On the Successful Migration of Languages Resources from one Repository to Another

Claus Zinn and Thorsten Trippel
Department of Linguistics
University of Tübingen
Keplerstr. 2, 72074 Tübingen, Germany
claus.zinn@uni-tuebingen.de
thorsten.trippel@uni-tuebingen.de

Abstract

More than five years ago, we crafted a detailed scenario for migrating our research data from our locally-maintained, departmental repository to an external, institutional repository for which we had only little control over. Now, with the rising cost of updating and maintaining our repository software to the latest version, personnel fluctuation, and the opportunity to use data services of a newly founded Digital Humanities Center, we decided to put into practise the scenario step by step. This paper describes the actual challenges we encountered in the migration process, the deviations from the original scenario and the compromises we needed to make, and finally, how we succeeded to get all data transferred in a safe and information-preserving manner.

1 Introduction

The maintenance of a research data repository comes with substantial costs. While a large part of the effort is devoted to data curation, metadata annotation and ingestion as well as to the communication with data depositors and consumers, there is a significant workload involved in keeping the repository software up-to-date. Security updates are a major concern; at short notice, they must be deployed in a running repository system to make it less vulnerable to external threats. From time to time, old versions of repository systems are deprecated and stop benefiting from security patches. In this case, one has to perform a major upgrade to a newer version of the software. Costs related to software maintenance are rising, and some institutions may consider migrating their research data to an external, infrastructural organisation that is experienced with research data management (RDM), already hosts research data from a variety of other disciplines, and has trained staff. Scaling up certainly helps to keep expenditures in check. Ultimately, the turnaround of key personnel triggered our migration process. The process was aided by a migration workflow that was started more than five years ago (Trippel & Zinn, 2018, 2021). We have now executed the workflow, and report on the actual challenges and difficulties we encountered, some of which were anticipated, others newly arose unexpectedly from new, unforeseen technical requirements.

2 Background

In the absence of proper research data management services at a university-wide level, in 2010, our department kickstarted its own repository system. The "Tübingen Archive for LAnguage Resources" (TALAR) was targeted at researchers of two Collaborative Research Centres[1] (CRC-441, CRC-833) to provide them with a centralised storage solution for all data they created. Also, the data created by CRC-external activities of our institution got a new, central archiving home. We started with a system based upon Fedora Commons[2] (version 3), which we extended with a number of essentials such as a shell-based environment to support data ingestion and rights management, and an OAI-PMH port[3] to make

[1] German:"Sonderforschungsbereich", abbreviated as "SFB"

[2] https://fedora.lyrasis.org

[3] See https://www.openarchives.org/pmh/

available metadata to external parties. Due to security reasons, we later updated the system to version 4 of Fedora. Security patches available for this version where applied whenever possible.

Having control over your own repository comes with a significant amount of responsibility, but it also opens up a design space around many aspects of research data management, *e.g.*, how to best describe research data with metadata; and what research data should be accepted for ingestion (only internal data stemming from your own institution, or data from other institutions, or only data passing some quality threshold)? In the past decade, we have fitted the Fedora repository system with a number of bells and whistles, namely, a web-based, graphical user interface to serve as a front-end of the repository for both users and archivists (Dima, Henrich, et al., 2012), the Bagman software for researchers to help them describing and transferring their data to the archive (Zinn, 2022), and the ProFormA editor to help them annotating their research data with metadata (Dima, Hoppermann, Hinrichs, Trippel, & Zinn, 2012). We have also defined, and redefined, a good number of CMDI profiles to have adequate means to describe different types of resources with rich metadata, and we have also implemented crosswalks between CMDI to Dublin Core and MARC-21 (Zinn, Trippel, Kaminski, & Dima, 2016).

The repository software used to run on a designated, virtual, Unix-based machine at the university's computing centre. It benefited from regular back-ups, and all its content was regularly mirrored onto a system running in a different physical location.

Our repository has been awarded with the *Data Seal of Approval*[4] in 2013 and 2015 and with the Core Trust Seal[5] afterwards (last renewed in 2023). From its initial days, our repository has also been a certified CLARIN-B centre.[6] The TALAR repository took part in the CLARIN harvesting infrastructure.[7] The CLARIN OAI-PMH harvest manager contacted TALAR at regular intervals to download its 670+ dataset descriptions. The harvesting result showed up in the CLARIN Virtual Language Observatory (VLO).[8] VLO users could browse all metadata in the VLO and could click on the handles that pointed to the dataset's landing page or to individual resources of the dataset.[9]

More than five years ago, within the NaLiDa project[10], a detailed migration workflow was crafted to migrate all research data to another repository, maintained by the university library of the University of Tübingen (Trippel & Zinn, 2018, 2021). Due to changes in personnel, and the arising opportunity to store research data in the newly-founded Digital Humanities Center, we are now putting the migration workflow into practice.

