Lessons Learned: On the Challenges of Migrating a Research Data Repository from a Research Institution to a University Library

Thorsten Trippel and Claus Zinn
University of Tübingen
Wilhelmstrasse 19, 72074 Tübingen, Germany
{thorsten.trippel, claus.zinn}@sfs.uni-tuebingen.de

Abstract

The transfer of research data management from one institution to another infrastructural partner is all but trivial, but can be required, for instance, when an institution faces reorganisation or closure. In a case study, we describe the migration of all research data, identify the challenges we encountered, and discuss how we addressed them. It shows that the moving of research data management to another institution is a feasible, but potentially costly enterprise. Being able to demonstrate the feasibility of research data migration supports the stance of data archives that users can expect high levels of trust and reliability when it comes to data safety and sustainability.

Keywords: Research Data Management, Data Repositories, Data Migration

1. Introduction

Good scientific practice requires that research data created and studied by scientists is archived. The sustainable archiving of research data is a complex manner as it includes the entire life cycle of data. In the different phases of this life cycle, many human factors are involved, often placed in different organizational structures, and making use of many technological frameworks.

The sustainable management of research data is a noble aim, but there is no single golden path to sustainability. Also, the path might suddenly encounter a road block, when for instance, an existing archival infrastructure faces a discontinuation because a research institution faces closure or a new research orientation. Here, the sustainability of research data management depends on another institution being able and willing to take over the data.

Though the scholars are partly responsible for archiving, they can hardly be responsible for running the archive. Research infrastructures and networks of institutions claim that they fulfill this responsibility, using certified technical infrastructure and processes. The cooperation between institutions in such infrastructures and networks strengthens the overall reliability of each partner, but the expression of intent can be severely tested if one partner discontinues its service.

In this paper, we report on the following use case: research data has been collected, evaluated, catalogued, and made accessible by a research institution in an exemplary manner. It is assumed that this data centre is discontinued, but that the research data should remain available. All research data therefore needs to be migrated, in the given case from a linguistics department to a discipline independent data facility operated by an institutional infrastructure, here a university library and computing centre, which takes care of all data and guarantees its access for the foreseeable future.

For this use case, we have devised a migration concept that has uncovered a number of challenges that need to be addressed to make such as hand-over of research data management a success.

2. Background

The work reported in this paper stems from the NaLiDa project, which was divided into two main phases. The first funding phase aimed at the construction of an infrastructure for the long-term archival of linguistic resources with technology and workflows that are manageable and sustainable. The infrastructure was to be built within the research institution that creates all data, the department of linguistics at the University of Tübingen. In the second phase, the NaLiDa project took on board the two infrastructural units of the University of Tübingen, the university library and the computing centre. The aim was to explore how to best transfer the management of research data to these units for the long-time archiving of linguistic resources. Also the university library wanted to learn about the processes required to ingest all resources' metadata into their catalogues. With library catalogues connected with the research data repositories of the computing centre, users would profit from easy-to-use access points.

It turned out that the transfer of research data management is no easy matter, and that many technological and organizational hurdles exist and need to be dealt with. In the remainder of this section, we describe the management of research data at both the research institution and at the university library. The aim is to identify the commonalities and differences of Research Data Management (RDM).

2.1. RDM at the discipline specific-data centre

RDM has a technical and an organizational perspective.

2.1.1. Technical Backbone

The technical backbone at the time of the migration process comprised the following four key components:

- a Fedora Commons 3 repository (which in the meantime has been ported to Fedora Commons 4), see [U6].
- ProAI: a repository-neutral, Java web application supporting the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), see [U7].

- ProFormA: a form-based editor for metadata management (Dima et al., 2012b) (in the meantime replaced by Comedi (Lyse et al., 2015)), and
- ERDO: a web portal for research data ingestion and maintenance (Dima et al., 2012a).

While the first two items are open-source applications, the latter two are in-house developments. The ProFormA editor is targeted at users to easily instantiate CMDI-based metadata schemas; and the ERDO web portal is used to support the data ingestion workflow, see below.

2.1.2. Workflows

The discipline oriented data centre closely cooperates with the data providers, usually the researchers who created or gathered all data. All parties involved are committed to follow FAIR, a set of guiding principles to make data Findable, Accessible, Interoperable, and Re-usable (Wilkinson, 2016), see also [U11].

