

POSTPRINT

Thorsten Trippel¹
Claus Zinn¹

Lessons learned: on the challenges of migrating a research data repository from a research institution to a university library

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 reorganization 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

✉ Claus Zinn
claus.zinn@uni-tuebingen.de

Thorsten Trippel
thorsten.trippel@uni-tuebingen.de

¹ University of Tübingen, Wilhelmstrasse 19, 72074 Tübingen, Germany

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 et al. 2016), see also <https://www.force11.org/group/fairgroup/fairprinciples>.

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), see [Web resources](#)).

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 in 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., CC-BY 4.0, see <https://creativecommons.org/licenses/by/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 researchdata.reading.ac.uk/deposit_agreement.html, but clearly, such (legally binding) documents must be drawn up on a case by case basis.¹

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. At the discipline-specific data centre, the Handle system is being used <https://www.handle.net>. Equipped with 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 <https://www.openarchives.org/pmh>. 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>
```

¹ As a starting point, depositing agreements can draw upon templates that are prepared by infrastructure providers. The legal evaluation of an instantiated template, however, often depends on the very instantiations, that is, specific criteria that involve the character of the data, the legal status of depositor and depositee, third parties etc. To avoid any form of liability, templates are rarely shared across institutions. If templates are shared, then with an explicit disclaimer (“do not use it as is”) and the strong suggestion to seek for independent professional legal advice.

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 lucene.apache.org/solr,
- *Docuteam Packer* for the creation of packages of archival files <https://www.docuteam.ch/en/products/it-for-archives/software>,
- 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.

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, researchers gather and structure their data into a machine-readable package, and describe 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 <https://www.loc.gov/ead/>.

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 users log into the portal system, their 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 Encoded Archival Description (EAD), 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 main 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 need to 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

```

<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

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

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 and ISO 24622-2 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 conversion of discipline-specific metadata to generic bibliographic metadata profits from the use of *authority records* (Trippel and Zinn 2016). Such records are widely used in bibliographic catalogues, but so far they are rarely used in the CMDI world. An authority file record gives the name of a person or institution a standardized representation. Often, the entities' preferred name is complemented with alternative forms or spellings, and becomes associated with a unique resource identifier. Many libraries use authority files for identity management, e.g., the German National Library <https://www.dnb.de/EN/Standardisierung/GND/gnd.html> or the Library of Congress id.loc.gov/authorities/names.html. Often, the International Standard Name Identifier (ISNI, isni.org) is also used. The Virtual International Authority File viaf.org links together the national authority files of

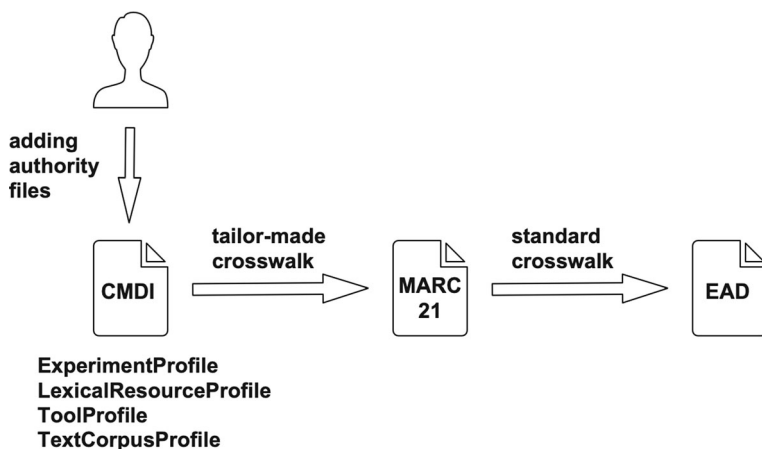


Fig. 2 Conversion of metadata from CMDI to EAD

many libraries to a single virtual authority file. Each VIAF record is associated with a unique resource identifier and aggregates the information of the original authority records of participating national libraries.

Before the conversion of CMDI metadata to bibliographic metadata, we have manually added authority file information to many persons and institutions mentioned in the description of the research data (e.g., as authors, creators, project partners). Enriching metadata with authority records helps bridging the different expectations between discipline-specific metadata and bibliographic metadata.

