A Repository for the Sustainable Management of Research Data

Emanuel Dima, Verena Henrich, Erhard Hinrichs, Marie Hinrichs, Christina Hoppermann, Thorsten Trippel, Thomas Zastrow and Claus Zinn

Department of Linguistics, University of Tübingen
Wilhelmstr. 19, 72074 Tübingen, Germany
firstname.lastname@uni-tuebingen.de

Abstract

This paper presents the system architecture as well as the underlying workflow of the Extensible Repository System of Digital Objects (ERDO) which has been developed for the sustainable archiving of language resources within the Tübingen CLARIN-D project. In contrast to other approaches focusing on archiving experts, the described workflow can be used by researchers without required knowledge in the field of long-term storage for transferring data from their local file systems into a persistent repository.

Keywords: primary research data repository, archiving workflow, archiving support

1. Introduction

Large amounts of research data are currently being stored by their resource creators on various devices, often leaving them inaccessible to the research community at large and sometimes even to the creators themselves. These data graveyards may contain treasures of results, possibly not only for the individual compiling the resource. With the help of a repository system, the challenges of archiving can be overcome, respecting the privacy and property rights of the researcher. Persistent accessibility of research data represents the main purpose of a repository system, but it also enhances the visibility and searchability of the data. This is implemented by a repository system in the following manner:

- persistent archiving of resources: data is stored and maintained in a consistent form within dedicated infrastructures with no interference by other processes;
- addressing the resources from outside of the repository, making them citable and hence visible to the academic public by maintaining resolvable identifiers;
- describing the resources consistently, allowing for searches over their descriptions using metadata;
- sharing of resources, either by providing download options for the general public or by permitting user access for authorized users only.

The infrastructure presented in this paper has been built by the Tübingen CLARIN-D project (see http://clarin-d.net), which is part of the European ESFRI CLARIN initiative (see www.clarin.eu). This federation of CLARIN centeres provides a digital infrastructure for language resources and tools in the humanities and social sciences. Since these centers, among other objectives, strive to make resources available in a persistent manner, it is necessary to establish an adequate foundation for the sustainable management of research data. This has been put into practice by developing the ERDO repository system (Extensible Repository System of Digital Objects) within the Tübingen CLARIN-D infrastructure.

In view of the data's visibility, it is necessary to describe resources by means of metadata. Therefore, ERDO supports the Component Metadata Infrastructure (CMDI, (Broeder et al., 2010), (Broeder et al., 2012)) which has been built in the context of the European CLARIN project (see http://www.clarin.eu/cmdi). Characterized by its flexibility, CMDI is particularly suitable for the description of different types of linguistic resources (Barkey et al., 2011a), such as text or speech corpora, lexical resources, experiments, or software tools. Both metadata and research data are ingested into the ERDO repository system and made subsequently accessible via various search tools. Examples provide (combinations of) full-text or facetted search using catalogues, such as the CLARIN Virtual Language Observatory (VLO, see http://www.clarin.eu/vlo/) or the facetted browser developed by the German NaLiDa project (see http://www.sfs.uni-tuebingen.de/nalida/en/catalogue.html, (Barkey et al., 2011b)). First implementations of ERDO have already been used by CLARIN-D and the Collaborative Research Center 833 (see http://www.sfb833.uni-tuebingen.de/).

The remainder of this paper is structured as follows: Sect. 2. gives an overview of related archiving work. Sect 3. describes the archiving infrastructure with special focus on the system architecture, the ERDO workflow for transferring data from the researcher's desk to the repository and its user interface. Sect 4. discusses the repository policies. Finally, Sect. 5. concludes and gives an outlook on future work.

