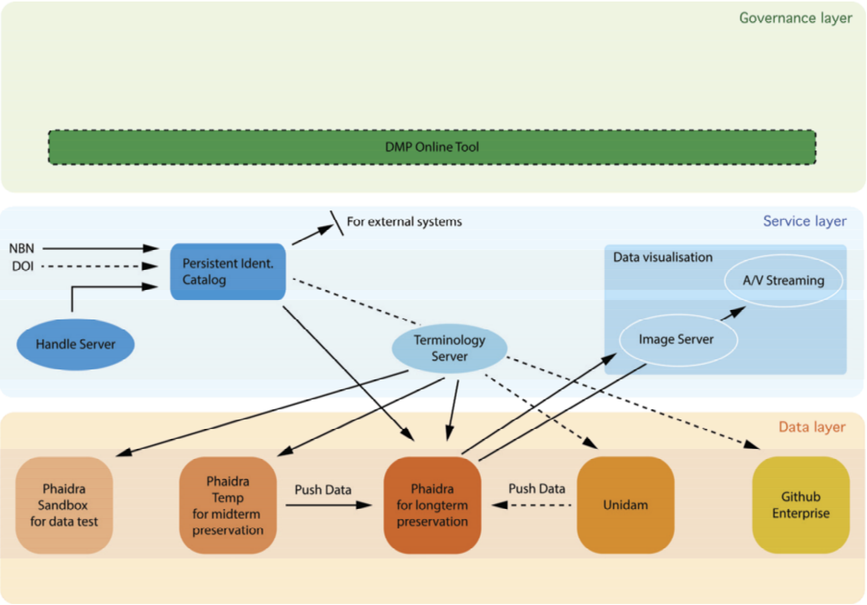

Figure 4. Ecosystem of Phaidra

3.1. The data layer

At the University of Vienna, the management started to create infrastructure for special services and special archiving systems. The team started with Phaidra, the long-term archiving system for generic data, where different bulks of heterogeneous data can be stored. In this repository, all metadata and controlled vocabularies were administrated. Per definition, the long-term archive Phaidra provides a persistent identifier for data which cannot be deleted.

According to certain requirements, some kind of data may be deleted after a defined period of time. Therefore, and in addition, a second repository for midterm archiving was established, where data can be deleted and, in the future, seamlessly transferred to the long-term archive. Currently, an automatic deletion of data after a specified time is not possible. Such a feature requires a better data management plan tool in a machine-readable format a related policy.

As a service, we also provide a repository for testing data, so that users can perform quality testing on data. This repository is the so called Sandbox, and it is mainly a clone of the long-term archiving system. In total, we operate three repositories for generic data with different purposes.

A further repository system, called Unidam, which was first created by two faculties of our university has recently been fully integrated to the central data management infrastructure. This gives repository users the possibility to get more features for their data, particularly in the field of digital humanities.

Based on the nationwide survey “Researchers and Their Data. Results of an Austrian Survey” (2015), which was directed at practically all Austrian researchers (36000 persons), we identified that nearly 25% of research projects use software developed during the process [7]. Looking at the software developed, and using what we know about well-established repositories for this purpose, it was possible to implement a Github Enterprise repository for such research and to integrate it into our ecosystem. This enables data to be linked to a software release, which could also be identified by a persistent identifier.

3.2. The services layer

In the services layer, re-use is the greatest value. For this layer, we reengineered the architecture of our Phaidra repository and integrated an API to enable other applications to dock on Phaidra. This change helped us to integrate an image server for presenting large images over the web and a streaming service for audio and video material, which is stored at the repository.

A further part of the service layer provides tools for managing data. We implemented a terminology server for controlled vocabularies, based on the SKOS [8] standard. This gives our users the possibility to choose controlled vocabularies on a wider range. A handle server creates persistent identifiers throughout the entire digital ecosystem, allowing consistent object referencing.

