On the Importance of a Research Data Archive

Benedict Wright[1] and **Oliver Brunner**[2] and **Bernhard Nebel**[1]
BrainLinks-BrainTools, University of Freiburg, Germany
[1]{bwright, nebel}@informatik.uni-freiburg.de [2]oliver.brunner@merkur.uni-freiburg.de

Abstract

As research becomes more and more data intensive, managing this data becomes a major challenge in any organization. At university level there is seldom a unified data management system in place. The general approach to storing data in such environments is to deploy network storage. Each member can store their data organized to their own likings in their dedicated location on the network. Additionally, users tend to store data in distributed manner such as on private devices, portable storage, or public and private repositories. Adding to this complexity, it is common for university departments to have high fluctuation of staff, resulting in major loss of information and data on an employee's departure. A common scenario then is that it is known that certain data has already been created via experiments or simulation. However, it can not be retrieved, resulting in a repetition of generation, which is costly and time-consuming. Additionally, as of recent years, publishers and funding agencies insist on storing, sharing, and reusing existing research data. We show how digital preservation can help group leaders and their employees cope with these issues, by introducing our own archival system OntoRAIS.

Introduction

A common problem in research at university level is the loss of data as a result of missing or wrong data management and leaving of essential staff members. Data is often stored on laptops, private computers, flash drives, or other storage devices which are not part of a central system. Even if data is stored on a network storage, it is usually stored in individual home drives and organized according to individual preferences. This leads to the situation in which it is known that certain data has already been created, but it is nearly impossible to retrieve. One solution to this issue is digital preservation. In this short paper we will first discuss the topic of digital preservation from the perspective of different stakeholders. The second section will then introduce an archiving software we are currently developing called OntoRAIS. Finally, we will discuss some general archiving issues.

What is Digital Preservation?

Digital preservation deals with all aspects of preserving digitally created or digitalized data. This ranges from preparing the data for long term storage to the issues of how to keep data accessible over a long period. As we are talking about digital preservation of research data, we consider the required archival time to be around 10 years. This is in contrast to other digital preservation areas such as cultural heritage, where data needs to be stored and accessible for an indefinite, potentially infinite duration.

Digital Preservation from a Researcher's Perspective

As researchers, we produce a large amount of data required for our work. This data consists of many types depending on the research field we are active in. In the case of Computer Science, this will usually be source code, experiment results, images, and videos. We have grown accustomed to using a central network storage for storing individual documents. For collaboration, multiple repositories and shared folders are used. This structuring of files usually results in a non-uniform storage of data in different locations. This works as long as the data is actively being worked with, and the creators of the data are still working at the work group. Problems arise when the authors have left, and one tries to access data created in the past. Generally there exists a rough notion of where the requested data lies, however, retrieving it from multiple repositories and storage locations on the internal network is tedious, and not always possible. Often the data will be stored on devices no longer functional. Additionally, many smaller data sets and scripts may be stored on the researcher's device, which may become unavailable once the researcher has left the work group. All these issues pose major obstacles to the activity of researchers and the group as a whole. New employees face the difficulty of getting to know the previous work going beyond what has been published, data and source code needs to be reconstructed, and time-consuming experiments need to be rerun. Therefore, a central archive containing all data in a retrievable manner is essential for any work group. Once such an archive has been established, it can be used to not only hold data for internal use, but also for sharing data with people from other organizations or to the public. This also helps in accommodating the requirements stated by funding agencies, publishers and good scientific practice of storing and sharing scientific data (German Research Foundation (DFG) 2013).

Digital Preservation from a Publisher's and Funder's Perspective

Publishers and funding agencies are becoming more and more aware of the necessity to encourage and enforce proper handling of research data, in the context of data reuse and sharing. We will now give a brief overview of a few major stakeholders, and their views on data archival and preservation.

The German Research Foundation (DFG) defines research data to be an essential foundation for scientific work (German Research Foundation (DFG) 2015). They encourage research groups, applying for funding through the DFG, to install means of archiving research data, by providing resources on this topic, as well as financial support for implementing such a system.

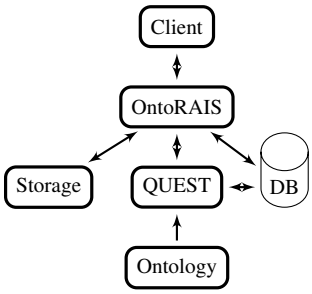
The European Commission has identified open access and data management as a key issue for funding of future projects within its Horizon 2020 funding program (European Commission 2016). The funding via the Horizon 2020 program requires publications to be provided in an open access fashion. Providing the underlying research data is still voluntarily under this program, however, financial support is provided if the research group decides to provide research data in an open access fashion.