2. Background: Related Archiving Work

Linguistic resources have been archived for quite some time. The Max Planck Institute for Psycholinguistics in Nijmegen (MPI, see http://www.mpi.nl/) holds a terabyte archive of linguistic (and other) data. It has developed software such the IMDI editor (see http://www.lat-mpi.eu/tools/imdi/) and LAMUS (Language Archive Management and Upload System, see http://www.lat-mpi.eu/tools/lamus/) to describe, organize and upload data into the archive. Both appli-

cations are web-based and thus easily accessible to the user. A metadata editor called ARBIL (see http://www.lat-mpi.eu/tools/arbil/) has been implemented to better support the evolving CMDI-based metadata infrastructure. These tools, however, are targeted more at archivists or librarians rather than individual researchers.
The QSS Dataverse Network (see http://dvn.iq.harvard.edu/dvn/) of Harvard University allows researchers and institutions to create *dataverses* to hold and manage research data on servers that are "backed up in perpetuity by the Henry A. Murray Archive" (ibid). Dataverses' customization allows parties to retain control of all data so that "all the scholarly credit, web visibility, and access control for the data devolve to you, but all the work, preservation guarantees, and software and hardware upgrades and maintenance are taken care of by IQSS" (ibid).
Previous work in the project BW-eSci(T) (http://www.bwescit.uni-tuebingen.de/) provided insights for further developments on user needs in the archiving process (Hinrichs et al., 2010). This will be described in the following section.

3. An Infrastructure for Archiving and Managing Language Resources

In contrast to some existing archiving approaches, the infrastructure presented in this paper is targeted towards individual researchers without required knowledge in the field of digital preservation. The system's application only focuses on linguistic resources. An evaluation of other research areas was neither intended nor conducted. In the following, the various components of the infrastructure, which are represented in different layers, will be introduced in a theoretical part as well as their use in practice.

3.1. Fedora-Commons Repository

The Fedora-Commons repository software (see http://fedora-commons.org/) is the core component of the infrastructure for the sustainable management of linguistic resources serving as back-end for storing both data and metadata. The software is open-source, followed by a large community, and used by many institutions to manage different types of digital content.
The digital object model of the Fedora-Commons repository allows content items of any type or format to be bundled as *data streams* into a single digital object. The digital object's content data streams are complemented by several reserved data streams. The *Dublin Core (DC)* data stream holds metadata about the object; the *AUDIT* datastream automatically records all changes made to the object; the *POLICY* data stream encapsulates security restrictions. The *RELS-EXT* (i.e. external relationships) and *RELS-INT* (i.e. internal relationships) data streams allow users to describe relationships between digital objects. Each digital object is associated with a unique and persistent local identifier and an extra set of object properties that describe and help managing it within the repository. Access to the digital objects and data streams is given via authorization mechanims individually defined on the respective level. For Dublin Core metadata, the system supports the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH, (OAI-PMH, 2008)) natively, allowing the repository's administrator to distribute metadata of digital objects to the public. The full power of the Fedora-Commons' digital object model is deliberately not used, as it would increase the complexity of the user interface. For instance, the *RELS-INT* and *RELS-EXT* features to establish or represent arbitrary relations between the data streams of a digital object, or between the digital objects themselves, are not supported by the workflow wizard. Researchers are expected to make such relations explicit in the CMDI-based metadata file describing the digital object. For public access to the repository's (CMDI-based) metadata, the Fedora OAI provider service has been adapted accordingly.

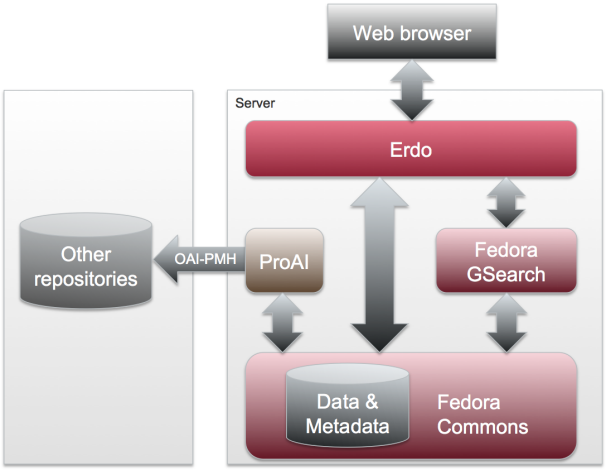

Figure 1: The ERDO System Architecture.

