One Core Preservation System for All your Data. No Exceptions!

Marco Klindt
Zuse Institute Berlin (ZIB)
Takustr. 7, 14195 Berlin
Germany
klindt@zib.de

Kilian Amrhein
Zuse Institute Berlin (ZIB)
Takustr. 7, 14195 Berlin
Germany
amrhein@zib.de

ABSTRACT

In this paper, we describe an OAIS aligned data model and architectural design that enables us to archive digital information with a single core preservation workflow. The data model allows for normalization of metadata from widely varied domains to ingest and manage the submitted information utilizing only one generalized toolchain and be able to create access platforms that are tailored to designated data consumer communities. The design of the preservation system is not dependent on its components to continue to exist over its lifetime, as we anticipate changes both of technology and environment. The initial implementation depends mainly on the open-source tools Archivematica, Fedora/Islandora, and iRODS.

General Terms

Institutional opportunities and challenges; Infrastructure opportunities and challenges; Preservation strategies and workflows; Innovative practice.

Keywords

Contexts of preservation, data model for management and access, preservation strategies, infrastructure, Archivematica, iRODS, Fedora/Islandora.

1. INTRODUCTION

Digital data and information is not only ubiquitous but also more and more the foundation for research, education, and dissemination of cultural heritage. A lot of effort is put towards digitization of cultural artefacts in galleries, libraries, archives, and museums (GLAM), but the lack of institutional resources to keep these substantial investments not only safe but usable for future generations raises the demand for preservation services significantly. Even though the core business of these institutions is to preserve their physical holdings and collections, some of them are unable to provide and maintain adequate custody/stewardship of *digital* objects.

For that reason, we identified a need for a long-term digital preservation archival information system (DPS) that supports cultural heritage and research institutions. They can use it as a service that offers the possibility to safeguard their digital artefacts without putting many resources towards implementing the rather complex requirements for best-practice digital preservation themselves. The preservation system we describe is intended to be trustworthy in the sense that it is transparent in its functionality, documentation, and policy and is also aligned to the Open Archival Information System (OAIS) functional model as standardized in ISO 14721:2012[9]. Clients of black box preservation systems may experience difficulties in risk-assessing the underlying processes and services, and scheming exit strategies. For a more thorough explanation of concepts and terminology refer to the OAIS magenta book[7] or Lavoie's Introductory Guide[4].

This paper will first establish the background and requirements, then give a description of and reasoning for the architectural design, followed by details about the implementation and tools chosen, discussing the implementation details in regard to the design and conclude with future work and discussion.

1.1 Background

The Zuse Institute Berlin (ZIB) is a research institute for applied mathematics and computer science and operates a regional super computing center that requires facilities to store more than five petabytes of data on disk and nearly a hundred petabytes on tape. Data needed for and generated by super computer runs are expensive and therefore great effort is made to ensure reliable data retention. Our working group utilizes the available infrastructure to design and build an archival information system.

The DPS should offer deposit, curation, and preservation services for any kind of data that data producers (either inhouse or external) would want to keep safe-guarded. A data-agnostic view should enable us to utilize a single core workflow system for digital objects from various domains. The system combines existing free and open-source components including Archivematica[12], iRODS[3], and Islandora[5] to vertically integrate the existing infrastructure.

We also provide a generic access layer to the submitted data for administrative and preservation watch purposes subject to access control mechanisms. The architecture is designed

to support future access mechanisms for external users.

Data from cultural heritage institutions is perhaps the most universally approachable data, because it is suitable for reuse not only in scientific work but is also of wide interest for educational or creative purposes. In contrast to research data in the natural and life sciences, where data often comprises unique numeric data sets only usable in very specialized domains, and the humanities, which often is concerned with textual data, the standardization of metadata descriptions for cultural heritage objects is paramount for discovery, comparability and reuse in the semantic web. We try to accommodate data deposits from all LAM institutions by accepting *Metadata Object Description Schema*[8] (MODS) for libraries, *Encoded Archival Description*[11] (EAD) for deposits from archives, and *Lightweight Information Describing Objects*[2] (LIDO) for museum object descriptions as preferred formats.

1.2 Requirements

In addition to the special cases of cultural heritage the system should be able to handle data from all fields in research and education and to help them to maintain the viability of the vast corpus of digital materials either already amassed or in the process of being generated.