As a major publisher for research work in multiple disciplines, NATURE.com requires authors, to make their data available to the editors and peer reviewers, and provide information on how to access the underlying data (NATURE 2016).

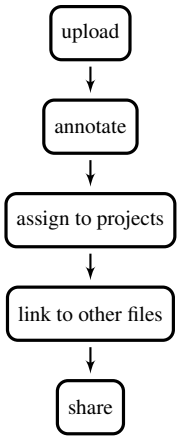
These initiatives and requirements indicate a strong growth of the awareness on digital preservation and its benefits for the future of academic research. Research groups will therefore have to invest into creating archives, or repositories, for their own data, not only for internal use, but also to accommodate open access requirements from publishers and governmental funding agencies.

OntoRAIS

In this section we describe our approach to solving above mentioned issues and challenges. For this we have developed an archiving system (OntoRAIS) consisting of an archive server and multiple interfaces for user interaction. The main goal of this development, was to create a system that approaches digital preservation from a user's perspective, instead of focusing on professional archivists. This approach seems beneficial to the scientific setting, where multiple smaller work groups conduct experiments and produce research data, of which no member is actually an expert in digital preservation. Our initial experiments showed that this corresponds to the real world settings in academia. In this section we will present our software, and give an overview of our preliminary experience with the system. OntoRAIS consists of two major parts, the server and multiple interfaces for interaction. The server is responsible for storing files together with their related metadata, and providing access for the user interfaces. The main user interface was developed as a web application, providing access to the data, and functionality to add new files. Additionally, a desktop client was developed to provide faster file transfer for larger documents. An overview of the system's architecture is given in Figure 1a with our back end which uses ontology based data access (OBDA), using the QUEST (Bagosi et al. 2014) library. OBDA combines the expressiveness of ontologies and the data storage capabilities of databases. This enables us to define semantics over the data stored in the database tables. With this we are able to capture not only metadata describing the actual files, but also in which context these files were created. This context consists of links to other entities of the archive. These links are modeled in the ontology in the form of OWL object properties such as *isAuthorOf* linking files to agents. Other such links are *ContributesTo*, *isDocumentedBy*, *isImplementedBy*, *isReferencedBy*, and *isRelatedTo*, which show relationships between archived files, or links to projects. The usage of an ontology for the data layer also enables us to reason about the objects in our archive. This helps us verify if archive objects are entered correctly and match the definition. Objects that do not validate in the ontology, can be marked for later review, adaptation, and completion. This provides the user with the option to archive a document without fully specifying all the metadata, and later review this document, and complete it according to the ontology definition.



(a) OntoRAIS Architecture with ontology based data access in back end



(b) OntoRAIS general work flow: Upload file, annotate with metadata, assign to related projects, link to other files, set access level to share

The basic work flow of using the archive is to first transfer the files to the archive using HTTP upload or git clone from the web application, or SFTP from the desktop client. The following steps are then performed in the web application. Each file needs to be annotated with metadata. This metadata is file type specific and contains very basic elements such as title, author, and additional annotations such as programming language for source code, or publisher for publications. After the annotation, one can optionally assign the archived file to a project for better organization. Finally, files can be linked to each other, and shared with other users in the archive. The following subsections describe these steps

in more detail.

Storing Files

The main purpose of an archiving system is to store files in a way that they are accessible to the target users over a long period. For this, two factors must be fulfilled: A reliable hardware infrastructure must be present, and a description of the data, which makes search and retrieval as easy as possible, needs to be provided. Providing reliable hardware for such a task is not in the scope of our work, as this is achieved by investing in proper storage and server hardware. Instead, we focus on providing means of adding and retrieving documents from the archive. To support this, each document stored in the archive, is annotated with a set of metadata, describing the underlying file. One type of metadata that can be associated with all files is the *financier* entry. This supports researchers in gathering all related documents for writing project reports, as the archive can be searched by this identification. Additionally, to the metadata describing file types, document names, authors, keywords, and other data type related information, documents can also be linked to one another using descriptive links such as *is documented by*, *is related to*, and many more. With this type of linking a network of documents is created, providing easy access to related data of any given file. This is especially important when a set of files belong together, documenting, for example a whole experiment with raw data, scripts, questionnaires, and results.