Both parties initialize the archiving process in a cooperative manner. The data providers decide on the granularity of the data to be archived, but get help from the data centre staff. Such decisions can be helped by consulting, for instance, the criteria of the ISO 24619 standard on the assignment of persistent identifiers to language resources (see (ISO 24619, 2011), section 6).

The next step is to collect and upload all relevant individual files to the repository system. Depending on the type of the research data (*e.g.*, lexical resources, experimental studies, or text corpora), an appropriate CMDI-based metadata schema is selected (ISO 24622-1, 2015). An initial provision of the metadata is given by the research data creators, who presumably know their data best. However, as the data providers are not necessarily archiving experts, they consult with the data centre's archivist to answer any questions. Usually, metadata provision is an iterative process between both parties, where research data providers add missing pieces of information, and where archiving experts may curate the data.

The discipline oriented data centre is committed to open access. In practise, however, there are often cases where language related research data is subjected to restrictive data usage licenses. This is the case, for instance, when research data makes use of third party data, which is turn is published under a restrictive license, or when research data involves potentially privacy infringing data collections. In some cases, researchers would like to choose an open license for their data (*e.g.*, `https://creativecommons.org/licenses/by-nc-nd/4.0/`) but want to be consulted before it is given to an interested party. In any case, the data providers – in close consultation with the archivists – assign appropriate license and access rights to the data, varying from "open to the general public" to "protected, individual permission required per dataset".

Whenever the data provider is *not* affiliated with the data centre's institution, the rights and duties between depositor and depositee are laid down in a depositing agreement. The agreement specifies, for instance, that the depositor (i) is the owner of the intellectual property rights of the data, (ii) warrants that the dataset (and its metadata) does not contain false or misleading information, (iii) assures that the dataset does not violate or infringe any copyright, trademark, patent or intellectual property rights of third parties, and so on. The agreement grants the depositee, for instance, the rights to distribute the dataset in electronic form, to make available its metadata records through its catalogues, or to assign digital object identifiers that link metadata records with the data they describe. A good example for a deposit agreement is given in [U12], but clearly, such (legally binding) documents must be drawn up on a case by case basis.

Once the data providers have finalized the provision of metadata and access restrictions, the archivists take over, adding elements unknown to the data providers, including technical information on the submitted files (*e.g.*, checksums, file sizes, storage locations) and references to authority files if available, see (Zinn et al., 2016). Additionally the archivists start a quality assurance process for all files.

At the end of the quality assurance phase, the archival objects receive a persistent identifier according to ISO 24619. In our study the Handle system is being used [U5]. With the persistent identifiers, the archival objects are finally archived in the repository system. This process involves the publication of the metadata via the OAI-PMH protocol, see [U8]. The metadata now also includes access information to the data such as location, contact information, and license/access rights.

To honor the access rights attached to research data, the archival system implements a system for authorization, which is based on the built-in access control system by Fedora Commons. XACML authorization rules define users with name and password, and a role allowing or denying specific operations for archived objects:

```
<user name="guest" password="xxx">
   <attribute name="fedoraRole">
      <value>user</value>
   </attribute>
</user>
```

2.2. RDM at the University Library and Computing Centre

2.2.1. Technical Backbone

The technical infrastructure at the university library and computing centre makes use of the following software:

- the Fedora Commons 4 repository,
- the software Apache Solr/Lucene for indexing [U9],
- *Docuteam Packer* for the creation of packages of archival files [U10],
- ingest software for the archival and validation of digital objects, and
- portal software for research data access and rights management.

The latter two items are in-house developments.

```
<Policy>
  <Subjects>
     <AttributeValue
         DataType="http://www.w3.org/2001/XMLSchema#string">
        A04_admin
     </AttributeValue>
  </Subjects>
                                          ==>
  <Resource>
     <AttributeValue
       DataType="http://www.w3.org/2001/XMLSchema#string">
       FID:183
     </AttributeValue>
  </Resource>
<Policy>
```

Handle-PID	ACL Expression
10900.1/FID:183	+u:A04_admin

Figure 1: From NaLiDa-ACLs to library and computing centre's ACLs.

2.2.2. Workflows

At the university library, the archival process starts with a pre-ingest of the research data using the Docuteam Packer software. During the pre-ingest, the researcher gathers and structures her data into a machine-readable package, and describes it with metadata. In this process, researchers are supported by the staff of the library and computing centre. As a result, a Submission Information Package (SIP) is created that contains all research data and its metadata in the EAD/METS format (Encoded Archival Description / Metadata Encoding Transmission Standard), see [U3].