The crosswalk from CMDI to EAD is depicted in Fig. 2. It makes use of an existing conversion from CMDI-based metadata profiles to Dublin Core <https://www.dublincore.org> and MARC 21 [U2], see (Zinn et al. 2016). This conversion is hand-tailored to all profiles used in the NaLiDa repository (*ExperimentProfile*, *LexicalResourceProfile* etc.) to limit the loss of information for these profiles.

Once the CMDI-based metadata is converted to MARC 21, an existing, well-established and well-documented crosswalk from MARC 21 to EAD is being performed <https://www.loc.gov/ead/ag/agappb.html#sec4>.

3.3 Persistent identifier management

Persistent identifiers (PIDs) ensure permanent access to a digital object by providing access to it independently of its physical location (ISO 24619 2011). Consider Fig. 3, sketching the simple case. Here, the target repository (“new archive”) does not create new PIDs during the ingestion of migrated data, but simple “re-uses” those of the source repository.

For this, the new archive simply updates the PID-URL mapping so that the PID now points to the new resource location.²

² It is possible that the new archive may move the resources at a much later point in time to yet another location, so the capability to manipulate PID-URL mappings should be transferred from the giving archive to the receiving archive.

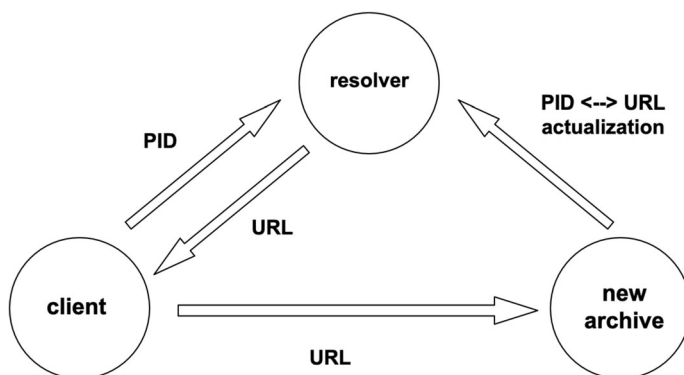


Fig. 3 Using PIDs to ensure that a resource remains accessible after it changed location

In the NaLiDa project, we faced a setting where the new archive *does* create new PIDs during their data ingestion workflow. Also, while both giving and receiving 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). 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. The nature of the PID migration is depicted in Fig. 4.

The mapping of persistent identifiers is a labour-intensive process. First, a list of all PIDs that are in use by the source repository needs to be created. For each source PID, a target PID needs to be created that is then used by the target repository to refer to the research data object being moved. Then, the source PID needs to be redirected to point to the resolvable target PID.

While the use of PIDs facilitate the migration of resources from one archive to another, persistent identifiers can cause problems, too. Here, the use of *part identifiers* in PIDs, often impacted by different policies with regard to the granularity of digital objects, may cause major migration pain.

A PID part identifier unambiguously identifies a part or segment within the containing resource the PID is referring to. Such part identifiers could identify fragments of a MPEG resource, part of a text in a textual resource etc. In the NaLiDa context, part identifiers are used to refer to the individual data streams of a digital object (the resource). When using PIDs with part identifiers, it is the responsibility of the archive to maintain and provide such fragments. This is because the Handle system does not support direct use of part identifiers itself; it