By and large, the main issues of the migration workflow outlined at the time are still the main issues to be tackled: user authentication and authorization, metadata harmonisation, and persistent identifier management. At the time, the migration would have benefited from a common technological base as both the source and the target repository system were based on Fedora. In the meanwhile, however, the university library has expanded its eScience services into a newly founded *Digital Humanities Center (DHC)*.[11] In due course, the DHC staff also deployed new repository software, which was now based on InvenioRDM.[12] This has complicated the migration process along all dimensions.

Our migration process was driven by the technical state and content of TALAR (henceforth, *source repository*), and the technical and organisational requirements of the target repositories.[13] We now give a more detailed description of the source repository.

[4] http://www.datasealofapproval.org

[5] https://www.coretrustseal.org

[6] https://www.clarin.eu/content/certified-b-centres

[7] https://centres.clarin.eu/oai_pmh

[8] At https://vlo.clarin.eu, select collection "Tübingen Archive of Language Resources (TALAR)".

[9] We also used to provide an HTML-based representation of all metadata with a sitemap through our institutional web-server, supporting researchers to discover all data more easily.

[10] http://www.sfs.uni-tuebingen.de/nalida/en/

[11] See https://uni-tuebingen.de/forschung/forschungsinfrastruktur/digital-humanities-center/.

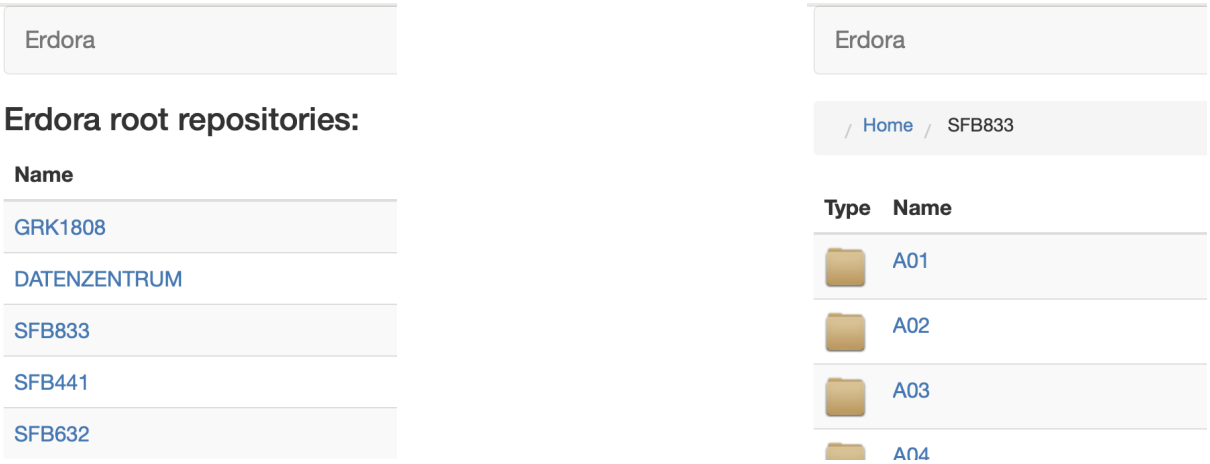
[12] https://inveniosoftware.org/products/rdm/

[13] As will be explained further below, we decided to use the DHC repository for genuine research data and to use the Zenodo repository for all other data.

3 The TALAR Repository

The TALAR repository was operational from 2010. At the time of migration, it contained 673 datasets, totalling hundreds of gigabytes of data. Each dataset came with CMDI-based metadata, for which we have developed a number of CMDI profiles to accurately describe the various types of research data we host: text corpora, speech corpora, lexical resources, tools, web services, experimental data, and teaching material.

Each dataset was addressed by a persistent identifier using the handle system[14]. Part identifiers were used to address individual files inside a dataset. Handles without part identifiers pointed to the HTML-based *landing page* of a dataset, which the TALAR website automatically rendered by using a CMDI-2-HTML based transformation of the dataset's CMDI file. [15].



(a) Root Structure

(b) WP Structure

Figure 1: The ERDORA GUI

Fig. 1a shows the top-level organisation of all research data in TALAR. In fact, most of our research data originated from two collaborative research centres, SFB441 and SFB833. But the repository also hosted resources that we have developed in our own department such as the lexical-semantic wordnet GermaNet (Hamp & Feldweg, 1997), and treebanks such as TüBa-D/Z[16], and other annotated text corpora such as the Index Thomisticus[17]. Those resources were hosted in the Fedora node "DATENZENTRUM". Resources from a *Graduiertenkolleg* and from an institution-external CRC were stored in GRK1808 and SFB632, respectively.

In TALAR, all research data were hierarchically structured. Fig. 1b shows a fragment of the SFB833 tree, mirroring the working packages of this collaborative research centre. Usually, a work package node contained multiple datasets, and often, each dataset itself exhibited a complex directory structure. Rarely was such data compressed in archive files, say in ZIP format. In fact, the CRC-833 research data had quite a few datasets that consisted of hundreds of individual files, each of which was addressable by a handle-based part identifier. Moreover, the underlying Fedora Commons software made is possible to assign access permissions at individual directory and file levels. In TALAR, those *Access Control Lists* were used to give authenticated individual users (usually, members of the two CRCs) read access to (non-public) datasets.