3.2. The Extensible Repository System of Digital Objects (ERDO)

On the basis of Fedora-Commons, the Extensible Repository System of Digital Objects has been developed for the sustainable archiving and management of language resources. The subsequent sections will focus on three fundamental aspects: the system's architecture itself, its underlying workflow and the workflow's representation in the user interface.

3.2.1. The System Architecture

ERDO's layered system architecture is shown in Figure 1. The core component, i.e. the already introduced Fedora-Commons repository software, serves as back-end for the purpose of storing both data and metadata. Moreover, it builds the foundation for three further layers: the component for distributing metadata (i.e. ProAI), the front-end (i.e. ERDO) and a search functionality (i.e. Fedora GSearch) used by ERDO.
Metadata is made available to other data repositories via the OAI-PMH protocol(short for: Open Archives Initiative Protocol for Metadata Harvesting, see http://www.openarchives.org/pmh/). ERDO uses ProAI (cf. http://proai.sourceforge.net/), a Java web application to enable the use of the protocol within Fedora-Commons. By supporting various kinds of metadata formats, ProAI extends the built-in OAI-PMH features, which

only disseminates Dublin Core, and allows the distribution of other metadata schemas such as CMDI. As OAI-PMH is the common protocol used for metadata harvesting by a large number of data repositories around the world, it is possible to share metadata on a large scale.
For accessing resources and creating new digital objects, the infrastructure uses ERDO as its front-end. This layer serves as the user interface that guides researchers through the process of archiving resources (i.e. creating digital objects) and exploring existing resources stored in the back-end for which the user has access rights. In the context of exploration, the ERDO interface also provides users with a search functionality using Fedora GSearch, a component of the Fedora Service Framework. ERDO is accessed via a web browser after authentication and authorization (i.e. user login). In order to make the access permissions easy to maintain, a rights management system based on individuals and groups is being used, comparable to UNIX systems.

3.2.2. Workflow

The workflow for the ingestion of linguistic resources into the repository system consists of three main phases that will be described in turn below.

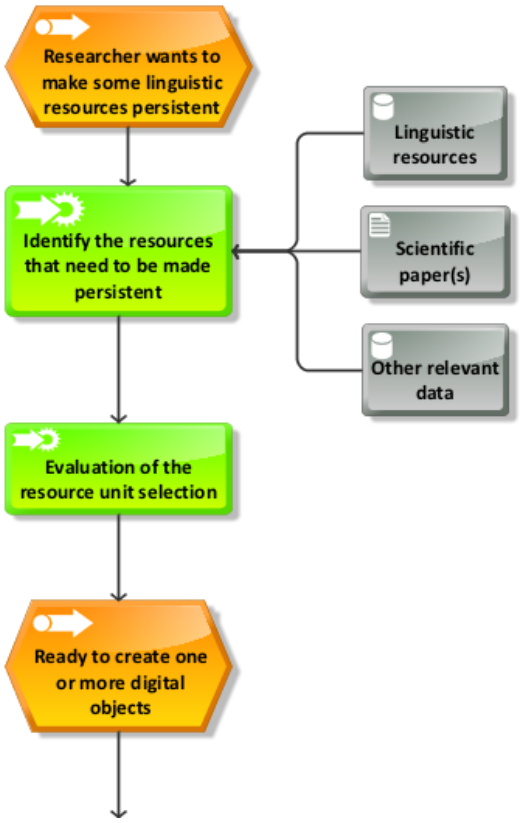

Figure 2: Workflow: Preparatory Phase (Phase 1).