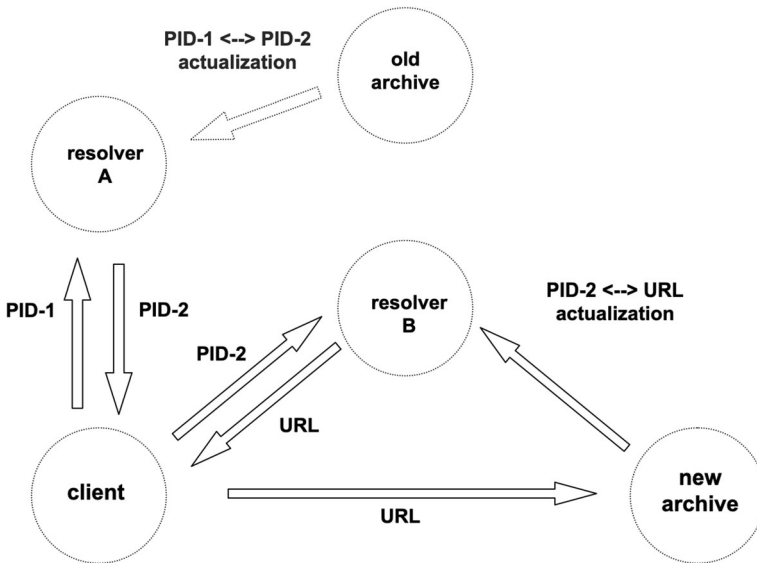


Fig. 4 PID mapping in the NaLiDa context

will only encode a part identifier as part of the URL-ified, resolvable handle. It is the task of the archive repository to manage such complex URLs (usually following the HTTP protocol). For part identifiers to survive the migration process, the target repository would need to mimic the way they are handled by the source repository.

3.4 Other issues

There are many other issues that need to be taken into account when migrating research data across repository systems.

3.4.1 Granularity of digital objects

Varying policies regarding the granularity of digital objects in repository systems are likely to create migration obstacles, especially when such policies are tied to the use or the rejection of using part identifiers in persistent identification systems. In an extreme case, consider a source repository where the individual prompts of an experiment were stored in individual files, and where each file was addressable via a PID suffixed with a part identifier. Consider further that the target repository's policy, however, is to containerize such prompts into a single zip file, which in turn should receive its own PID. Here, the target repository's policy is that each data stream of a digital object must get its own PID assigned, and that part identifiers are not supported.

Such conflicting policies with regard to granularity and the use of PID part identifiers create much hassle. The restructuring of digital objects require an actualization of their associated metadata, and the management of PID-URL

mappings can create almost insurmountable problems (see Sect. 3.3). In essence, the new archive would need to mirror the granularity of the old archive to keep migration costs within reasonable bounds. The granularity issue, often related with the use of part identifier, should be a pivotal criteria for the selection or exclusion of a target repository.

3.4.2 *Revision and curation*

The re-evaluation of research data and its metadata with regards to obsolescence is an integral part of data' life cycle. In an ideal world, such re-evaluations should not depend on a data migration step (or other external causes such as hardware and software changes), but taken at larger regular assessment intervals. In any case, the migration of research data to another repository should trigger the activity to revise and curate all data and their metadata with utmost attention. First, the source repository might contain old research data that has been archived for a long time, and where log files monitoring repository access show that it attracted little interest from the community. Also, there might be other reasons to decide that the archival of some research data shall end prior to the migration process (e.g., where data represent old test case entries).³

The source repository might also have “old” research data that is worth keeping alive, but whose digital objects contain data streams in outdated data formats. Often, the discipline-specific source repository has the expertise to convert such outdated formats to newer formats. While existing outdated formats should be kept around (existing processing workflows might rely on them), the digital object should be enriched with the modern format variants.

Also, digital objects might have data streams in proprietary formats. Here, the source repository might want to consider to convert them to open-source formats (e.g., from Excel table sheets to comma-separated files). Such conversions should be handled with care as they might yield a loss of information (e.g., when the Excel tables contain formulae to compute table cell values).

Revision and curation also affects the metadata. With the source repository's established contacts to data depositors, it is clear that all metadata must be updated before it is transferred to a new archive. This includes updated personal details (e.g., email addresses and phone numbers) for all data contributors. Also, data contributors should be encouraged to relax restrictions for access and to use open licenses whenever possible.

3.4.3 *Transparency*

Research data depositors should be informed at an early stage of the migration effort. All other users should be informed in due time that research data is taken

³ In the Fedora repository, the deletion of a digital object yields a “tombstone”. The PID associated with this object then points to a tombstone notifying users that the resource has been deleted. Note that tombstones still require migration, meaning that PIDs still need to resolve to inform users that their associated digital objects have been removed.

over by another archive. Also, organizations that harvest metadata from the source repository should be informed within some notice period.

When research data is moved from a discipline-specific archive to an generic archive, there are bound to be service digressions but also service improvements. The generic archive will often only make available bibliographic metadata such as Dublin Core or MARC 21 for OAI-PMH harvesting. Detailed CMDI-based metadata, for instance, will hence be no longer available to search engines such as the CLARIN Virtual Language Observatory. On the other hand, less-rich metadata will enter library catalogues so that information about language resources suddenly appear to new user groups and become more widely visible and accessible. Both service digressions and progressions should be clearly sketched out and publicized.