Note that the CMDI-based description of a dataset listed all its resources and defined their handle-based address space, see Fig. 2a. Handles with part identifiers allowed users to directly download in-

[14] https://handle.net

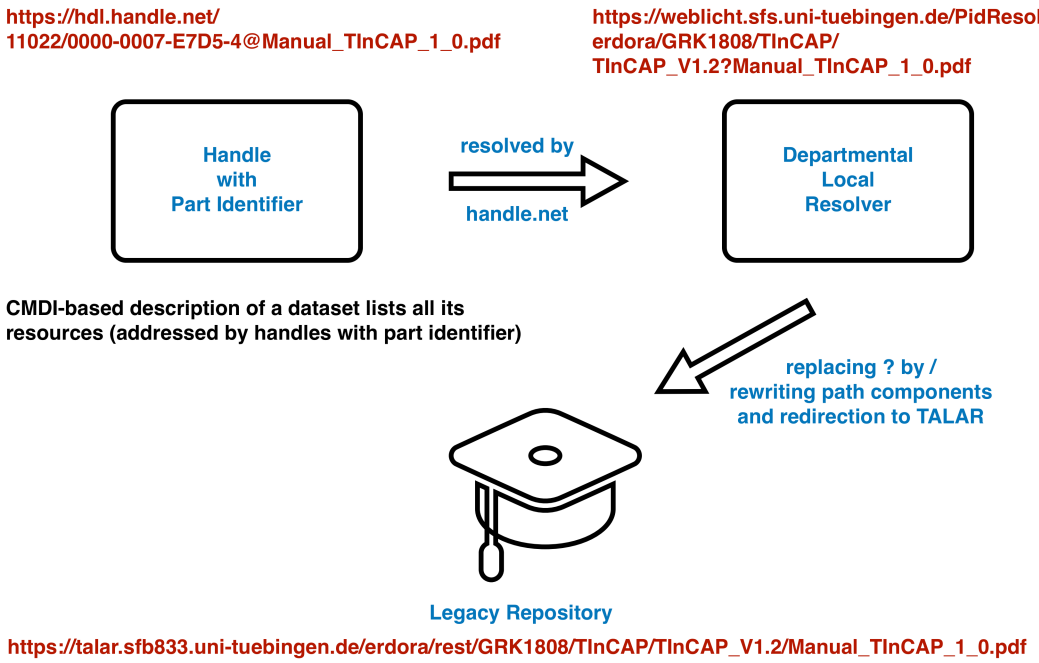
[15] For this purpose, a processing instruction in the XML file invoked an XSL transformation.

[16] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/corpora/tueba-dz/

[17] https://uni-tuebingen.de/fakultaeten/philosophische-fakultaet/fachbereiche/neuphilologie/seminar-fuer-sprachwissenschaft/arbeitsbereiche/allg-sprachwissenschaft-computerlinguistik/ressourcen/corpora/index-thomisticus-baumbank/

```
<cmd:ResourceProxyList>
  <cmd:ResourceProxy id="TInCAPexportv12xml">
    <cmd:ResourceType mimetype="text/html">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://hdl.handle.net/11022/0000-0007-E7D5-4@TInCAP_export_v1.2.xml</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="ManualTInCAP10pdf">
    <cmd:ResourceType mimetype="application/pdf">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://hdl.handle.net/11022/0000-0007-E7D5-4@Manual_TInCAP_1_0.pdf</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="landingpage-11022-0000-0007-E7D5-4">
    <cmd:ResourceType mimetype="text/html">LandingPage</cmd:ResourceType>
    <cmd:ResourceRef>https://hdl.handle.net/11022/0000-0007-E7D5-4</cmd:ResourceRef>
  </cmd:ResourceProxy>
```

(a) Handles with part identifiers in the CMDI file



(b) Handles and PID Resolving for TALAR-based Resources

Figure 2: Handles and how they were resolved

dividual files of a research dataset. Fig. 2b depicts the process that took place when users clicked on such persistent identifiers. Note that the rewriting behaviour of part identifiers has been configured per handle prefix, where `@partIdentifier` was rewritten to `?partIdentifier`. Also note that the main part of the handle did not point directly to the source repository but to a locally maintained local resolver, which rewrote the URL fragment weblicht.sfs.uni-tuebingen.de/PidResolver/erdora/ to talar.sfb833.uni-tuebingen.de/erdora/rest/ and replaced the question-mark announcing the part identifier with a forward slash, see the URL pointing into the source repository at the bottom of the figure.

Data Curation In normal operation, once a dataset was ingested into a repository, there was little, if any, work, associated with it. Occasionally, researchers contacted us to update the metadata with new contact information, and when a new version of a dataset was released, a new dataset and its updated metadata was ingested into the repository (equipped with a new handle and a number of part identifiers). Licences were rarely changed.