Phase 1: Preparatory Phase Finding the right level of granularity for the organization of research data into digital objects is often difficult; the question of what constitutes *one* resource is all but trivial and usually does not have a unique answer. To address this issue, a set of recommendations is developed that helps researchers to make an informed decision. These recommendations are based on the size of a resource, possible units for distributing a resource, organizational questions, etc., which can lead to conflicting answers. The decision on the granularity of a resource needs to be taken within the first phase of the workflow (cf. Figure 2) before the actual data can be included in the repository by performing an upload (i.e. the second phase of the workflow). As a general rule of thumb, one resource should be one unit which is citable by a persistent identifier (PID, (Schroeder, 2009)) and which is independent of other resources in terms of "making sense" as one individual unit (see also ISO 24619:2011). Once a unit is identified, it is advised to distinguish its constituents into three distinct classes (see below).

Phase 2: Digital Object Creation Phase The second phase of the workflow concerns the creation of the digital object containing both primary data and metadata, also including administrative functions such as access restrictions. This phase is supported by the ERDO front-end.

Hierarchical association with the resource tree. At the beginning of the second phase of the workflow, the user must login to the ERDO repository system. Prior to the definition of the resource type, the location of the new digital object to be created in the repository needs to be specified. Since ERDO allows a hierarchical organization of digital objects, the digital object has to be created at the appropriate position in the resource tree. For instance, one way of organizing the resource tree is to mirror the organizational structure of the institution that hosts the repository. In the case of the University of Tübingen, the top node is associated with the university as a whole, the top node's children with the university's faculties. This structure is made explicit down to departmental and sub-departmental levels, and also includes temporary organizational units such as collaborative research centres and externally funded research projects.

Resource type selection and instantiation with data. To help researchers define the digital objects that will hold the linguistic resource, the workflow first enquires about the type of the resource in question (cf. Figure 3). Currently, the following resource types are anticipated, each being reflected by a metadata schema which should be used to describe the particular resource: text/speech corpus, lexical resource, experimental study, grammar, software tool, web service and images/audio/video recordings.
The material to be included in the repository is assigned to one of the following three categories:

- *research data:* the file contains research data;
- *documentation:* the file represents a scientific publication or technical documentation;
- *other data:* any other file.

The user is asked to upload these different types of files successively, resulting in a first draft of the digital object which is meant to be archived.
In general, it is left to the user to decide which kinds of data should be put into the repository, as the user's requirements may vary. Documentation, for example, could include preprinted articles or technical manuals. Some might