The main requirement is to maintain deposited information as self-contained and self-describing *archival information packages* (AIP) using the preservation metadata dictionary as described by PREMIS[10]. In an ideal world AIPs should not depend on the existence of any component of DPS itself and therefore enable the exchange of AIPs within a collaborative federation of OAISs as described in [13]. Unfortunately this is not the case; interoperability in terms of AIP exchange between systems is a complex problem which remains to be solved.

The system should furthermore use or adapt existing and available open source tools and open standards and by doing so benefit from community best-practices and advancements. Implementing a DPS from scratch and keeping up with research is not feasible.

1.3 Review of existing solutions

Existing, commercially available preservation systems do not fully meet our workflow requirements due to lack of access to the source code and availability of publicly accessible documentation, or mainly depend on cloud infrastructure. Open-source OAIS aligned systems seemed too complex to easily change components without losing functionality. Integrating different components into one system offers the opportunity for clear responsibilities, audit, and documentation and therefore trustworthiness.

2. ARCHITECTURAL DESIGN

Because of our requirements to be aligned to OAIS and have self-contained AIPs for interoperability, we designed the architecture to be a single, modular core pipeline of existing tools linked up by in-house developed data conduits.

The system is therefore composed of loosely coupled components with strictly defined responsibilities. As we anticipate the software components to become obsolete during the lifetime of the system, this modularity enables us to find or develop tools to substitute waning functionalities.

Loose coupling also means that redundancies with regard to data and metadata are necessary to achieve the goal of independent, functional exchangeable modules. The self-contained, homogenous information packages do also support a more streamlined, automatical migration of archived content into other DPSs as an exit strategy.

2.1 Aiming at moving targets

The whole system design and implementation is regarded as a living system that has to be cared for and adapted to a changing environment and requirements. We try to achieve this goal by utilizing a single core preservation workflow for all information packages to reduce complexity and increase sustainability. The DPS as a whole consists of the stages *preingest* (see 2.2), *ingest* (2.3), *management* (2.5), and *access* (2.6). Only the first and last stage (deposit and access) will be customized to meet the requirements of different types of data, the core stages (ingest and management) will process the data from all producers the same way. An overview of the system architecture is shown in figure 1.

A consistent and well-defined data model is a fundamental prerequisite that allows for not only treating all submitted data with a single workflow but also, more importantly, to supersede tools in the future. Our data model distinguishes between *preservation description information* (PDI) and *descriptive information* (DI) as described by the OAIS. The PDI in our case also contains descriptive and administrative information about the data producer such as identifiers, contract numbers, contacts, and more. See figure 2 for an overview of the data model.

2.2 Preingest

The ingest functional entity described in the OAIS model is split into two phases within our architecture: deposit and ingest. The deposit phase covers the preingest process, i.e. the preparation and transfer of data as a submission information package (SIP) into a quarantine staging area. The ingest phase covers the preparation of the archival information package (AIP).

The DPS does require data to be organized and described in a certain way in order to treat the deposit's content agnosticly independent from origin and purpose. The deposit workflow ensures completeness of administrative metadata and transforms the data formats into the data model expected by the ingest process. The original data is preserved to maintain authenticity.

Information packages preserve only the included information. To preserve the ability to independently understand the contained information requires that the data can be rendered as information by the consumers. Therefore the producer has the responsibility to ensure that the data is suitable for re-use by a designated community. The responsibility of the DPS is to maintain that renderability of the data. With research data sets this often conflicts with the need for widely supported representation information, i.e. file formats, that are easily rendered with standard soft-