Our system provides two possibilities to add files to the archive. The first being a simple upload via the web interface, followed by entering metadata, and linking to other existing documents as shown in Figure 2. Alternatively, for larger files which can not be reasonably uploaded using a browser, a desktop client was developed. This client provides a simple interface for selecting a file type, setting the title, and selecting the files. The files are then transferred to the server using SFTP, a more sophisticated transfer protocol than HTTP. Finally, metadata can be added at a later time using the web interface. For convenience, we also added this tool to the context menu of the windows file manager for simple single click file archival. It is noteworthy that objects that are added to the archive can not be deleted or edited at a later stage, as this would defeat the purpose of an archive. Therefore, great care needs to be taken when entering the metadata. For future work, one might want to add a final review step, providing edit functionality until the archived object has been marked as finalized. The final archived object can then be retrieved and viewed in the web interface as depicted in Figure 3. This interface supports basic browsing functionality, providing easy access to all the documents of the user, as well as to documents with which he is affiliated with, either via projects or authorship. Additionally, a basic search functionality was implemented, enabling users to search by author, title, keywords, or financier. For future work, we are planning on implementing a plain text search of the whole archive.

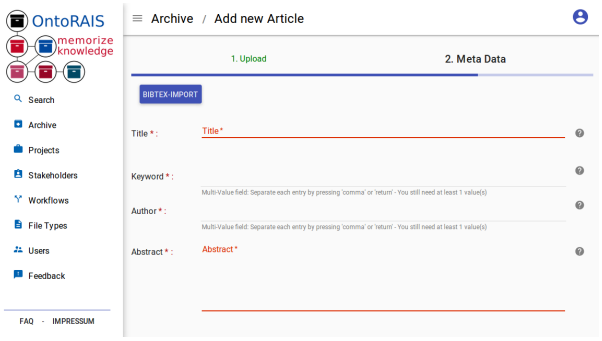


Figure 2: File Ingest

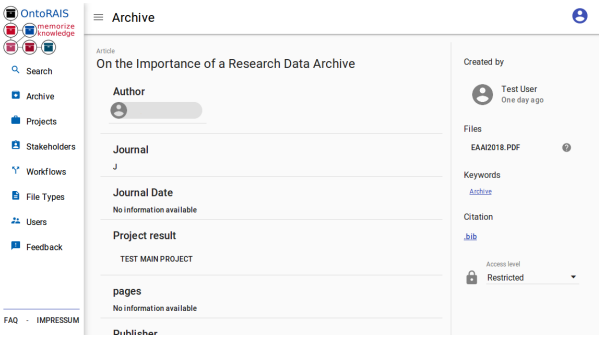


Figure 3: Object View

Organizing Files

When an archive grows it becomes important to be able to organize documents in a more sophisticated manner than simply archiving them and linking related documents to each other. Defining projects in the archive seems to be a natural way of organizing documents, as they can reflect the actual research projects carried out by the work group. By adding the possibility of defining sub projects, an even finer granulated organization can be realized. This can be used to create sub projects for each individual experiment, adding structure to a larger project. To reflect real world project plans, we also added the support of milestones, and assigning archived files to them. A screenshot of the project overview in the web interface is shown in Figure 4.

Sharing Files

One requirement from funding agencies and publishers is to share and publish data. Sharing data within the own work group or other organizations can be realized by providing access to a project under which the data was archived. Currently, this requires the accessing party to have a valid login to the system. However, we are working on a feature to publish data for public access. To manage who can view which documents, access management based on user permissions has been implemented. Currently, there are three levels of access that can be set on a per file basis. *Restricted*, providing access to only the authors and special members of related projects of the document; *Project*, providing access to all project members; and *Internal*, which grants access to

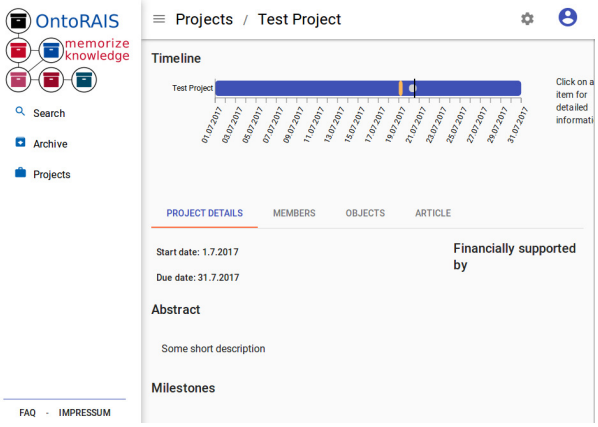


Figure 4: Project Overview

the document to anybody with a valid user account. In the future, a fourth access level will be added, providing access to the file without a login, and thus for public viewing.