In-house software is then used to read the resulting package and validate its content for correctness and completeness. Upon successful validation, the package is being ingested into the library's digital archive, which is based on the Fedora Commons 4 repository system. As part of the ingestion process, each resource is assigned a unique persistent identifier (PID) of the Handle system. Also, all metadata is being ingested into an Apache–Solr Server.

All access to archived digital objects is performed via a purpose-built portal software. The portal gives a web-based access, and for this it makes use of a database that holds information about access rights to digital objects, and information about users and their authorization records. Authorization is defined via *Access Control Lists (ACL)*. The database associates with each PID an ACL.

When a user logs into the portal system, his user data is retrieved from the database. Users will only be able to access a resource when their credentials feature in the resource' ACL. Here, the authorization system distinguishes three access categories: roles, users, and groups. An ACL can contain any number of instructions along these categories that are either tagged as "+ (grant)" or "- (revoke)". For authentication, the central authentication server of the University of Tübingen is used. The user ids from the authentication server correspond to the user ids in the portal's database.

2.3. Commonalities and Differences

Our two data centres share common features, but there are also differences. Both rely on the same repository backend and hence share a large common technological ground. Also both centres use handle-based persistent identifiers, though with different prefixes. Major differences exist in the workflow, which is more generic at the library and computing centre; here, the research institution, naturally, offers more discipline-specific support. This is also exemplified by the different metadata schemas; here, the discipline-specific institution makes available CMDI profiles for different types of resources, while the library makes use of EAD (Encoded Archival Description), which does not discriminate against resource types. Both data centres also manage access rights differently. Here, the library-based archive has a stricter regime in place, which is also embedded in the university's authentication system.

3. Migration Concept

We outline migration issues along three dimensions.

3.1. Authentication and Authorization

The authentication and authorization procedures for accessing research data differ considerably. At the web portal of the library and computing centre, authentication is embedded in the university's central LDAP server whereas the discipline-specific NaLiDa repository uses a proprietary authentication procedure that is captured by locally maintained XML-based documents.

To address this issue, the library and computing centre needs to complement the usage of the central LDAP server with a local server that will also be consulted for user authentication. The local LDAP server will register all NaLiDa users that do not have a valid university-based id (that is, their id is not part of the central LDAP server).

With regard to authorization, both approaches use a role-based access management based on Access-Control-Lists (ACLs). However, there are differences in the use of ACLs, and where they are stored. In the library and computing centre, no user of the web portal is granted write access to digital objects. Once a digital object is ingested, it cannot be changed. In the NaLiDa repository, the archiving workflow allows ERDO users to modify the digital objects prior to their publication by the archivists. It is clear that digital objects, once transferred to the repository of the library and computing centre's repository, cannot be changed thereafter. Any NaLiDa-based access rights that grant the writing of digital objects will be revoked.

Technically, the NaLiDa repository used the functionality for access management as provided by the Fedora Commons 3 software (xacml-2.0-policy-schema). The library and computing centre repository uses a different approach that is decoupled from the repository software, the aforementioned database-driven approach. This approach helps migrating the NaLiDa-based access rights as any ACL can be mirrored to a corresponding database entry. Here, all *XACML Subjects* are mapped to users of the library and computing centre; the roles for administrator and user also have their correspondence in the repository of the library and computing centre, and all XACML READ Actions can be transformed into equivalent *grant* and *revoke* statements. All WRITE statements will not be migrated. Fig. 1 illustrates the migration of access rights from one repository system to the other.

3.2. Metadata Harmonization

There is a profound difference in the metadata used to describe research data. While the NaLiDa team uses the CMDI-Framework, which follows the ISO 24622-1 standard, the library and computing centre uses the Encoded Archival Description (EAD) scheme. The transfer of research data must hence include a transformation (crosswalk) from one metadata scheme to another.

The crosswalk is based upon an existing conversion from CMDI-based metadata profiles to Dublin Core [U1] and MARC 21 [U2], see (Zinn et al., 2016). The conversion is hand-tailored to all profiles used in the NaLiDa repository (*e.g.*, for the description of corpora and tools), and aims at limiting the loss of information for these profiles.