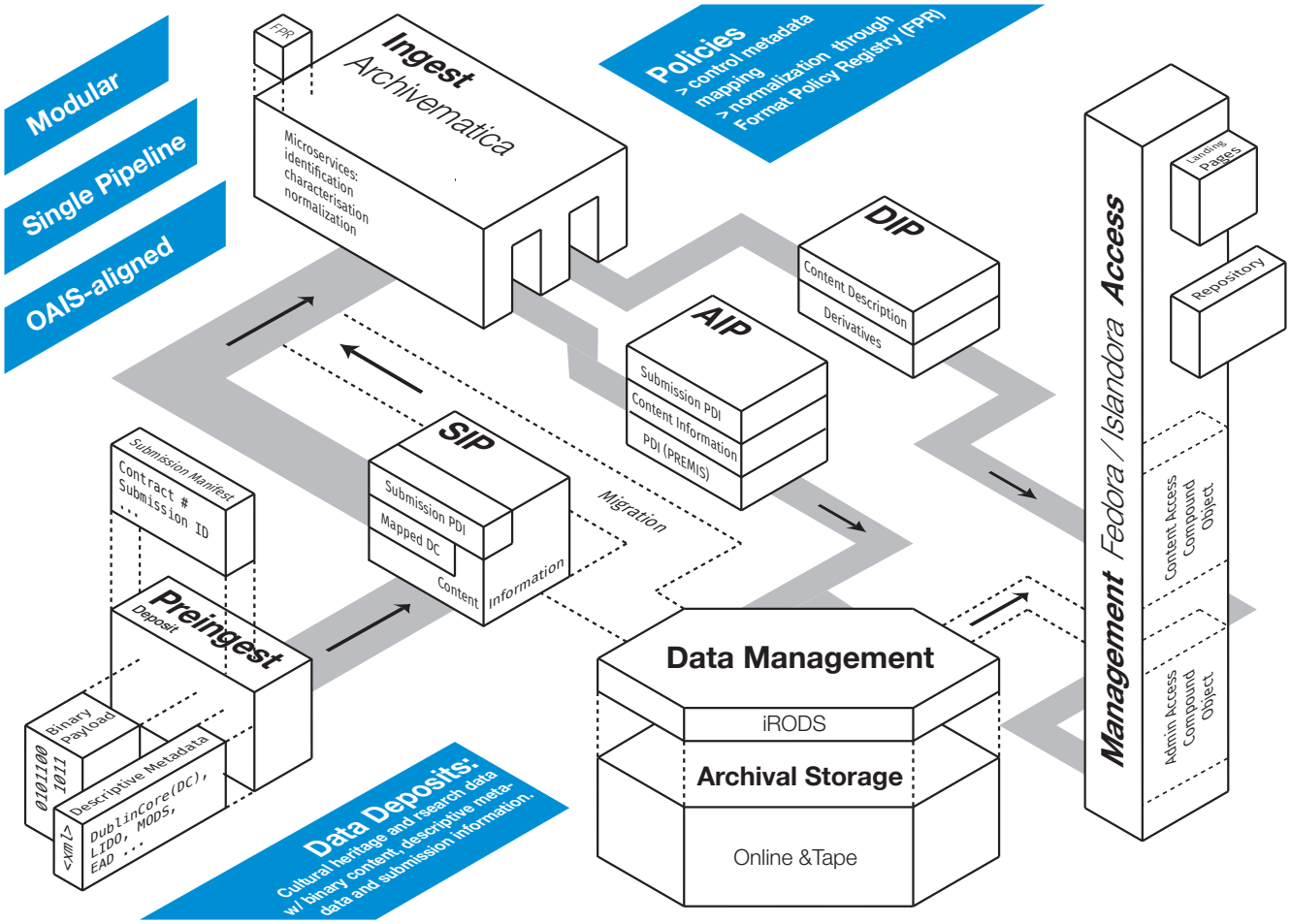

Figure 1: Digital Preservation System Architecture Overview

ware. To support the preservation of any data regardless of format, our preservation system differentiates only two levels of preservation: passive and active. Data perceived to be at the passive preservation level will be preserved at the bit-level in addition to retaining structure and metadata, and the DPS promises a best effort to describe the contained data. Data perceived to be at the active preservation level requires the best effort of a depositor to comply with policy-published archival formats and the DPS therefore promises to ensure the renderability through migration.

The system will accept any digital material for deposit but rejects any submission for ingestion that does not satisfy the submission agreements. Preservation activities will differ based on the assessment of a preservation level on ingest. The perceived level of preservation is however not a static one, but is an outcome of re-examination of the supplied or extracted technical metadata. It can change from passive to active preservation as a result of an updated format policy following actionable observations during preservation watch.

Context information, which might be useful for understanding the deposited material in the future but is not part of the information object and therefore does not have to be considered for preservation actions, can be declared to be submission documentation. This data will be captured in the AIP but is not included in the DIP.

During the deposit phase all necessary metadata is gathered that is needed for managing and accessing the data within the archive.

2.2.1 *Data Deposit Registration*

Prior to depositing the content information itself, the producer must initiate a data deposit session by requesting a submission identifier through a web portal. The submission identifier will act as the reference for the data to be deposited and will be used to attach the transfer data to the deposit agreement negotiated between the producer and our archive. The deposit agreement itself comprises a legal contract for transfer of custody including responsibilities of the producer and the DPS, and a technical policy (the submission agreement) describing workflows, procedures, and actions based on the types of data objects. Upon registration the producer will have the choice either to input mandatory and optional preservation description information (PDI) via a web form or by selecting to include said information as a submission manifest with the data transfer. The PDI is necessary for managing the data by providing information about relation to contracts and submission agreements.