Preliminary Experience

Currently, we are testing our system in cooperation with 7 work groups and institutes working in the research areas of epilepsy, radiology, engineering, robotics and computer science. All together there are currently 21 test participants consisting of professors, researchers, and students. During multiple presentations and discussions many additional features were identified to comply with the unique requirements of each individual discipline. Some of these features regarded the supported file types, others required a more detailed structuring of archived documents, and some regarded input methods and UI optimizations. One mayor positive feedback from all participants was about the user interface, which was conceived as very intuitive and streamline. The main negative remark was on archiving in general, as it adds workload to the researchers, for which no immediate gain is perceptible. Group leaders and professors did not regard this added work as negative, as they had the long term goal of the research group in mind, in contrast to students and PhDs whose working scope are often of a more short term nature. This led to the conclusion, that means of increasing the acceptance of an archive must be implemented. Initial ideas on achieving this were to implement a top down approach, where the supervisor requires the employees to archive their work before leaving the department. Alternatively a more user-friendly approach can be taken, by implementing features from games such as achievements and rewards for participating in the archiving process. These ideas however are for future consideration.

Related Work

In this section we present two of the more prominent digital preservation systems and discuss how our work compares to these existing systems. But first we give a short introduction to the Open Archival Information System *OAIS* standard (The Consultative Committee for Space Data Systems 2012) as it is implemented in nearly all modern digital preservation systems. *OAIS* defines six responsibilities a preservation system must fulfill:

- **Ingest**: Defines submission information package (SIP), which describes the physical files to be archived, as well as some initial checks and conversions that need to be executed before the files can actually be archived as an archival information package(AIP).
- **Storage**: States how the AIPs should be stored on the physical medium as to provide reliable data storage.
- **Management**: Describes how the data can be accessed, identified, and searched.
- **Access**: Provides security measures to protect data from unauthorized access.
- **Preservation Planning**: Describes how data needs to be transformed to stay accessible over time. This is an essential part of long term preservation.
- **Administration**: Deals with the overall administration such as configuration and access management.

RODA(RODA 2017) is an open source digital preservation solution. It follows the *OAIS* and other standards such as *Dublin Core* (Wolf et al. 1998) and provides functionality for storing, transforming, and retrieving data. It supports multiple ways of ingesting new data such as file upload via a web interface, a client application for offline archiving, and batch import using SIPs. Additionally, it provides a client API which can be used to implement own interfaces for ingesting new files. One shortcoming in our setting is that *RODA* requires the user to have a firm understanding of digital preservation and its processes. This could be overcome by implementing a more user-friendly interface for non experts using the client API.

DSpace(DSpace 2017) was originally developed by Hewlett-Packard (HP) and MIT Libraries and is now being maintained as an open source project by Duraspace. It complies with multiple standards such as *OAIS* and *Dublin Core*. The archive is structured in communities representing departments. Each community can contain multiple collections, grouping together related objects. Users add documents to the archive by uploading files and adding metadata to them. This upload is then reviewed by an archivist before it is accepted to the collection. Standard browsing and search functionality is provided by a web interface.

For an in-depth analysis of existing digital preservation systems we would like to refer to the recent survey carried out by Rosa, Craveiro and Domingues (Rosa, Craveiro, and Domingues 2017).

All the systems we analyzed come from a preservation background, therefore not integrating well into the everyday research work performed at a university level. During development of *OntoRAIS* we focused on the user's perspective making the interface as easy as possible to use for non-preservation specialists. Also, we reflected project-based research in our archive structure. Another feature, missing in classical archive systems, is the possibility to link documents to each other, giving more information about the context of a file.

Archiving Guidelines

Before one can start archiving research data, it is necessary to first answer the following questions: What data is required to be archived and what data can be left out of the archive? Who is responsible for archiving data and maintaining the archive? What are the legal settings under which the data is being created? In this section we provide some preliminary, brief answers to the above questions, in the scope of our environment. However, this discussion can also serve as a starting point for discussions about one's own archiving system, as each research group or institution will have its own unique settings and requirements.

Scope of the Archive

The first question to be answered is about the scope of the archive. Some data will be identified as crucial for future work, while others will be of less importance, or will be unpractical to archive due to it's size. In the case of files being too large, one can try to think of ways of reducing the size, such as archiving the means of how the data was generated, or filtering the data, reducing it to the subset actually used in an experiment. This, however, will depend on the actual data and the environment in which it was created and used. In the scope in which we developed and deployed our archiving system, we identified the following categories of data which we want to archive:

- Scientific publications: although most publications are stored in multiple locations, our goal is to create a single point of access for all related data. Storing the publications is also required for linking related documents to one another.
- Any document required to reproduce published results. This consist of raw data collected during experiments, images, protocols, scripts and source code.
- Project related documents such as reports and proposals.
- Negative results: sometimes it can also be beneficial to store information about negative results, as they are usually not documented in any publication, but can guide future researchers towards a more successful outcome of their research.