The metadata harmonization makes use of those conversions by first converting CMDI-based metadata to MARC 21. Then, a crosswalk from MARC 21 to EAD is being performed. This crosswalk is well documented and used in the library world [U4]. The conversion of discipline-specific metadata to generic bibliographic metadata profits from the use of authority records (Trippel and Zinn, 2016). The ingestion process at the library and computing centre will need to make use of the conversion service.

3.3. Persistent Identifier Management

Persistent identifiers can cause problems. While both repositories make use of Handle-based PIDs, they use different identifier prefixes and different local handle servers to resolve them. The NaLiDa-based PIDs are resolved using a local Handle-Server at the GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen), and the library and computing centre uses its own handle server with their own prefix (10900.1). This server will need to take over the resolving process for handles. Here, the new PIDs automatically created during the ingesting workflow at the library and computing centre's repository will need to be mapped to the existing PIDs that stem from the NaLiDa repository. The PID stemming from the GWDG will then point to the new library and computing centre based PID, which in turn will point to the corresponding research data in the library and computing centre. Note that such PID assignment involves a third party (at the GWDG) so that the PID mapping can only be partially mechanised.

4. Discussion and Conclusion

The authors are not aware of reported similar efforts in linguistics or related research areas. The migration of research data from one data repository to another is bound to take place from time to time in many institutions, but seems to get rarely reported and published. Also note that there is large degree of freedom that governs such enterprises. Readers who managed to migrate from, say the Fedora 3 repository system to its Fedora 4 successor will be aware of the many technical subtleties and intricacies involved, even if such migration is taking place within a single institution.[1] As a result, many design and migration decisions may well differ across institutions, which makes it hard to generalize. In this case study, we outlined the migration of research data from one data repository to another one. We assumed that all data is being migrated, and ignored a potential step to re-evaluate all data with regard to data obsolescence. The migration study profited from a common technological base as both archives used the Fedora Commons repository system. Still, there were issues that we needed to address.

Access restrictions, once imposed to research data, can become a significant hurdle for data migration. Here, we advocate a strong commitment to Open Data. Restricted access to data should be avoided, potentially at the cost of moratoria where restrictions must be lifted after a limited period of time. Ideally, legal agreements in favour of open access should be drafted when research data is deposited for the first time as it might be harder to amend any agreements at the time of the data migration.

The authentication and authorization infrastructure (AAI) that we described in this paper was limited to the level of the discipline-specific data centre at the institution level, or the wider university-wide level for RDM at the university library and computing centre. Ideally, digital repositories should support users from the outside, too. It should be possible, for instance, to make available resources with a "CC BY-NC-ND 4.0" license (permitting the non-commercial use of research data) to other researchers world-wide. Here, the migration of data to the more generic infrastructure will likely increase data accessibility as the university library and computing centre is in a better position to support an international authentication infrastructure such as `https://www.eduroam.org`.

The conversion of metadata formats can also be a challenging undertaking. Here, research institutions might have very expressive means to describe their research data, whereas library institutions often strictly adhere to bibliographic metadata standards such as MARC 21 or EAD. In our case, the information loss is significant as the conversion process went from CMDI to MARC 21, and from MARC 21 to EAD. Also, only the EAD description enters the library catalogue. To a large part, the value of a repository is rooted in the metadata that is used to describe its content. If the migration of research data implies a degradation of its corresponding metadata quality, then this is very unfortunate, in particular, when so much effort has been

[1] See `https://wiki.duraspace.org/display/FF/Training+-+Migrating+from+Fedora+3+to+Fedora+4`.

undertaken to describe research data is the most descriptive way possible. Here, we advocate to keep the rich original metadata attached to the digital object so that a maximum amount of information about the research data is preserved.

The migration of data from one data centre to another is a non-trivial undertaking. Migration costs can be significantly lowered when both data centres make use of good practices and standards. A common technological base eases the migration process considerably, but parties should be aware of issues such as access rights, metadata conversion, and the new resolving of persistent identifiers. There are other issues that might be taken into account, for instance, when the new data centre requires all research data to be bundled and ingested at a different level of granularity.

The migration of research data from one data centre to another needs to be carefully planned, and sufficient time and personnel should be allotted to ensure a smooth transition. From our description in Sect. 3, it should be clear that only parts of the migration process can be fully automated so that migration costs increase almost linearly with the number of digital objects to be migrated.