Descriptive information (DI) in Qualified Dublin Core can be entered or included as CSV[1] or as XML for Qualified Dublin Core in the data transfer. If the descriptive metadata is included as either *Encoded Archival Description* (EAD) for deposits from archives, *Lightweight Information Describing Objects* (LIDO) for museums, or MODS for libraries, the necessary fields will be automatically extracted and mapped to DC during the restructuring stage.

The submission session can be interrupted and resumed at any time to allow for thorough preparation of data objects and content information. If for any reason the producer decides to abandon the session altogether he can also terminate it and discard entered and uploaded data.

2.2.2 Data Transfer

The data objects can now be uploaded to the staging area which is provided alongside the submission identifier. The data can either be uploaded via a web browser, transferred to the staging area by other network protocols or via sneakernet on external hard disk drives, USB thumbdrives or optical media. To ensure integrity and completeness during transfer the data must reside either in ZIP archive containers, be put into a BagIt structure or be referenced in a METS files with checksum information in the file section.

2.2.3 Submitting and Compliance Testing

In order to conclude the submission session, the producer must initiate a compliance test by clicking the submit button. This checks for completeness of the required preservation description information and the presence of descriptive information. If ambiguous data is detected during the automatic metadata extraction, the submission is considered not to be compliant.

Furthermore the integrity of all data objects is checked by testing the zip containers, validating the BagIt integrity or checking the uploaded files against the file section of the METS files.

If the submitted data deposit successfully passes these tests, the data is accepted and transferred to the restructuring step.

2.2.4 Restructuring

The quarantined data is restructured into either a single or multiple submission packages. Some producers with poor data management facilities or legacy applications choose to bulk export data sets and upload corresponding content information and rely on the archive to bundle the appropriate files and metadata into information packages. The rules for breaking up bulk deposits into multiple information packages are specified in the submission agreement.

The descriptive information, either as entered in the registration process or extracted from metadata files, and the preservation description information are transformed into a format suitable for ingest. The original metadata files are treated as data objects and bundled with the remaining data objects. METS files that were used for transfer and do not contain descriptive information are put into the submission documentation area.

After restructuring all SIPs are of equal structure and ready to be ingested through a single ingest workflow.

2.3 Ingest

The ingest phase creates an identifier for the AIP and assigns identifiers to all data objects for reference within the AIP. The ingest workflow identifies common file formats and extracts necessary technical information. Non-archival file formats are normalized if the delivered content is identified to be in a set of known formats for which format policies exist. Only content already in a set of archival formats or conforming normalized versions can be actively preserved. Content not identified in ingest will only be preserved passively, i.e. at the bit-level. This enables the archive managers to easily identify the need for migration preservation actions and their planning.

PDI including rights statements, DI, logical and physical structure of the SIP, fixity information, PREMIS events of identification and normalization is captured into a single METS file that will be the authoritive source of information about the AIP for managing the archive. All files are compiled into a BagIt structure that is saved as a single archive file to allow for easy transfer within the data management layer.

Access or dissemination copies of either the normalized data objects or derivatives are also created during ingest and transferred along with DI and PDI to the combined management and access repository.

A submission report will be sent to the contact person (producer) by email. The data deposited is now under stewardship of the archive.

2.4 Archival Storage

The AIPs are transferred to archival storage by a data management middleware layer. The middleware abstracts from the physical resources and is responsible for not only storing AIPs as multiple replications to on-site and potential off-site locations but also to retrieve the physical AIPs independent of residence. A subset of the PDI is attached as administrative metadata to the AIPs such as the type of information package, the identifier, submission and contract identifiers and fixity information. Aside from using this metadata in the management of the archive, it can also be used as the database for generating reports on storage usage or item count reports.

2.5 Management

The higher-level data management operations (see 3.4) are not based on information stored in the data management middleware, but based on the PDI and DI stored in a repository alongside the dissemination copies of the content information. There are two main roles for accessing the information contained in the DPS: data managers and data consumers. Data managers are entrusted with keeping data usable. Data consumers want to search, discover, and retrieve information. To address the different needs of these

[1] Comma-separated values. See IETF RFC 4180.

roles the repository provides two different views to the same AIP realized as separate entities within the repository object store: an administrative entity providing the PDI and a descriptive entity providing the DI. Management activities like monitoring, reporting or performing preservation actions often require selection of AIPs through information contained in the PDI, which is accessible through the administrative view. The DI entity is responsible for access by data consumers.