3.4.4 Legal issues

When third party researchers deposit their data in an archive, they will usually do so based on a legal agreement between depositor and archive.⁴ Such depositing agreements need to be respected when research data is handed-over to another archiving institution. Ideally, the receiving end will honor such prior agreements, but issues related to sensitive research data have to be dealt with extra care.

The overall migration should be based on a legal agreement between giving and receiving end, addressing all issues with regard to sustainability (how long should all data be stored and made accessible), access (through which channels should the data be available), support (is research data being curated in the new archive at regular intervals to ensure that it can be viewed and processed using modern software and data formats) etc. The legal agreement will sometimes need to tackle complicated legal issues, especially when source and target archive are based in institutions that belong to different legal entities. While open licenses on digital objects (e.g., CC-BY 4.0) are easily dealt with, stricter licenses that need to be enforced by the target repository are more complicated to deal with. Extra time should be allotted. A white paper on using Creative Commons licenses for language resources discusses some of the intricacies involved (Kamocki and Ketzan 2014).

3.4.5 Transition and testing period

It is advisable to plan for an extensive testing and transition period, and that both source and target repository operate simultaneously over a sufficient amount of time before the source repository is deprecated. Test scripts help to ensure, for instance, that PIDs resolve as intended and that all AAI-related issues are dealt with as required. During the transition period, a combination of human-level tests and automatic scripts should also establish that all data has been transferred to the new repository system in a complete and correct manner. When source and target repository system use different checksum algorithms, additional efforts are required to ensure the data's integrity.

⁴ Usually, researchers working in the same organization that also hosts the archive do not need depositing agreements.

4 Discussion

The authors are not aware of reported similar migration 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 a 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.⁵

Our migration study profited from a common technological base as both archives used the Fedora Commons repository system. This will not always be the case. As a result, many design and migration decisions may well differ across institutions and their technological and organizational settings, which makes it hard to generalize. Nevertheless, there are a number of issues that the stakeholders involved need to address, of course, with varying degrees of complexity given the many different factors.

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 at least with a “CC BY-NC-SA 4.0” license (permitting users to copy and redistribute the material in any medium or format, and to remix, transform, and build upon the material for any non-commercial uses) 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>, say to check whether users are registered with an academic (non-commercial) institution.

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.

⁵ See <https://wiki.duraspace.org/display/FF/Training+-+Migrating+from+Fedora+3+to+Fedora+4>.

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 undertaken to describe research data in 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 preservation of rich metadata serves multiple purposes. First and foremost, it limits the disappointment of the data depositors who created such rich metadata in the first place. Also, once a dataset has been identified with the (lossy) bibliographic metadata of the library catalogue, the attached CMDI record allows metadata consumers to learn about research paradigm, experimental setups and other information they find essential to judge the datasets' relevance for their studies.

If the receiving institution is unable to maintain the original rich metadata format, it should explore whether information relevant to its consumers can be crosswalked, say, to a MARC 21 catch-all field that collects it in an unstructured form. While this harms the semantic interoperability across formats, its information-preserving nature helps swallowing the bitter medicine.

There is also the option that the receiving archive ports the research-specific metadata format (such as CMDI) to generic formats used by many other communities (such as some micro-format <https://en.wikipedia.org/wiki/Microformat>). When users are unable to identify relevant research data using the search capability of the library catalogue (which is driven by MARC 21 records), they can use generic web search engines that exploit semantic mark-up added to HTML tags. Here, for instance, a CMDI-based metadata record for some research data can be converted into an HTML page that displays the record in a user-friendly way. During the conversion the HTML source is semantically enriched by referring to, say, concepts of the schema.org ontology schema.org.⁶ The archive could thus create a site-map of its holdings, allowing search engines to access and index their site. Search hits would then direct users to the web-page holding all metadata including information where the corresponding research data can be downloaded. Library archives might be more inclined to maintaining such a mirroring structure rather than the discipline-specific metadata format.