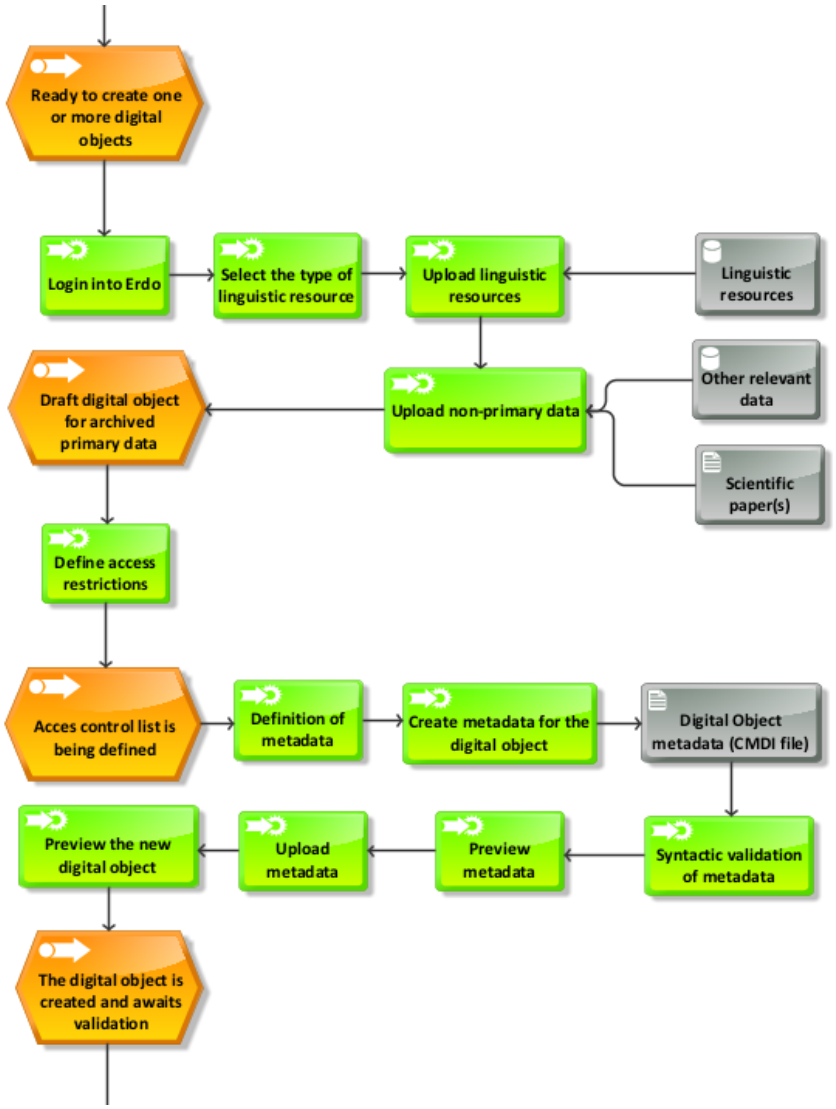

Figure 3: Workflow: Digital Object Creation Phase (Phase 2).

be essential for understanding the data (e.g. the code book for psycho-linguistic experiments) while others might be the result of the analysis (the publication). Each data stream can have its own access restriction.

Temporary files created during data processing, which are not to be analysed further, or back-up files should not be included. Such data files do not meet the requirements of an archiving system which is intended to preserve persistent versions of resources. This means that the resources' status should be final. For backup purposes of intermediate research results, the redundant storage of data on servers is advised.

Access restrictions. At this stage, the researcher specifies whether the status of the digital object holding the resource, or any part of it (the data streams), should be *public* (accessible by the general public), *private* (accessible to the resource creator only), or *group* (accessible to a defined group of researchers).

Metadata provision. Once the researcher has specified the type of resource as well as all resource-specific files, it is necessary to describe the resource. For this purpose, the Component Metadata Infrastructure (CMDI) has been chosen as the underlying metadata schema. CMDI is based on three fundamental concepts: *data categories* (i.e. *field descriptors* or *metadata fields*), *components* and *profiles* (i.e. *metadata schemas*). Data categories are defined and made persistently accessible by the ISOcat Data Category Registry (ISO 12620:2009, see `www.isocat.org`). These are then grouped as semantically similar units into components that are stored in the Component Registry (see `www.clarin.eu/ds/ComponentRegistry/#`). Finally, components form building blocks for profiles that, in turn, serve as templates for the description of different resource types.

For the provision of metadata itself, there are two options available. The first and preferred option is the use of a metadata editor (Dima et al., 2012). This editor will display a form that is based on the metadata schema for the given resource type. The resource type in question has previously been selected by the user (cf. the first action in phase 2). Some of the displayed information can automatically be filled in on the basis of existing information avail-