4. Outlooks

In the future, we plan to integrate a service for data management plans based on the DMP Online Tool [9] from DCC (Digital Curation Centre based at the University of Edinburgh) and the recommendation from RDA (Research Data Alliance) [10] regarding actionable data management plans. These are data management plans which are provided in both a human-readable and machine-readable way. Machine-readable output can further be used in tools for data stewardship. This allows more control over the data, its provenance and context, all relevant for re-use.

Currently, software development takes place in the research community, which poses a challenge regarding infrastructures and coordination. The question is, if software developed by research projects constitute a part of data preservation, and if so, how can software be maintained after a project ends? This challenge shows the need for technical consulting for researchers from the beginning of a project.

Important steps for the digital ecosystem are not only to provide a good working infrastructure, but to connect with the research community and maintain links to other infrastructure projects. Therefore, it is essential to our services to maintain the yet existing links to projects such as OpenAIRE[3] and Europeana[4] and OAPEN[5]. We are in regular contact with GÉANT[6] [link 5], and observe the European Open Science Cloud[7] [link 6] and large Austrian infrastructure projects, such as the Vienna Scientific Cluster.

[3] OpenAIRE: https://www.openaire.eu/

[4] Europeana: http://www.europeana.eu/portal/en

[5] OAPEN: http://www.oapen.org

[6] GÉANT: http://www.geant.org/

[7] European Open Science Cloud: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

Acknowledgements

Web references of named organizations

[link 1] e-infrastructures Austria: https://e-infrastructures.at/en
[link 2] OpenAIRE: https://www.openaire.eu/
[link 3] Europeana: http://www.europeana.eu/portal/en
[link 4] OAPEN: http://www.oapen.org
[link 5] GÉANT: http://www.geant.org/
[link 6] European Open Science Cloud: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

References

[1] Realising the European Open Science Cloud. First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud. Available from: https://ec.europa.eu/digital-single-market/en/news/first-report-high-level-expert-group-european-open-science-cloud : 12

[2] PARSE.Insight report from the FP7-2007-223758. At LIBER [homepage on the Internet. cited 2017 March 18]. Available from: http://libereurope.eu/wp-content/uploads/PARSE-Insight_D3-5_InterimInsightReport_final.pdf : 3

[3] Susan Reilly, Wouter Schallier, Sabine Schrimpf, Eefke Smit, Max Wilkinson. Report on Integration of Data and Publications. Opportunities for Data Exchange (ODE). FP7 Grant Agreement number 261530 : 19

[4] The Consultative Committee for Space Data System Practices (CCSDS). Reference Model for an Open Archival Information System (OAIS). Magenta Book, June 2012, CCSDS 650.0-M-2 Available under: https://public.ccsds.org/pubs/650x0m2.pdf : 2-2

[5] Mark D. Wilkinson, Michel Dumontier, Barend Mons. The FAIR Guiding Principles for scientific data management and stewardship. nature [homepage on the Internet. cited 2017 March 18]. Available from: http://www.nature.com/articles/sdata201618

[6] J. C. Burgelman. European Open Science Agenda. Presentation. Brussels, 15 January 2016. Available from: https://www.era-learn.eu/events/annual-joint-programming-2015-new-date-2016/topic-3-strategies-for-fostering-open-knowledge-and-open-access-in-research/01_2016OpenScienceAgendaERALEARNconference.pdf : 13

[7] e-infrstructures Austria. Researchers and Their Data. Results of an Austrian Survey. Report 2015. Available from: :24

[8] SKOS Simple Knowledge Organization System. W3C [homepage on the Internet. Cited 2017 March 19]. Available form: https://www.w3.org/2004/02/skos/

[9] DMPonline. [homepage on the Internet. Cited 2017 March 19]. Available form: http://www.dcc.ac.uk/dmponline

[10] Tomasz Miksa. Information integration through Actionable Data Management Plans. RDA 8th Plenary https://www.rd-alliance.org/system/files/documents/RDA_8thPlenary_Miksa.pdf: 2

List of Figures