Responsibility

The second question that needs to be answered is who is going to be responsible for what: Who is responsible to archive what data? Who is responsible for maintaining the archive? Who is responsible for financing the hard and software of the archive? There are basically two approaches to defining who is responsible for archiving the data. First, the scientist responsible for generating the data may be required to archive all files. Alternatively, a dedicated staff member can be assigned. In the second case, this person will need to gather all required information, and make sure that all data is handed over. Archiving the data can be done once a work package is completed. This can be once an experiment has been conducted and evaluated, after a publication, or after some milestone within a project has been reached. As an archive should store and provide access to data over a long period, it is of utmost importance that a plan for maintaining the archive is created. For this, it is recommended to designate a single responsible person, who will service and extend the hardware, configure the software in compliance to changing requirements.

Data Requirements

Finally, when it comes to the actual data, there are a few issues that need to be considered before it is archived. The first being the legal issues concerning the data. Things to keep in mind here are ownership: is the data owned by the scientist/work group? Was it created in partnership with another work group or even an industry partner? Does it contain confidential or sensitive data? These questions must be answered, and depending on the environment of the archive and its capabilities, the data may or may not be suitable for archiving. A second requirement towards the data is its size and format. Some data might be too large to be considered for archival; in this case the archivist should consider compressing or filtering the data, to reduce it's size. Sometimes data was generated by a simulation, in that case it may be better to archive the generating software or source code. Also, to take into account is the file format, as some may become obsolete in the future, rendering the information in the file inaccessible. Therefore, one should preferably use file formats which are industry standards across software and platforms. For example, it is preferable to store text as PDF or plain text over Microsoft or OpenOffice.org word files. Preservation planning deals with this issue of file type obsolescence, by defining what original file types need to be converted to what new file type, as to ensure continued file access.

Conclusion

As data is ever growing, the need for storing data in an accessible way for future use is constantly increasing. Additionally, as experiments become more and more data intensive, reuse of existing data can significantly increase productivity and lower costs of conducting research. One solution to this is to deploy an archiving system, which stores and organizes data and provides unified access to the files. We have given a short introduction to our own archiving software OntoRAIS and discussed some issues concerning archiving in general. We showed how our system can be used to archive files together with its context, by annotating it with metadata and linking them to one another. Moreover, introduced the concept of *projects* to help organize large archives which reflects real world research structures. Finally, this section shows how the files can be shared with other researchers. The second topic of this paper is the design of archiving guidelines. We discussed some preliminary questions that need to be answered before any archival initiative can be started. We also pointed out some issues that need to be taken into consideration during the usage of an archive.

Acknowledgements. This work was partly supported by BrainLinks-BrainTools, Cluster of Excellence funded by the

German Research Foundation (DFG, grant number EXC 1086).

References

Bagosi, T.; Calvanese, D.; Hardi, J.; Komla-Ebri, S.; Lanti, D.; Rezk, M.; Rodríguez-Muro, M.; Slusnys, M.; and Xiao, G. 2014. *The Ontop Framework for Ontology Based Data Access*. Berlin, Heidelberg: Springer Berlin Heidelberg. 67–77.

DSpace. 2017. DSpace. http://dspace.org. Accessed: 2017-09-11.

European Commission. 2016. Horizon 2020 online manual. http://ec.europa.eu/research/participants/docs/-h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm.

German Research Foundation (DFG). 2013. Proposals for Safeguarding Good Scientific Practice.

German Research Foundation (DFG). 2015. Dfg guidelines on the handling of research data. http://www.dfg.de/-download/pdf/foerderung/antragstellung/forschungsdaten/-guidelines_research_data.pdf.

NATURE. 2016. Nature announcement: Where are the data? http://www.nature.com/authors/policies/-availability.html?foxtrotcallback=true#data.

RODA. 2017. RODA. https://www.roda-community.org. Accessed: 2017-09-11.

Rosa, C. A.; Craveiro, O.; and Domingues, P. 2017. Open source software for digital preservation repositories: a survey. *CoRR* abs/1707.06336.

The Consultative Committee for Space Data Systems. 2012. Reference Model for an Open Archival Information System (OAIS).

Wolf, M.; Kunze, J. A.; Lagoze, C.; and Weibel, D. S. 1998. Dublin Core Metadata for Resource Discovery. RFC 2413.