https://doi.org/10.2352/issn.2168-3204.2017.1.0.74

Developing ARCLib – An Open-Source Solution for a Bit-Level and Logical Long-Term Preservation

Andrea Miranda; The Czech Academy of Sciences; Prague, Czech Republic

Zdenek Hruska; Moravian Library; Brno, Czech Republic

Abstract

This poster informs about the Czech ARCLib project. One of the main goals of the project is the development of an open-source solution for a bit-level and logical preservation of digital documents, respecting the national and international standards as well as the needs of all types of libraries in the Czech Republic. The mission of the ARCLib project lies, among others, in creating a solution that will allow institutions to implement all of the OAIS functional modules and entities, considering institutions' information model. The architecture is planned as open and modular and the final product will be able to ingest, validate and store data from a majority of software products used for creating, disseminating and archiving libraries' digital and digitised data in the Czech Republic.

Motivation

ARCLib project follows a series of activities of the Czech libraries from the last fifteen years. Motivation to deal with long-term preservation mainly came to the Czech libraries with digitised data. Individual libraries have gradually begun to run their own digital libraries and archive the scanned data. The National Library of the Czech Republic created the National Digital Library (NDK) project, combining digitisation, long-term preservation of existing and new data with an adequate access to them.

Another motivation for the project is connected to Strategy of Czech Libraries for the years 2011-2015. In its part dealing with long-term archiving, requirement for testing freely available solutions had been made. As a result of this, LTP Pilot (LTP — Long-Term Preservation) project was undertaken. Its goal was a pilot implementation and testing of a low-barrier Archivematica-based system for long-term preservation of digital data, including its possible integration into the infrastructure of CESNET [1] digital storage. The project also wanted to motivate Czech libraries to increase their level of data preservation.

Problem

Even though libraries in the Czech Republic have official Czech translations of the ISO 14721 [2] and ISO 16363 [3] standards available, a clear implementation methodology concerning the Czech environment, used systems and formats is missing. In terms of the initial conditions for successful long-term preservation in the Czech Republic, the existence of a digitisation standard NDK format can be considered as a great move. Since its definition, it is followed in all major digitisation projects and in many smaller activities. It is necessary to ensure for these data a bit-level preservation (preventing physical data loss, alteration or corruption of digital files and media) as well as logical preservation (protection against adverse effects of changes, obsolescence of information technologies and data formats).

Nevertheless, easily attainable software solutions for archiving are not available. The digital preservation issue was up until recently an exclusive domain of large institutions such as national libraries and national archives, which had the necessary mandates, finances and human resources. These institutions typically focus on developing complex customized solutions, built often on commercial systems. For most Czech libraries, complex and rather expensive LTP solutions are out of their reach. Advances in theory and practice of digital preservation, along with the growing need to address the long-term archiving of digital data in smaller institutions led to realizing that even with limited resources one can start creating their own solution using freely available software (e.g. POWRR [4]).

Since the concept of open source is relatively widespread and previously undertaken projects brought promising results, it was quite clear the coming project would continue on this path.

Approach

ARCLib was inspired by some of the projects developing systems for long-term preservation of digital data, which for this purpose also use open-source software (e.g. Archivematica [5], iRODS [6], Islandora [7] or RODA [8]). For the needs of the project, the open-source system Archivematica would probably be one of the first choices, since it is dynamically developed and implemented in many projects around the world. However, it does not cover all the OAIS functional entities [9]. Since it lacks a well elaborated data management and focuses on critical archiving functions only (transfer, ingest, create SIP/AIP/DIP [10]), data curators are missing not only a powerful system for long-term preservation, but also a tool for an effective administration of archival data.

ARCLib project is being financed from the applied research support programme of the Ministry of Culture of the Czech Republic [11].

It pursues four main goals:

- developing a complex LTP solution ARCLib for logical as well as bit-level data preservation
- creating methodology for a logical preservation of digital documents for the specific Czech environment in respect to the international standards (ISO 14721 and ISO 16363 in particular)
- creating methodology, solution for storage of large amounts of data and ensuring their bit-level preservation
- verifying functionality of the entire solution in practice in the form of pilot system in at least one of the participating institutions

The final product is planned to be open source, free to download and use for any library, accompanied with the documentation and set of guides for easy implementation and use.