In the past decade, several archive manager took care of research data management, and consequently, their are subtle differences in archiving standards.[18] During the times, we have improved the CMDI

[18]We found some of the research data poorly described; we occasionally encountered CMDI-based metadata that, for instance, used obscure titles, failed to specify resource creators, or had ambiguous licence statements.

profiles in use, clearly separating administrative metadata (such as project-related information) from resource-type specific information (such as for the description of lexical resources or text corpora). In the last years, Bagman, a web-based software that supports researchers and archive managers alike to ingest research data into the repository, was developed and deployed. With Bagman, users pack their data using the BagIt format (Kunze, Littman, Madden, Scancella, & Adams, 2018) and fill out forms to describe it with metadata. All this information is then used to automatically instantiate the appropriate CMDI profile. The automatic generation of CMDI files improved the quality of the metadata considerably and proved superior to the manual creation of such data where XML editors such as ProForma or simple text editors were used.

The migration of the entire repository created a situation were we stepped back and took a bird's-eye view, seeing all the datasets accumulated so-far for the first time in their entirety. This perspective revealed a number of organisational issues that the migration sought to address:

- a significant part of the the CRC-833 collection have datasets that consist of only a single text file (in all cases some scientific text in PDF format). Those 265 PDF articles played the role of a reference collection to support scientific research across the many members of the CRC-833. While some of those papers have been published in journals, others were manuscripts submitted for review or were in press. For these reasons, all datasets were assigned a non-public licence.

- we found thirteen datasets that should have received their own root node. Each dataset contains teaching material (presentation slides, exercises, data files) targeted at graduate and PhD students. These datasets were assigned an open CC-BY licence.

- the repository also contained 54 service descriptions for the WebLicht workflow engine (Hinrichs, Zastrow, & Hinrichs, 2010). WebLicht contacts a harvester[19] whose task is to source available service descriptions from our repository and others parties via their respective OAI-PMH endpoints.[20] It is paramount that those CMDI-based WebLicht service descriptions in TALAR must continue to be available. Note, however, that these services were ingested into TALAR not for archival but for pragmatic reasons, namely, for the sole purpose to be discoverable via an OAI-PMH port.

- the repository also contained 28 publications from the TLT-13 conference "Treebanks & Linguistic Theories", which took place in Tübingen in 2013. Their ingestion into TALAR made them publicly accessible via a persistent identifier. As no research data was attached to any of the papers, a simple preprint server would have been a better match than TALAR.

- there were two sparsely populated root nodes; the subdirectory SFB632 (CRC-632) had only two datasets, and the directory GRK 18080 ("Graduiertenkolleg 1808") had only 5 datasets.

The bird's-eye view revealed that our datasets consisted of four different types: genuine research data, scientific publications, teaching material, and software configuration data. We decided to only move genuine research data to FDAT, the repository of the Digital Humanities Center.[21] All other data must be moved to different homes.

We started the migration process with the following goals in mind: (i) preserve the hierarchical structure of research data; (ii) migrate all research data (independent of its age and quality); (iii) ensure that there is no information loss in terms of metadata; and (iv) strive for minimal service disruptions.

4 Migration

Originally, we thought that the repository of the Digital Humanities Center of the University of Tübingen (FDAT) should receive all of our TALAR data. However, we decided that our CRC-833 collection of

[19] http://weblicht.sfs.uni-tuebingen.de/apps/harvester/resources/services

[20] At last count, 572 services are harvested in total, from many dozen repositories. The URL request for our repository (TALAR) is https://talar.sfb833.uni-tuebingen.de/erdora/rest/oai?verb=ListRecords&metadataPrefix=cmdi&set=WebLichtWebServices.

[21] FDAT is an acronym built from the German word 'Forschungsdaten'.

research articles, despite their non-public licence, as well as the collection of teaching material and the collection of TLT-13 papers should be migrated to elsewhere and that Zenodo[22] was found a better fit. For the genuine research data, we created designated *communities* for data from the two main CRCs in FDAT, and for our own research data (and the one from SFB632 and GRK1808) we defined the TALAR community there. Since neither FDAT nor Zenodo offers an OAI-PMH port, none of the CMDI files carrying WebLicht services descriptions could be migrated. Instead, we extended the source code of WebLicht, which is software we developed in house. The workflow engine has been enabled to also "harvest" service descriptions from subdirectories attached to WebLicht's source code distribution; here, a new subdirectory was created to host the 54 service descriptions.

Table 1 depicts the new organisation.