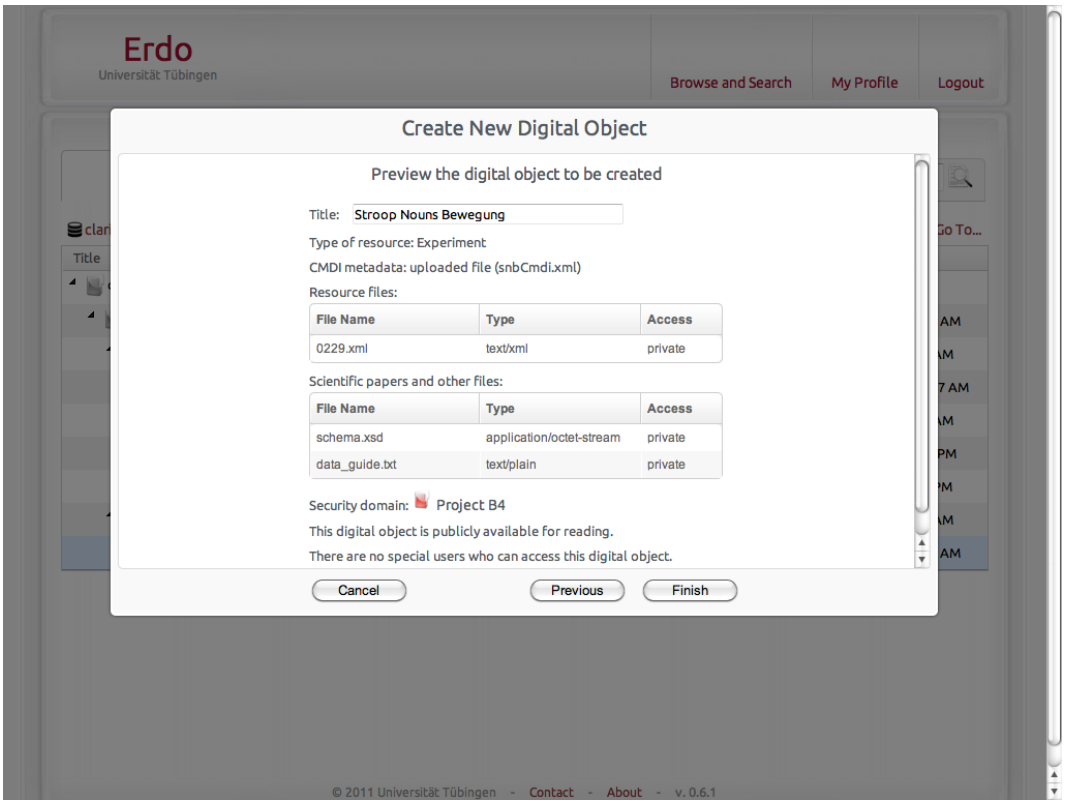

Figure 4: ERDO Preview.

able in ERDO, e.g., the researcher's name and affiliation as well as the references to the data streams (files) selected for the digital object. Further, the editor assists users in supplying information on all other metadata fields still requiring values. The use of the form based metadata editor is primarily intended to be used in cases for which metadata do not exist and thus need to be created from scratch. It also provides support for non-expert users as the editor does not require any (XML) knowledge in terms of metadata creation. In contrast, the second option for metadata creation in ERDO is via upload of already existing metadata files. This is, for example, the case for new versions of a resource for which the metadata is largely similar with some adjustments. These adjustments can be achieved by using the metadata editor or another form of editing.

For both cases, the next step within the workflow is the metadata's syntactic validation, which is performed in the background without being visible to the user. As a result, the user is provided with a preview of the resource description which applies both to uploaded and newly created metadata. At this stage, it is still possible to edit resource descriptions by going back in the workflow.

Data preview and submission. After the digital object has been fully defined, the user is provided with a preview before this part of the workflow is finalized. Here, it is still permitted to edit the digital object by returning to a previous step in the workflow. Figure 4 shows such a preview for a psycho-linguistic experiment. Accepting the preview is the last step performed by the researcher in archiving the digital object. Afterwards, only status updates of the archiving process are provided to the researcher, such as upload progress and PID assignment.

Phase 3: Validation and Archiving Phase The last phase of the workflow (cf. Figure 5) is concerned with the archive manager finalizing the archiving process and does not require any further involvement of the researcher providing the data. This phase starts when the archive manager receives a notification for a new submission. This needs to be checked in terms of completeness and correctness as well as in view of the compliance of the data's organization with the defined policies (see Section 4.). Completeness and correctness refer on the one hand to the metadata description, which is available to the archive manager. On the other hand, there are also a number of verification steps that the archive manager will run semi-automatically, such as ensuring that the uploading of all files did not lead to corrupt files. However, if it led to corrupt files, the archive manager is then able to contact the data provider and ask for elaboration. The tests performed can be on various levels, depending on access restrictions. If the data streams are readable for the archive manager, the tests can be more elaborate then if the archive manager is barred from reading the files, in which case only simple tests based on the size of a file, etc. can be performed.