2.5.1 *Preservation actions: migration*

Following a change of policy regarding a certain file format, the PREMIS records are checked for occurrences of that format and the corresponding AIPs are selected for migration. The change in format policy affects the normalization step in the ingest workflow for all subsequent submissions. Migration is performed by re-ingesting selected AIPs as SIPs into the ingest workflow preserving the identifiers and amending the PREMIS event trail.

2.6 Access

The access repository as mentioned contains not only representational copies of the data objects but also the corresponding DI and PDI. Access to information in the PDI and the retrieval of AIPs residing in archival retention is restricted to data managers. Data consumers on the other hand have different requirements: the finding of information inside the archive based on the descriptive information provided in DC and a representation of the content information suitable for their needs. The access to DIPs is restricted based on the access rights information from the PDI. To provide different designated communities or other users a higher level of precision and recall for retrieval, subsets of DIPs can be transferred to access systems or repositories that provide a more specialized integration and understanding of the original descriptive metadata schema instead of generic Dublin Core.

If the access to a data package is assigned an open license, a Digital Object Identifier (DOI) can be provided for persistent access to research data sets at the DIP level.

3. IMPLEMENTATION

We chose to make use of existing and freely available tools as much as is feasible to achieve the architectural design goals and to keep in-code customization of existing tools and in-house development to a minimum. The transfer and the access stages are obviously the most challenging because of our goal to accept data deposit from various domains. After restructuring the delivered information packages and mapping of metadata, a *single* pipeline is used for preservation and management actions.

The normalized deposits are processed by a single Archivematica pipeline for AIP generation, handed off to an iRODS data management grid for archival storage and transfer, and the PDI and a subset of DI and CI (i.e. derivatives where appropriate) are ingested in a Fedora object storage, from where they are accessible through an Islandora front end for both management, discovery, and retrieval functionality.

The implementation is guided by the overall architectural

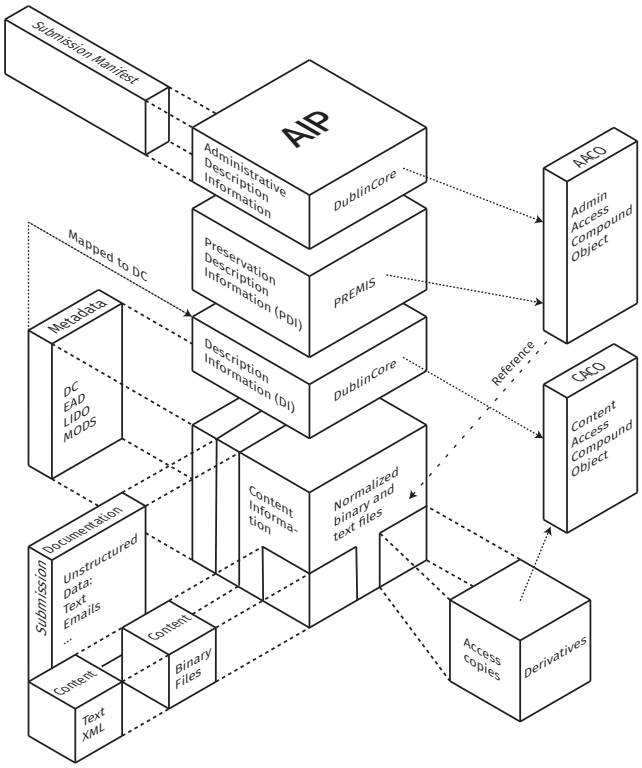

Figure 2: Data Model and AIP

design and adapts to lessons learned during the development. Some stages are not completely functional yet.

3.1 Preingest

While the self-deposit of data will be available in the future via a web portal as described above, most submitted data arrives currently by external media or secure file transfer (sftp) or copy (scp). The integrity is checked by verifying a bag structure or zip archive created by the producer. The data resides in a quarantine storage area (and temporary backup) and is available for the data extraction and restructuring tools.

The administrative PDI metadata (an excerpt is shown in table 1) for the submission is stored in a submission manifest file in YAML[2], which is human and machine readable. Based on the field `MetadataFormat`, a customized python script is selected to extract descriptive DI metadata for content discovery from `MetadataFile`. Supported metadata schemas are EAD, LIDO, MODS, and qualified DC-XML. Extraction of DC from a METS container will be supported in the future. The scripts also check for the completeness of the payload that is referenced in the metadata file.