In an ideal world, the receiving end has a fully functional technical and organizational setup in place, but in reality many universities have only started to establish eScience centres that must cater for the needs of many different disciplines. When data migration must happen during the start-up of such an eScience centre, extra time must be allocated.

Also, be prepared that many smaller issues may materialize well after the actual migration. Here, data depositors might feel the most important service deterioration. When those researchers handed over their data to the discipline-specific data centre, they were given their data to colleagues they know, and despite well established workflows, their were informal communication channels were it was easily possible to amend data, metadata, or access rights. When research data is now managed at the infrastructural institution of the university, those informal settings are replaced by official routes. Here, the depositee is not both a linguist and archivist (who caters for the few), but just an archivist (who caters for the many). With greater distance to the respective discipline, it is likely that discipline-specific metadata schemes (such as CMDI profiles for speech corpora) will be superseded by generic bibliographic metadata. This makes it harder for others to find and evaluate research data for their studies. In consideration of this factor, it is advisable to choose the data-receiving archive with care. If possible, choose an archive that values and enforces rich metadata and which guarantees that all data and metadata are indexed in a widely known and searchable resource.

5. Acknowledgments

This work has been supported by the German Research Foundation (DFG reference no. 88614379), and the SFB 833 data management project INF (DFG reference no. 75650358). The data centre cooperates closely with the CLARIN-D centre in Tübingen which is funded by the German Federal Ministry of Education and Research (BMBF). We would like to thank the anonymous reviewers for their valuable feedback.

6. Web Resources

[U1] The Dublin Core Metadata Initiative, see www.dublincore.org.

[U2] The MARC 21 standard, see www.loc.gov/marc/bibliographic.

[U3] The EAD standard, see https://www.loc.gov/ead/.

[U4] The MARC to EAD crosswalk, see http://www.loc.gov/ead/ag/agappb.html#sec4.

[U5] The Handle system, see https://www.handle.net.

[U6] The Fedora repository platform, see http://fedorarepository.org.

[U7] ProAI, see http://proai.sourceforge.net.

[U8] The OAI-PMH protocol, see https://www.openarchives.org/pmh.

[U9] Apache Lucene and Solr, see http://lucene.apache.org/solr.

[U10] Docuteam packer, see https://www.docuteam.ch/en/products/it-for-archives/software.

[U11] The FAIR principles, see https://www.force11.org/group/fairgroup/fairprinciples.

[U12] Example of a deposit agreement (University of Reading, UK), see http://researchdata.reading.ac.uk/deposit_agreement.html

7. Bibliographical References

Dima, E., Henrich, V., Hinrichs, E., Hinrichs, M., Hoppermann, C., Trippel, T., Zastrow, T., and Zinn, C. (2012a). A Repository for the Sustainable Management of Research Data. In N. Calzolari, et al., editors, *Proceedings of the 8th. International Conference on Language Resources and Evaluation (LREC'12)*. ELRA.

Dima, E., Hoppermann, C., Hinrichs, E., Trippel, T., and Zinn, C. (2012b). A Metadata Editor to Support the Description of Linguistic Resources. In N. Calzolari, et al., editors, *Proceedings of the 8th. International Conference on Language Resources and Evaluation (LREC'12)*. ELRA.

ISO 24619. (2011). Language resource management - Persistent identification and sustainable access (PISA). International Standard.

ISO 24622-1. (2015). Language resource management – Component MetaData Infrastructure – Part 1: The Component Metadata Model. International Standard.

Lyse, G. I., Meurer, P., and Smedt, K. D. (2015). Comedi: A component metadata editor. *Selected Papers from the CLARIN 2014 Conference, Linköping University Electronic Press*, 116(8):82–98.

Trippel, T. and Zinn, C. (2016). Enhancing the quality of metadata by using authority control. In *5th. Workshop on Linked Data in Linguistic (LDL-2016) at LREC'16*.

Wilkinson, M. D. *et al.*. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(160018). http://dx.doi.org/10.1038/sdata.2016.18.

Zinn, C., Trippel, T., Kaminski, S., and Dima, E. (2016). Crosswalking from CMDI to Dublin Core and MARC 21. In N. Calzolari, et al., editors, *Proceedings of the 10th. International Conference on Language Resources and Evaluation (LREC'16)*. ELRA.