Type of resource	**Target**	URL of the Community	Number
PDF Articles (CRC-833)	Zenodo	https://zenodo.org/communities/sfb-833-literature	265
Teaching Material	Zenodo	https://zenodo.org/communities/talar-teaching-material	13
TLT	Zenodo	https://zenodo.org/communities/tlt13	28
Research data (CRC-441)	FDAT	https://fdat.uni-tuebingen.de/communities/crc441	28
Research data (CRC-833)	FDAT	https://fdat.uni-tuebingen.de/communities/crc833	227
All other research data	FDAT	https://fdat.uni-tuebingen.de/communities/talar	13
WebLicht Service data	–	`transferred to WebLicht Github Repository`	54

Table 1: Overview of Target Repositories

4.1 Migration to Zenodo

Like FDAT, Zenodo is a repository system that is based upon InvenioRDM, and hence, it also allows the organisation of research data into communities. For our purposes, three communities were created, one to hold the literature from the CRC-833, a second one to take on TALAR's teaching material, and a third one to hold the TLT papers.

The ingestion of literature data was rather straightforward. With each PDF file being complemented with a CMDI-based metadata description, we wrote an XSLT stylesheet to extract the relevant metadata into the required DataCite[23] fields, namely, author, title, publication date, and description. It showed that the CMDI files did not have more information that needed to be preserved, and therefore, no information loss incurred. Consequently, we did not ingest any CMDI files to the Zenodo CRC-833 community. Also, due to copyright issues, it was required that all research data in the Zenodo community "sfb-833-literature" was restricted. Interested parties can contact the community curator for which we have created a new special-purpose email account.[24]

The teaching material had a complex nature. They usually consisted of many files, some of which were hierarchically structured, and used a variety of different data formats. We found their CMDI-based description rather shallow, not making use of the potential that the CMDI profile "CourseProfile" offered. As a result, we also omitted the ingestion of CMDI profiles to Zenodo. The teaching material of each dataset was archived in ZIP format to preserve their hierarchical structure, and subsequently ingested into the Zenodo community "talar-teaching-material" with a CC-BY licence.

The TLT data consisted of 28 PDF files with corresponding CMDI-based annotations that carried little information other than DataCite fields. Without loss of information, only the TLT papers were ingested into the TLT community, but not their CMDI metadata.

In sum, 306 datasets (research articles, teaching material and TLT-13 papers) left the realm of Tübingen University and found their new home in the Zenodo repository. All SFB-833 literature data was automatically ingested into Zenodo using a Python-based script that makes use of the Zenodo developer

[22] https://zenodo.org

[23] https://datacite.org

[24] The email account data-steward@semsprach.uni-tuebingen.de also answers requests from the other Zenodo community and the newly created FDAT communities.

API.[25]. The teaching material as well as the TLT-13 data was manually ingested into Zenodo.

4.2 Migration to FDAT, the Institutional Research Data Repository

There were less than 300 datasets still to be taken care of. In terms of content and size, they constituted the "real" research data. To mirror the high-level structure of the TALAR source repository, we have created three communities on the new institutional FDAT repository, see Tab. 1: (i) the CRC-441 community to host all data stemming from the Collaborative Research Centre "Linguistic Data Structures", (ii) the CRC-833 community to host all data originating from the Collaborative Research Centre "Emergence of Meaning", and (iii) the "Tübingen Archive for LAnguage Resources" (TALAR) community.

The target repository defined a number of hard constraints that we needed to deal with:

- CMDI-based metadata is not an accepted metadata standard for the description of research data; all research data must be described using DataCite;
- the size of each dataset is limited to 100 GB and cannot contain more than 100 individual files;
- FDAT only supports lists of resources, not hierarchically structured ones;
- access to a dataset is either restricted for all users or available to all. No individual rights can be associated with the files of a dataset; and
- all datasets must be addressed with DOI-based persistent identifiers; existing handle-based persistent identifiers cannot be reused.

Apart from these constraints, FDAT offered two benefits: it gives support for the versioning of datasets as such dependencies between datasets can be spelled out explicitly; and it supports the creation of communities so that research data that stemmed from different players could be easily grouped together.

While there was little data curation necessary for transferring literature data or course material from the source repository to the three Zenodo communities, the situation was different for the remaining, genuine research data.

Moreover, many of our research datasets have a deep directory structure, which is not supported by FDAT. In these cases, the hierarchy was "flattened", usually by replacing them with ZIP archives so that their unarchiving reestablishes the hierarchy. Moreover, some datasets had information duplicated. Sometimes, the files of a dataset were complemented with a ZIP archive that also held all files. Such redundancy was removed, consistently in preference for the ZIP archive.

Also, some datasets failed to attribute the agency that funded the project producing the research data. In these cases, the funder for the two CRCs was manually added to the DataCite metadata. Last, while some CMDI-based metadata used authority file information from GND (https://gnd.network) and VIAF (https://viaf.org) to uniquely identify persons associated with the research data, no such information was given for the organisations the persons work for. In DataCite, we have complemented information about organisations with their ROR identifier.[26]

Given the complexity of the research data, the migration to FDAT required us to review each dataset individually. During the review, we observed other, mostly minor, oversights or flaws in the metadata, which we corrected in due course. We also contacted some of the researchers that produced the research data to review our migration work. This made us realize that an acceptable translation from CMDI to DataCite is far from trivial, and must take into account a few subtle but important aspects. In the CMDI metadata, for instance, we find the CMDI component `Project`[27] , which contains information about the project in which a resource was created. In the first iteration, we falsely associated the person mentioned in the Project component as the *creator* of the resource (with role "Project leader"), where in fact, the

[25] https://developers.zenodo.org

[26] The German Research Foundation has the ROR identifier https://ror.org/018mejw64.