After having successfully gathered the PDI and the DI, all files will be sorted into either the object or the submission documentation directories of a single or multiple SIP depending on rules in the submission agreement. The script

[2]YAML is a human-friendly data serialization standard. See `http://yaml.org/`.

Field name	Dublin Core Mapping
SubmittingOrganization	dcterms:rightsholder
OrganizationIdentifier	dcterms:publisher
ContractNummer	dcterms:identifier
Contact	dcterms:creator
...	
AccessRights	dcterms:accessrights
License	dcterms:license
...	
Metadatafile	not mapped
Metadataformat	not mapped

Table 1: Submission manifest and mapping

also writes metadata to special files that are processed by the ingest workflow. The subsequent stages will operate on these now equally structured SIPs.

3.2 Ingest

The ingest phase will transform the submitted data and metadata, extracted technical metadata and the documentation of the actions taken into an AIP. After extensive market research and testing we found that Archivematica is a viable tool that meets our requirements quite well. Archivematica is an open-source application that combines various open-source tools into a distributed workflow pipeline to process digital objects into AIPs following the OAIS functional model.

3.2.1 Archivematica

The ingest phase will be executed by a *single* Archivematica pipeline. The processing workflow can be controlled with a web interface or through REST calls. Creating entirely new workflows or modifying the existing is cumbersome at the moment because it is stored in a database with heavy use of referencing. The existing workflow contains workflow decision points that can be influenced by presets, and these will be applied to all deposits. The preconfigured choices can be overridden by embedding a preset file into the transfer data directory. This allows, for example, to store AIPs in distinct storage locations if negotiated in a submission agreement.

Archivematica generates a METS structure to capture the references to files and file structure of the ingest, to attach metadata to individual files, and to document rights and executed preservation actions and their result in PREMIS events. Metadata is attached to files in the submission by creating a metadata CSV file that Archivematica then inserts into the respective descriptive metadata section (DMDSec) in the METS. We wanted to keep changes to the codebase of Archivematica to a minimum but also have control over metadata that end up in the generated METS to support the PDI from the submission manifest. The METS standard schema does not support the description of the METS file itself, so we use the METS generation script in Archivematica to attach the administrative data to the object directory level as a convention for the metadata to survive the ingest. As we have control over the ordering of the DMDSecs we also use the second DMDSec to store the DI about the submission. The structure of the resulting AIP is shown in figure 2.

Archivematica also generates DIPs that are not handed off to archival storage but are used to ingest data into the access and management repository (see 3.4.1).

3.3 Archival Storage

For archival storage we use the on-site data storage facilities available at ZIB. Persistent storage consists of a hierarchical Storage Archive Manager (SAM) that augments an online file system transparently with nearline redundant tape storage. ZFS is used for the online filesystem that is designed to prevent data corruption caused by bit-rot. Nearline access to tape storage means offline data is available in less than 30 seconds on average. The data is stored redundantly on two StorageTek 8500 tape libraries with currently installed tape capacity of around 100 petabyte, of which nearly 400 terabyte (800 terabyte with redundancy) are reserved for our archival system. The libraries have no physical connection to prevent tapes from being destroyed by tape recorder malfunctions. They also use automatic fixity checks and error correction to ensure data integrity. The tape libraries are installed in a special vault that is waterproof, can withstand an outside fire for around 10 hours, and has an additional CO_2-fire extinguisher system.

Archivematica supports different storage backends through the use of separate storage service application that abstract various services like local and NFS file systems, LOCKSS, Duraspace, and others. As we use iRODS (integrated Rule-Oriented Data System) to store and replicate digital objects, we extended the Archivematica storage service to expose an iRODS storage space which not only stores AIPs but also attaches administrative metadata to help with discovery and retrieval.

3.3.1 iRODS

iRODS is a distributed data-management system for creating data grids and persistent archives. It provides access to data objects organized in collection trees called zones with granular access control. Data in a zone can be accessed by authenticated users regardless of where the data is physically stored. Integrated rules manage replication to physical storage resources transparently to the user and can also act on user-supplied metadata attached to the data objects. Such replication is also possible to remote, off-site storage for geographical redundancy. The iRODS grid supports integrating storage resources and user bases of different organizations and thus can be used for federated archiving.

iRODS tiered resources are responsible for replicating AIPs to the SAM and back into online storage. iRODS maintains checksums for all AIPs that are used for fixity checks if AIPs reside online. Online storage is more expensive than tape so rules are implemented to trim redundant copies of AIPs if a threshold of disk usage is reached.