Schedule

The ARClib project has run since 2016 and will continue until 2020. Afterwards development and future extensions are more than desirable but it is out of scope of the original project.

Work schedule is as follows:

- Beginning of the ARCLib project (2016)
  - first analysis and system designs
  - collecting testing data
- Phase One: Design (2017-2018)
  - module's prototypes
  - programming
  - first versions of all modules
- Phase Two: Development and testing (2018-2019)
  - full versions of all modules
  - intensive debugging
  - performance tests
  - crash and recovery tests
  - finishing documentation for system administrators and other users
- Phase Three: Initial deployment
  - pilot installation for tests in one of the participating institutions
  - installation of pilot system in one of the participating institutions
  - performance tests on real data samples
  - debugging and updating documentation
- Phase Four: Future development (2020+)
  - final analysis and recommendation for future development of ARCLib system (Figure 1)
  - the end of ARCLib project

ARCLib description

Ingest. ARCLib does not include a pre-ingest or deposit module. Conversion of ingested data into a SIP is not presumed. In case of having raw or scanned data and metadata (MarcXML or Dublin Core), it is expected to make use of external systems (ProArc, Archivematica, DSpace, own scripts) for creating SIP in a required structure and choosing the right profile of ingested data. ARCLib Ingest requires import of data in a later stage of their processing meaning they already are in a form of a SIP defined by a standard of systems like ProArc or Archivematica, wrapped in BagIt. SIP entering ARCLib needs to have content information (data object together with representation information) as well as preservation description information. ARCLib Ingest adds to an entering SIP newly generated ARCLib AIP XML with information on validation and extracts few metadata from a SIP (ingest date, ID of data producers, original location and ID of SIP, fixity, format identification results, etc.). Ingest workflow also records the process in a database and generates XML and ARCLib AIP along with the SIP BagIt (in .tar format), and UUID.

ArcLib AIPs will consist of two main components: original SIP wrapped in newly created AIP XML METS (with PREMIS). In order any newly added metadata and/or metadata from the SIP BagIt can be indexed, they will become a part of the AIP XML METS. This XML will be an integral part of the ARCLib system. However, SIP stays intact, no changes will be made during the Ingest phase – there will be no normalization or creating user copies.

Data Management. ARCLib Data Management contains information on AIPs (or rather their parts) stored in the Archival storage. It provides an index and a search interface, ideally with reporting (over the stored AIPs and reporting on processing). From the Data Management search interface an event of DIP export can be initiated (for the time being DIP equals AIP) and ARCLib AIP XML can be viewed individually or in bulk. It will be possible to search descriptive, administrative and technical metadata created in ARCLib AIP XML. Queries can be combined, it is possible to search at a specific metadata element, saved as (logical) sets of results and used for data export to a working area (bulk export – request for the Access module). It is possible to edit ARCLib AIP XML or upload a newer version of XML that will replace the old one. In that case, a new version will be stored as ARCLib AIP XML_v2 (validated against the profile). This assumes locking the package in a database while writing the new version. This operation is available only to users with defined roles. Ideally, data management provides all services through the API, including the ability to edit ARCLib AIP XML and the creation of the new version.

Administration. ARCLib Administration module allows configure workflow for Ingest and it also controls state of system architecture. It secures communication with databases, controls availability of storage capacity, secures administration procedures like cache cleaning, work storage cleaning, checks after restarts and crashes. For purposes of workflow configuration it manages a set of different databases:

- register of SIP profiles and data providers,
- register of workflow's steps and scripts for Ingest,
- register of profiles for SIP validation,
- format register for identification and validation,
- storage location register – work and permanent storage pools.

Administration module contains register of users and their roles, so different rights can be set to individual users. ARCLib philosophy is to have only limited set of roles – Administrator (system and workflow configuration, troubleshooting), Analyst (problem solving – e.g. invalid metadata during validation), Editor (editing AIP metadata) and Data provider (starts new Ingest). All roles may be combined, so one user can have multiple roles in the system.