[27] See the CLARIN component registry at https://catalog.clarin.eu/ds/ComponentRegistry; select group name "NaLiDa" and search for "Project".

DataCite *contributor* with that role should have been used. As a result, the dataset's citation as generated by the FDAT GUI was misleading as it omitted the actual creators of the resource.

To avoid any loss of information given in the CMDI metadata description of a dataset, we made the CMDI file an integral part of the dataset itself. While the DataCite metadata can be altered after the publication of a dataset, this is not the case for the research data itself. Consequently, it was crucial to ensure that all CMDI-based metadata was in its final, publishable state. As a result, each CMDI file was diligently reviewed by the communities' curator before the entire dataset it describes was published.

4.2.1 User authentication and Authorization

While the source repository had an expressive rights management system in place, the target repository had no capabilities for user authentication and authorization to cater for such personalised access. Data still under publication embargo will continue to be inaccessible to *all* users in the target repository; and this includes the data creators. For most datasets, the embargo date has been set to September 30, 2026. Interested parties must contact the data steward of the FDAT communities (CRC 833, CRC 441, & TALAR), or the contact persons specified in the DataCite metadata. Once the embargo data passes, all research data in FDAT will become publicly available under a CC-BY licence.

4.2.2 Persistent Identification

The FDAT repository requires the use of DOI-based persistent identifiers.[28] In principle, the FDAT repository allows the addressing of individual files of a dataset. For this purpose, however, no DOI can be used, only the resolved URL. [29] However, the FDAT administrator cannot guarantee that the resolved URL, or its current path, will continue to be serviced in the future. If we were to continue supporting part identifiers, then we must continue to support our local resolver, see Fig 2b and maintain its mapping table to do the rewriting, and change the rewriting whenever the target URL of FDAT changes.

To minimize our commitments, and to keep things simple, we stopped our support for part identifiers; access to all research data is via their new FDAT-based landing page. All handles that we registered at `handle.net` now directly point to their respective DOI handles of FDAT, hence bypassing any local URL resolver, which we have therefore retired. Handles, with or without part identifiers used in CMDI files before (see Fig. 2a) have been replaced by their respective DOIs (see Fig. 3a). Users trying to invoke legacy handles with part identifiers, say because they have bookmarked them, will find themselves redirected to the landing page of the dataset in FDAT.

Fig. 3b illustrates the resolving of a handle-based persistent identifier with a part identifier, which gets redirected to doi.org. Note that the part identifier initiated with the "at"-sign "survives" the first redirection; the part identifier now appears as URL argument of the DOI-based address. However, when doi.org resolves the DOI-based URL, it ignores all its arguments so that the final URL is free from any path information.

4.2.3 Metadata Provision

The migration required us to convert CMDI-based metadata to DataCite. To minimize information loss, some resource-specific metadata (such as information about, say an experimental design, or the number of entries in a lexical resource) was written into one of Datacite's description fields. Moreover, the original CMDI-based description became part of the research data stream so that users can consult the metadata for information that cannot be (or has not been) mapped to DataCite.

The source repository offered OAI-PMH harvesting for metadata in DC, MARC 21, and CMDI. Zenodo and FDAT, however, only support metadata harvesting for DataCite. With the source repository being shut down, we had to search for an alternative solution to OAI-PMH servicing that provides CMDI-based metadata to the Virtual Language Observatory (VLO) and other interesting parties.

For this purpose, we have implemented our own OAI-PMH endpoint in the Prolog programming language.[30] The endpoint's content is sourced from a directory that contains the CMDI files of all data