If all tests pass, the archive manager accepts the submission and a persistent identifier as well as a time-stamp are assigned to the digital object. At this point, the resource is stored in the repository. Indices are updated to make the resource accessible and findable via full-text and metadata-based search through the resource's metadata file. The lat-

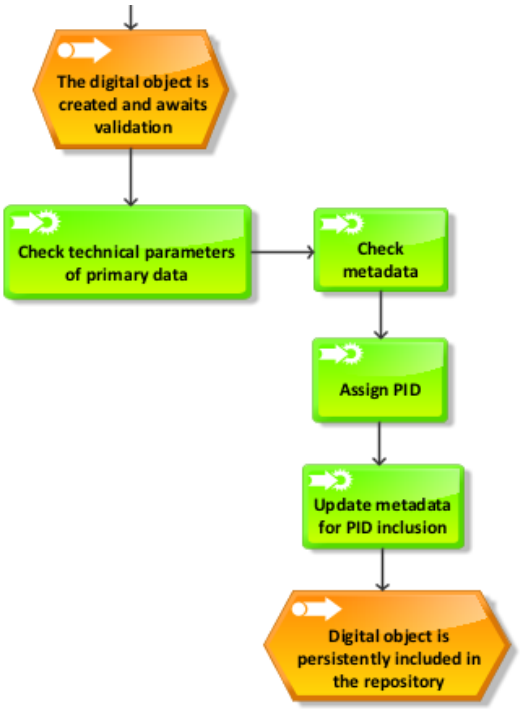

Figure 5: Workflow: Validation and Archiving Phase (Phase 3).

ter will also be open to OAI-PMH harvesting, access restrictions permitting.

This third phase of the workflow is facilitated by the ERDO-Browser, but in part also depends on external services such as PID assignment.

It is important to keep in mind that the workflow described in this section applies only to those types of resources that have previously been uploaded to the repository system. If new resource types are to be ingested into the system this may require intervention by the archive manager to assist the user and possibly also modifications to the workflow described above.

4. Discussion

From the archivist's and the data provider's points of view, considerable tension may arise especially regarding the policies of writing and updating resources, which also need to be regulated by appropriate processes. Archivists and data users rely on the principle that data stored in a repository will remain available permanently and will not be changed (Strathmann, 2009). For this reason, modifications to the data are under normal circumstances not permitted, and the deletion of digital objects should be impossible. Such data persistence is inter alia a necessary prerequisite for data provenance, the practise of documenting scientific results in such a way that their creation process can be reproduced by a third party.

For the present repository, deleting and updating are possible only in exceptional cases when there are strong external forces, such as legal court orders to withdraw the data, breech of copyright restrictions or blatant violations of an individual's right to privacy. Such exceptional cases require the archive administrator's intervention and are therefore not part of the regular workflow wizard.

As digital objects can be uniquely referenced by a persistent identifier, the present approach adopts the policies used in the publishing world and by ISO 24619:2011. The policies include, among others, that a digital object's metadata can be changed whenever necessary, as, for instance, by updating contact details. The same applies to data streams of type *documentation*, giving researchers the option of updating an existing publication (e.g. from submitted manuscript to camera-ready copy) or adding a new publication related to the research data. While researchers are also allowed to edit data streams of type *other data* (or add them to an existing digital object), they are not permitted to modify data streams of type *research data*. When primary data needs to be changed, i.e. modified or deleted, it is advocated to generate a new digital object, and thus a new persistent identifier in accordance with the rules on persistent identifiers. In the described cases, the archive manager can help researchers to update their digital object or to define a new version.