Archival Storage. Archival Storage represents a complex service for a bit-level preservation, allowing data replication into disparate geographic locations and use of several data storage technologies. Archival Storage offers object storage available through a simple REpresentational State Transfer (REST) interface. This interface inside the Archival Storage module also offers (among others) functions for a controlled access to data. Towards the storage, this module stores data in a defined structure on a distributed file system, into Ceph [12] object storage or any other suitable technology that enables secure storage and handling of data.

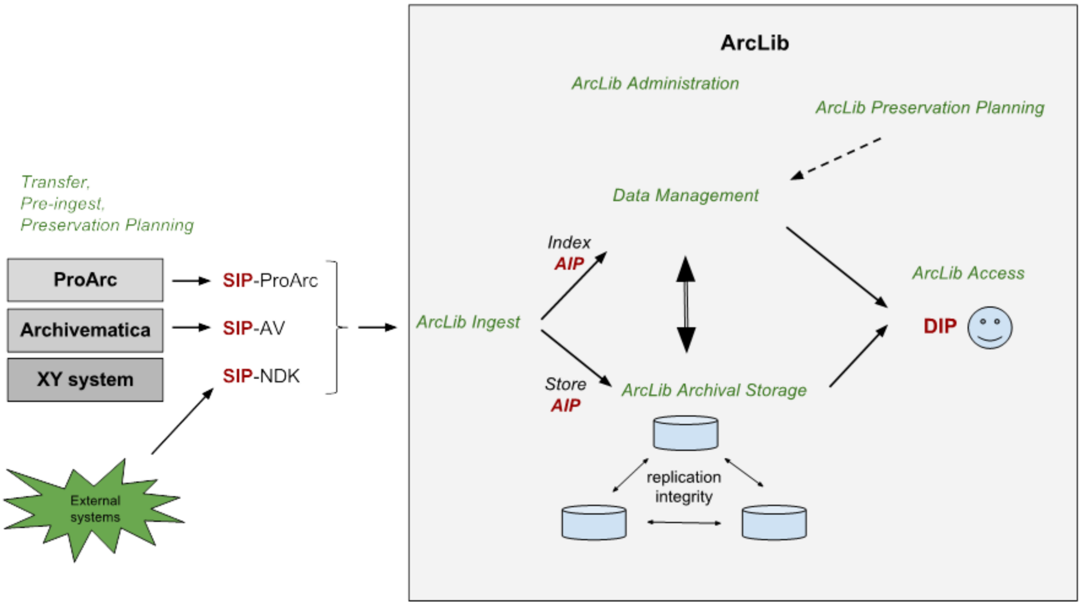

***Figure 1** - ARCLib System Schema*

Storage service procedures and functions are for administrators clearly transparent in a logical part of the ARCLib system (Data management, Ingest or Administration), meaning it offers storage and a package issuance functions, but it does not burden with implementation details like employed storage or numbers of copies stored. Archival Storage has its own database of stored objects and its own administrative interface.

Preservation Planning. A great deal of the preservation planning functions is done outside the ARCLib system. Defining and monitoring of a designated community as well as technology monitoring are functions of institutions focused on science and research, therefore their performance is in interest of many communities. In the Czech Republic the main role in the field of standardisation in libraries is carried out by the National Library of the Czech Republic. Part of its activities are performed at specific departments of the library and administrators of the ARCLib system are encouraged to follow their recommendations and published standards.

Each owner of an ARCLib system should have their own defined preservation strategies and policies based on long-term preservation standards. The aim of the ARCLib system is not in providing a testbed for data migration, however tools and system's options should allow re-ingest of packages edited in external systems, an effective data management and a trustworthy ingest processing. For running the ARCLib system as a trustworthy system for a long-term preservation, a methodology created in the ARCLib project will be followed.

Access. The philosophy of an ARCLib system presumes that users need to get the ingested data back in the original state. Access enables export of AIP as DIP, where AIP and SIP relationship is 1:1. Other potential processing for an access to end-users or AIP's content update (Kramerius conversion [13], metadata editing, AIP's content and structure change, repeated format validation or a new extraction of technical metadata) is made in other systems as ProArc or Archivematica.

ARCLib is a back-end application and is not aimed at end-users. It does not enforce a restricting policy to AIPs. Access rights metadata are part of the submitted SIP and checked at ingest, but not transformed into ARCLib AIPs.