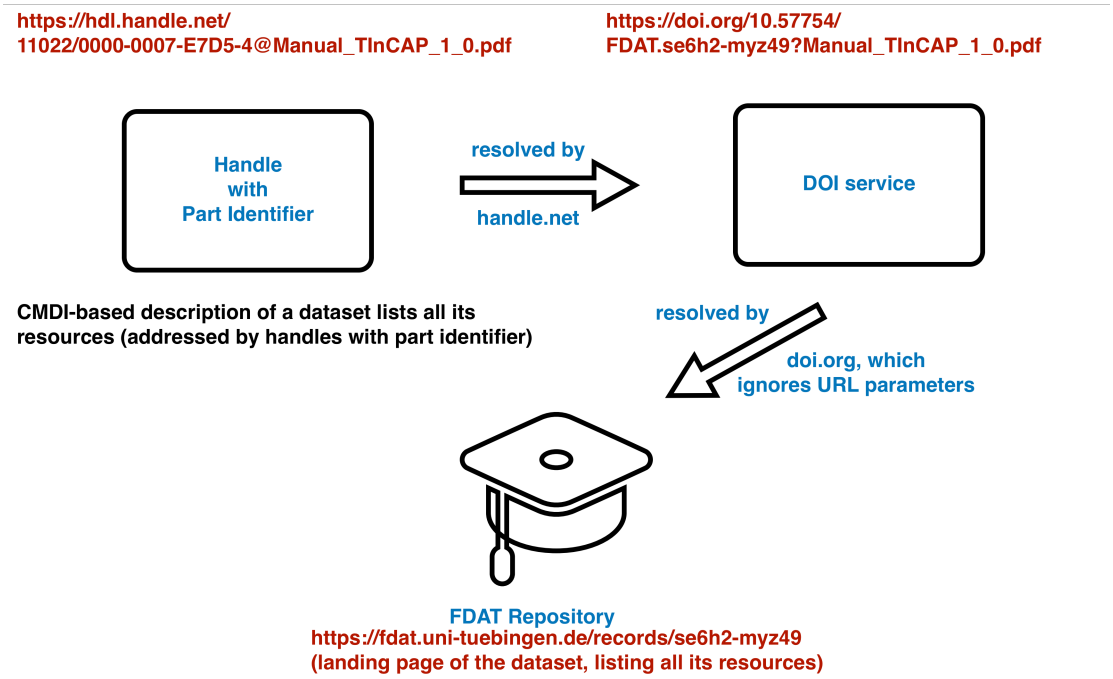
[28] The DOI https://doi.org/10.57754/FDAT.c0cvj-4vk83, for instance, resolves to https://fdat.uni-tuebingen.de/records/c0cvj-4vk83.

[29] For example, https://fdat.uni-tuebingen.de/records/c0cvj-4vk83/files/de-nn-com-sem_8005_compounds.txt

[30] http://textplus.sfs.uni-tuebingen.de:8088/api/oai?verb=Identify

```
<cmd:ResourceProxyList>
  <cmd:ResourceProxy id="TInCAP_export_v12xml-11022-0000-0007-DEBD-B">
    <cmd:ResourceType mimetype="application/xml">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://doi.org/10.57754/FDAT.se6h2-myz49</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="Manual_TInCAP_10pdf-11022-0000-0007-DEBD-B">
    <cmd:ResourceType mimetype="application/pdf">Resource</cmd:ResourceType>
    <cmd:ResourceRef>https://doi.org/10.57754/FDAT.se6h2-myz49</cmd:ResourceRef>
  </cmd:ResourceProxy>
  <cmd:ResourceProxy id="LandingPage-11022-0000-0007-DEBD-B">
    <cmd:ResourceType mimetype="application/xml">LandingPage</cmd:ResourceType>
    <cmd:ResourceRef>https://doi.org/10.57754/FDAT.se6h2-myz49</cmd:ResourceRef>
  </cmd:ResourceProxy>
```

(a) DOI-based identifiers in the CMDI file



(b) Handles and PID Resolving for TALAR-based Resources

Figure 3: Handles and PID Resolving for TALAR Resources (after migration)

migrated to FDAT. The name of each file follows a naming scheme that encodes two pieces of metadata, namely, the FDAT community, and the original publication date of the research data as well as a running number (*e.g.*, `sfb833_2020-10-26_0004.xml`). Each CMDI-based file is mirrored by a metadata description in Dublin Core, which has been automatically generated using an XSL-based crosswalk (Zinn et al., 2016). [31] – At September 30, 2026 latest, our simple OAI-PMH service will be transferred to another (yet to be identified) organization to ensure that all research data remains visible to the VLO for the foreseeable future.

Tombstoned data In contrast to our initial aim to migrate all research data independent of its age and quality, we encountered ten datasets where we made an exception. In the source repository, we found occasional datasets that were either incomplete, clearly did not merit long-time archiving, or constituted merely test data. Such data were tombstoned. For this purpose, we created a 'Metadata only' entry in FDAT, see https://doi.org/10.57754/FDAT.kvd3z-7w002. It lists all the tombstoned datasets via their respective handles, which in turn now all point to the aforementioned DOI.

[31] Recall that CMDI descriptions for all data migrated to Zenodo have been removed; the TALAR collection in the VLO will hence shrink to around 300 datasets of genuine research data.

Transitional period We have now migrated all research data to the new repositories. All handles have been redirected to DOI-based identifiers pointing to FDAT or Zenodo. The new OAI-PMH endpoint has replaced our legacy endpoint so that CMDI-based descriptions in the VLO now only have DOI-based identifiers (saving one level of redirection). Users with bookmarked handle-based identifiers, with or without part identifiers, will always be directed to a research dataset's landing page. Those interested in a particular part of the dataset may now be required to download a ZIP-based archive of the entire dataset to then extract the specific item of their interest.