In our scenario, research data were to be migrated from a research-specific institution to the library and computing centre, that is, a generic infrastructural organization. While such a scenario might be quite common across academic institutions, we did not anticipate it from the very start. That is, in the first phase of our project, we have built a research-specific institution on its own, without taking into account workflows, best practices, techniques, metadata standards etc. of the library world. On the one hand, this freed us from certain constraints, e.g., such as using the generic MARC 21 library standard, rather a well-adapted, semantically rich metadata such as CMDI. On the other hand, it created the aforementioned difficulties such as the lossy metadata conversion from CMDI to MARC 21. In hindsight, the authors would indeed favour a more cooperative model between research-specific specialists and the experts from the library world from the very

⁶ The ontology has, for instance, the concept 'MediaObject' which can be described with properties such as 'encodingFormat', 'bitrate', and 'duration', among many others.

start of any archival project, given that both partners have an open mind about the requirements from both ends of the spectrum. Such a cooperation will undoubtedly yield a number of compromises, but ideally results in the management of research data that is both sustainable and scalable as well as adoptable by research communities, e.g., by offering acceptable services for making data *findable*, *accessible*, *interoperable*, and *reusable* <https://www.force11.org/group/fairgroup/fairprinciples>.

5 Conclusion

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 gave their data to colleagues they know, and despite well established workflows, there were informal communication channels where 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 depositor 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.

Acknowledgements 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).

References

- Dima, E., Henrich, V., Hinrichs, E., Hinrichs, M., Hoppermann, C., Trippel, T., Zastrow, T., Zinn, C. (2012a). A Repository for the sustainable management of research data. In: Calzolari N, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (Eds) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, ELRA.
- Dima, E., Hoppermann, C., Hinrichs, E., Trippel, T., Zinn, C. (2012b). A metadata editor to support the description of linguistic resources. In: Calzolari N, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (Eds) *Proceedings of the Eight 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 (CMDI)—Part 1: the component metadata model. International Standard.
- Kamocki, P., Ketzan, E. (2014). Creative commons and language resources: general issues and what's new in CC 4.0. Tech. rep., CLARIN Legal Issues Committee (CLIC), White Paper Series. see https://www.clarin-d.de/images/legal/CLIC_white_paper_1.pdf.
- Lyse, G. I., Meurer, P., Smedt, K. D. (2015). Comedi: A component metadata editor. In *Selected Papers from the CLARIN 2014 Conference*, Linköping University Electronic Press 116(8):82–98.
- Trippel, T., Zinn, C. (2016). Enhancing the quality of metadata by using authority control. In *5th Workshop on Linked Data in Linguistic (LDL-2016) at LREC-2016*.
- Wilkinson, M. D. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, <https://doi.org/10.1038/sdata.2016.18>.
- Zinn, C., Trippel, T., Kaminski, S., Dima, E. (2016). Crosswalking from CMDI to Dublin Core and MARC 21. In: Calzolari N, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (Eds) *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, ELRA.

Web resources

- [U1] The Dublin Core Metadata Initiative, see <https://www.dublincore.org>.
- [U2] The MARC 21 standard, see <https://www.loc.gov/marc/bibliographic>.
- [U3] The EAD standard, see <https://www.loc.gov/ead/>.
- [U4] The MARC to EAD crosswalk, see <https://www.loc.gov/ead/ag/agappb.html#sec4>.
- [U5] The Handle system, see <https://www.handle.net>.
- [U6] The Fedora repository platform, see fedorarepository.org.
- [U7] ProAI, see proai.sourceforge.net.
- [U8] The OAI-PMH protocol, see <https://www.openarchives.org/pmh>.
- [U9] Apache Lucene and Solr, see 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] The Virtual International Authority File, see viaf.org.
- [U13] Example of a deposit agreement (University of Reading, UK), see researchdata.reading.ac.uk/deposit_agreement.html.
- [U14] Integrated Authority File (GND) at the German National Library, see <https://www.dnb.de/EN/Standardisierung/GND/gnd.html>.

[U15] The Library of Congress Control Number, see id.loc.gov/authorities/names.html.

[U16] The International Standard Name Identifier, see isni.org.

[U17] On micro-formats, see <https://en.wikipedia.org/wiki/Microformat>.

[U18] The Schema.org vocabulary, see schema.org.