Results

The new ARCLib solution will meet requirements of the OAIS functional and information model (preserving AIPs' information content along with all metadata) and offer tools supporting all the OAIS functional entities, including the preservation planning one. A users' community will maintain the knowledge base necessary for skilled decisions to preserve the information content (a format, rules, services, migration paths and tools database) and perform functions required by the OAIS' preservation planning.

Projected ARCLib architecture is open and modular in a way that it is possible to replace individual components (database, index, functional modules, micro-services, Business Process Management (BPM), tools for processing formats, etc.) without compromising the system as a whole, should it be needed anytime in the future.

ARCLib will be compatible with a commercial solution of the NDK (at the National Library) and enable transfer of AIPs between instances of this newly developed system and the NDK, and vice versa. With respect to the OAIS model, there is a possibility of creating a network of cooperating OAIS archives for exchanging packages, where DIP from one system should serve as a SIP to another system and with the order reversed. A two-way data exchange and interoperability with a commercial LTP (NDK) will significantly increase the level of security of preserved data in the Czech Republic. It can also fulfill the requirement of the National Library on the existence of an exit strategy.

Conclusions

The solution will be able to store data in the structures and formats that are already used in the libraries of the Czech Republic (NDK, older digitisation standards of sound, map collections, theses, academic publications, etc.). The architecture is open and modular and the final solution will be able to store data from all widely used software products for creating, disseminating and archiving libraries' digital and digitised data in the Czech Republic. It will support data and knowledge exchange, make use of previously created LTP-tools (e.g. Archivematica) without duplication of already developed parts. It will also support modular implementation of individual components for other institutions, not necessarily only in the Czech Republic.

References

1. CESNET - an association of universities of the Czech Republic and the Czech Academy of Sciences. It provides national e-infrastructure for science, research and education.

2. Anon., 2012. ISO 14721:2012: Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model, Geneva: ISO.

3. Anon., 2012. ISO 16363:2012: Space data and information transfer systems -- Audit and certification of trustworthy digital repositories, Geneva: ISO.

4. SCHUMACHER, J., 2014. From Theory to Action: Good Enough Digital Preservation for Under-Resourced Cultural Heritage Institutions: A Digital POWRR White Paper for the Institute of Museum and Library Services, 2014. Available at: http://hdl.handle.net/10843/13610.

5. Anon., 2016. Archivematica. Available at: https://www.archivematica.org/en/ [Accessed November 16, 2016].

6. Anon., 2016. iRODS – The Integrated Rule-Oriented Data System. Available at: http://irods.org/about/overview/ [Accessed November 16, 2016].

7. Anon., 2016. Islandora. Available at: http://islandora.org [Accessed November 16, 2016].

8. Anon., 2012. RODA. Available at: http://www.roda-community.org [Accessed November 16, 2016].

9. Anon., 2011. UML Activity Diagrams. Archivematica wiki. Available at: https://wiki.archivematica.org/UML_Activity_Diagrams [Accessed November 16, 2016].

10. SIP/AIP/DIP – information packages defined in the ISO 14721:2012.

11. Anon., 2013. NAKI II – Program na podporu aplikovaného výzkumu a experimentálního vývoje národní a kulturní identity na léta 2016 až 2022. Available at: https://www.mkcr.cz/program-na-podporu-aplikovaneho-vyzkumu-a-vyvoje-narodni-a-kulturni-identity-na-leta-2016-az-2022-naki-ii-857.html [Accessed February 21, 2017].

12. Anon., 2017. Ceph storage. Available at: https://ceph.com [Accessed February 21, 2017].

13. Kramerius is an open-source (GNU GPL license) Czech digital library project. There is over thirty individual Kramerius installations in Czech libraries. Source code is available at: https://github.com/ceskaexpedice/kramerius

Authors' Biography

Andrea Miranda received her PhD. in LIS from the Charles University (2014). Since 2006 she has worked in the Computer Center and Central Library at the Charles University. Her work has focused on digital repositories, their audit and certification, long-term preservation of digital data and digitisation.

Zdenek Hruska has worked in the Digitisation Department of the Moravian Library since 2014. He was involved in the LTP-Pilot project and the National Digital Library (NDK) project. He is interested in digitisation, digital repositories and long-term digital preservation.