Although federated data replication is also possible with LOCKSS, by using iRODS we maintain more control over data movement, residence and replication.

3.4 Management

High-level management of the data in our DPS consists of monitoring data integrity, triggering preservation actions,

and providing access to AIPs and DIPs. Preservation actions that migrate file formats are not yet implemented but one of the future releases of Archivematica will add a feature that will allow us to re-ingest AIPs back through the Archivematica pipeline. This allows for changing metadata in already ingested AIPs without changing the identifiers including amendment of the PREMIS trail. Extending the feature to re-normalize file formats for which the format policy has changed in Archivematica is planned for the future.

3.4.1 Fedora/Islandora

We ingest the DIPs generated by Archivematica into a Fedora Object Store and use Islandora as a front end to the repository for management actions. The DIP contains the same information in METS as the stored AIP including the PREMIS data and derivative representation (access copies) of the binary payload (where appropriate), and a transformation has been defined and has been carried out by the ingest stage.

We represent a single DIP with *two* Fedora compound objects: one for administrative and management purposes and another for content discovery and access purposes. The DIP METS is parsed on ingest and transformed into multiple Fedora METS files that are ingested as multiple Fedora objects: one METS for each binary payload file or access copy, and one METS for each of the two compound objects as parents. One of these parent objects will be ingested as an *admin access compound object* (AACO) and comprises all data streams contained in the DIP. The AACO gives access to the submission manifest data as main descriptive Dublin Core and refers to the PREMIS data and payload derivatives. It is used for administrative tasks involving contracts and deposits. Additionally the payload derivatives and the DI, i.e. the description of the digital object, are accessible through another object called *content access compound object* (CACO). This is used to discover and retrieve objects based on the mapped, generic DC metadata of the datasets. These two different "views" of a DIP try to separate administrative tasks like report generation of stored file formats or calculating the amount of data stored for each contract or year, and finding objects by their actual content while keeping the datastreams in a unified repository. The AIP is referenced through the Islandora front end by the identifier of the AIP and can be retrieved from the storage layer through the AACO view. The iRODS integration with Islandora and the data model is described in more detail in [1].

The DIPs managed by Fedora are stored in a filesystem that is backed up separately from archival storage. They can, however, be re-generated any time from the AIPs.

3.5 Access

The repository is currently accessible only for internal administrative users. Access to the AIPs for data producers is realized with management actions by the administrative staff to stage data in a location where it can be picked up; we provide no self service at the moment. With changing requirements of our clients we might have to implement an access repository for bulk self-service AIP exports in the future.

The descriptive information exposed through the repository is limited due to its DC-only design and therefore not adaptable to the different metadata descriptions from the various domains, we plan to support, and is therefore not suitable for discovery and reuse of the data within those domains. This is a consequence of the need for normalization of descriptive metadata for utilizing a single preservation system.

We do provide DIP presentation access for selected data sets not from within the internal access repository but by generating landing pages or handing data off to external content management systems (CMS) or repositories.

3.5.1 Landingpages

For some clients who have no means of providing access to their data themselves, we offer a service for generating customized, static landingpages for each AIP as a low-maintenance way to present them on the Web. The customization includes converting the used metadata schema to HTML templates and populating them with metadata and binary data from the DIPs. Discovery can be provided by generating a digital object identifier (DOI) for reference in publications or uploading metadata to specialized search portals, e.g. Europeana[3] or the Deutsche Digitale Bibliothek[4] (DDB, German Digital Library) for cultural heritage data. As the landingpages are static, they can easily be migrated to other hosting services as they are independent from our infrastructure.

4. FUTURE WORK

New data producers who want to use our infrastructure might require different protocols for data deposits. We will investigate the suitability of SWORDv2 (Simple Web-service Offering Repository Deposit)[6] or OAI-PMH[5] for data from institutional repositories or the S3[6] data cloud protocol for large research data sets.

Collaboration with other archives will be tested by taking over their AIPs by transforming their structure in our deposit stage and ingesting them into our pipeline. Other archives can also ingest our AIPs in their archives because the information packages are self-contained and do not depend on data residing in databases. Other archives could also choose to use only the ingest stages or utilize our archival storage through the federation and replication features in iRODS.

4.1 Object repositories

For clients who require the data to be stored in and accessed through a data repository, we may potentially offer Islandora instances that can be customized to their designated user communities and corporate designs. The CACO and its relevant child objects would then also be transferred to these repositories after ingest.