For new research data, our institution will continue to provide help to researchers who would like to archive their research data in a trusted, sustainable environment. For the data we accept, end-users can expect to receive the same quality of service as before; annotation of research data with CMDI profiles will continue. The CMDI metadata will be part of the dataset, and distributed via our OAI-PMH service. Upon the closure of the institute, all users will be directed to another CLARIN-B centre or asked to contact the FDAT archive manager. At this point, FDAT users will not be required to provide CMDI-based metadata annotations.

5 Conclusion

The migration of research data from one repository system to another is no easy matter and is bound to create issues that cannot always be resolved without making compromises. The actual effort for migrating all data involved numerous internal discussions, coordination with the FDAT manager, the authoring of XSL-based stylesheets for ingest mechanisation, the adaptation of the WebLicht software to fetch their CMDI-based service descriptions in a modified manner, the editing of CMDI files, and the reconfiguration of handles to point to new target DOIs. Our discussion shows that the migration of research data needs careful planning and execution, and that any migration efforts must be started well in advance.

The migration of research data freed us from maintaining a good number of software packages. ProForma, the Erdora shell and its GUI, the OAI-PMH plugin, the local resolver, and the entire Fedora repository has been retired. The only software that is run for a longer period of time is the OAI-PMH service to make available all CMDI files to the CLARIN VLO, ensuring that the visibility of our language resources stays high.

Our department has offered a repository system since 2010; it started at a time where research data management was underdeveloped in the infrastructural institutions of our university. At the time, we had little alternatives to perform research data management other than doing it ourselves. Supported by national and international funding from CLARIN, we developed an entire eco-system around RDM. This time is now coming to an end. The old repository, which served the CLARIN community well, has been dismantled. What remains are our Github repositories that continue to hold our source code for the software we built. In particular, we hope that the Bagman software attracts some interest from other parties as it provided tremendous support to our archiving workflow.

For us, our work in research data management does not stop. We will continue to advise researchers in these matters. We will continue to take on new research data, help getting it properly annotated with CMDI metadata, ingest it into FDAT, and making available its metadata via OAI-PMH to the CLARIN Virtual Language Observatory and now also to the Text+ registry.[32] Relieved from the burden of running the daily business of keeping a repository running and up-to-date, we can focus on other important aspects of research data management.

6 Acknowledgements

In the last decade, the TALAR repository profited from a number of research grants, in particular, the German Research Foundation within the NaLiDa project which ran von 2009 to 2016[33]. We also profited from funding within the national CLARIN-DE project and the pan-European CLARIN project.

[32] See https://registry.text-plus.org.

[33] https://gepris.dfg.de/gepris/projekt/88614379

The migration work described in this paper has been supported by the Text+ project[34], funded by the German National Science Foundation (DFG), project reference 460033370.

We thank the repository manager of FDAT, Dr. Steve Kaminski, for his invaluable (and continuing) support during the migration of our data.

References

Dima, E., Henrich, V., Hinrichs, E., Hinrichs, M., Hoppermann, C., Trippel, T., Zastrow, T., & Zinn, C. (2012). A repository for the sustainable management of research data. In *Proceedings of the eighth international conference on language resources and evaluation (LREC)* (pp. 3586–3592). ELRA.

Dima, E., Hoppermann, C., Hinrichs, E., Trippel, T., & Zinn, C. (2012). A metadata editor to support the description of linguistic resources. In *Proceedings of the eighth international conference on language resources and evaluation (LREC)* (pp. 1061–1066). ELRA.

Hamp, B., & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the ACL workshop automatic information extraction and building of lexical semantic resources for nlp applications*. Madrid, Spain.

Hinrichs, M., Zastrow, T., & Hinrichs, E. (2010). WebLicht: Web-based LRT services in a distributed eScience infrastructure. In *Proceedings of the seventh conference on international language resources and evaluation (LREC)*, ELRA.

Kunze, J., Littman, J., Madden, E., Scancella, J., & Adams, C. (2018). *The bagit file packaging format (v1.0)*. RFC 8493, DOI 10.17487/RFC8493. See https://www.rfc-editor.org/info/rfc8493.

Trippel, T., & Zinn, C. (2018). Lessons learned: On the challenges of migrating a research data repository from a research institution to a university library. In *Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018*, ELRA.

Trippel, T., & Zinn, C. (2021). Lessons Learned: On the Challenges of Migrating a Research Data Repository from a Research Institution to a University Library. *Language Resources and Evaluation*, *55*, 191–207. Springer.

Zinn, C. (2022). Bagman – a tool that supports researchers archiving their data. *Linköping Electronic Conference Proceedings*, *189*, 181–189. Selected papers from the CLARIN Annual Conference 2021. Ed. by Monica Monachini and Maria Eskevich.

Zinn, C., Trippel, T., Kaminski, S., & Dima, E. (2016). Crosswalking from CMDI to Dublin Core and MARC 21. In *Proceedings of the tenth international conference on language resources and evaluation (LREC)*, ELRA.

[34] See https://www.text-plus.org/en/home/.