5. Conclusion and Future Work

The presented infrastructure for managing linguistic resources is intended to support researchers as the creators and holders of all research data. Thus, the deployed tools are different from those that are oriented more towards archivists or metadata specialists. The ERDO front-end guides researchers through the entire process of archiving language resources. Thereby, it neither requires experience in archiving technologies nor (XML) knowledge in terms of metadata creation. ERDO also addresses possible privacy concerns by users by integrating configuration options for access restrictions.

Future work includes a better integration of the metadata editor into the workflow wizard, and extensive user tests elaborating the results of first evaluations by researchers using an alpha version of ERDO. The infrastructure's further development will be conducted within the Tübingen CLARIN-D project.

6. Acknowledgements

The infrastructure presented in this paper is based on previous work in the project BW-eSci(T) (`http://www.bwescit.uni-tuebingen.de/`) and is jointly funded by the CLARIN-D (Common Language Resources and Technology Infrastructure, `http://clarin-d.net/index.php/en/`) grant of the BMBF (Federal Ministry of Education and Research, `http://www.bmbf.de/en`), the SFB 833 (Collaborative Research Center 833: The Construction of Meaning - the Dynamics and Adaptivity of Linguistic Structures, `http://www.sfb833.uni-tuebingen.de/`) funding by the DFG (German Research Foundation, `http://www.dfg.de/en`) and the NaLiDa project (Sustainability of Linguistic Data, `http://www.sfs.uni-tuebingen.de/nalida/en/`) grant by the DFG in the LIS program (Scientific Library Services and Information Systems, `http://www.`

dfg.de/en/research_funding/programmes/infrastructure/lis/index.html).

7. References

R. Barkey, E. Hinrichs, C. Hoppermann, T. Trippel, and C. Zinn. 2011a. Komponenten-basierte Metadatenschemata und Facetten-basierte Suche: Ein flexibler und universeller Ansatz. In *Internationales Symposium der Informationswissenschaft (ISI 2011)*, Hildesheim. Universität Hildesheim.

R. Barkey, E. Hinrichs, C. Hoppermann, T. Trippel, and C. Zinn. 2011b. Trailblaizing through forests of resources in linguistics. In *Proceedings of Digital Humanties (DH), June 19-22, Stanford: Stanford University*, Stanford. Stanford University.

D. Broeder, M. Kemps-Snijders, D. Van Uytvanck, M. Windhouwer, P. Withers, P. Wittenburg, and C. Zinn. 2010. A data category registry- and component-based metadata framework. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*, Malta.

D. Broeder, D. van Uytvanck, M. Gavrilidou, and T. Trippel. 2012. Standardizing a component metadata infrastructure. In *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012)*, Istanbul.

E. Dima, E. Hinrichs, C. Hoppermann, T. Trippel, and C. Zinn. 2012. A metadata editor to support the description of linguistic resources. In *Proceedings of the 8th Conference on International Language Resources and Evaluation (LREC 2012)*, Istanbul.

E. Hinrichs, V. Henrich, and T. Zastrow. 2010. Sustainability of linguistic data and analysis in the context of a collaborative escience environment. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010)*, Malta.

ISO 24619:2011. 2011. Language resource management – persistent identification and sustainable access (PISA). Technical report, ISO.

OAI-PMH. 2008. The open archives initiative protocol for metadata harvesting. Technical report, OAI. Protocol Version 2.0 of 2002-06-14, document version 2008-12-07, http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm.

K. Schroeder. 2009. Persistent Identifier (PI) - ein Überblick. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, and K. Huth, editors, *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*, chapter 9.4. Verlag Werner Hülsbusch, 2. edition.

S. Strathmann. 2009. Einführung. In H. Neuroth, A. Oßwald, R. Scheffel, S. Strathmann, and K. Huth, editors, *nestor Handbuch: Eine kleine Enzyklopädie der digitalen Langzeitarchivierung*, chapter 3. Rahmenbedingungen für die LZA digitaler Objekte. Verlag Werner Hülsbusch, 2. edition.