For clients with existing CMS or repositories but without the infrastructure or resources for digital preservation we

[3] `http://www.europeana.eu/portal/`
[4] `http://www.deutsche-digitale-bibliothek.de/`
[5] Open Archives Initiative Protocol for Metadata Harvesting
[6] Amazon Simple Storage Service API, a protocol implemented in various technology stacks.

plan to hand-off the DIPs into their CMS or repositories so that they have an overview about the deposited data. These repositories could be used in the future to generate AIP delivery orders.

The original (not mapped) metadata description schema contained in the AIP could also be used for the development of sophisticated platforms to explore and discover the data because it takes advantage of the inherent complex data models they are based upon.

5. DISCUSSION AND CONCLUSION

Our data model for metadata supports digital long-term preservation within a single core workflow for ingest and management activities and allows for consistent description of widely varied content and a clear separation of PDI and DI. The resulting AIPs accommodate both submission requirements from multiple data producers as well as accommodate discovery opportunities for data consumers in addition to the information needed for administration, access control, preservation watch (using the PREMIS information), reporting, and billing for the management entity.

The core idea of AIP construction is to treat the AIPs as well-documented, atomic information objects that contain the full intellectual information about the preserved objects without external references. This obviously permits a simple exit strategy in case the DPS ceases to exist in the future.

The described model, design, and implementation may not be suitable for everyone. However, we hope that it enables us to offer preservation services to a whole range of different data producers because it reduces complexity of the infrastructure and therefore helps manageability and sustainability of the whole system. The modular architecture allows us to substitute software building blocks as reaction to technical issues related to software obsolescence. At the same time it deals with the intrinsic complexity and variety of data and contained information that has to be preserved in order not to deprive future users of possibilities and opportunities.

6. ACKNOWLEDGMENTS

This work has been supported by the Senate Chancellery Cultural Affairs of the State Berlin. The authors would like to thank Tim Hasler, Wolfgang Peters-Kottig, and Elias Oltmanns for valuable input and discussions.

7. REFERENCES

[1] K. Amrhein and M. Klindt. Islandora as an access system for iRODS managed information packages. Presented at the 10th International Conference on Open Repositories, 2015.

[2] E. Coburn, R. Light, G. McKenna, R. Stein, and A. Vitzthun. LIDO - Lightweight Information Describing Objects Version 1.0. Technical report, ICOM-CIDOC Working Group Data Harvesting and Interchange, November 2010.

[3] B. Fortner, S. Ahalt, J. Coposky, K. Fecho, A. Krishnamurthy, R. Moore, A. Rajasekar, C. Schmitt, and W. Schroeder. Control Your Data: iRODS, the integrated Rule-Oriented Data System. White paper, RENCI, University of North Carolina at Chapel Hill, 2014.

[4] B. Lavoie. The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition). *DPC Technology Watch Series*, Oct 2014.

[5] M. A. Leggott. Islandora: a Drupal/Fedora Repository System. In *4th International Conference on Open Repositories*, May 2009.

[6] S. Lewis, P. de Castro, and R. Jones. SWORD: Facilitating Deposit Scenarios. *D-Lib Magazine*, 18, January/February 2012.

[7] Reference Model for an Open Archival Information System (OAIS). Hosted at `public.ccsds.org/publications/archive/650x0m2.pdf`, 2012.

[8] Metadata Object Description Schema (MODS). Hosted at `http://www.loc.gov/standards/mods/`. Retrieved September 2014.

[9] ISO 14721:2012: Space Data and Information Transfer Systems – Open Archival Information System (OAIS) - Reference. Retrievable via `http://www.iso.org/iso/catalogue_ics`, 2012.

[10] Preservation Metadata: Implementation Strategies (PREMIS). Hosted at `http://www.loc.gov/standards/premis/`. Retrieved March 2014.

[11] B. Stockting. Time to Settle Down? EAD Encoding Principles in the Access to Archives Programme (A2a) and the Research Libraries Group's Best Practice Guidelines. *Journal of Archival Organization*, 2(3):7–23, July 2004.

[12] P. Van Garderen. Archivematica: Using micro-services and open-source software to deliver a comprehensive digital curation solution. In *Proceedings of the 7th International Conference on Preservation of Digital Objects, Vienna, Austria*, 2010.

[13] E. Zierau and N. Y. Mcgovern. Supporting Analysis and Audit of Collaborative OAIS's by use of an Outer OAIS – Inner OAIS (OO-IO) Model. In *Proceedings of the 11th International Conference on Digital Preservation, Melbourne, Australia*, October 2014.