11th Conference on Grey Literature and Repositories: proceedings [online]. Prague: National Library of Technology, 2018. ISSN 2336-5021. Available from: https://nusl.techlib.cz/en/conference/conference-proceedings

ARCLIB – LTP SOLUTION FOR LIBRARIES

Eliška Pavlásková

eliska.pavlaskova@ruk.cuni.cz

Library of the Czech Academy of Sciences

Zdeněk Vašek

zdenek.vasek@ruk.cuni.cz

Library of the Czech Academy of Sciences

Abstract

The presentation introduces project ARCLib. The project aims to create complex open source Long Term Preservation solution for libraries. ARCLib ensures long term preservation of digital data according OAIS guidelines and provides a free alternative to commercial software solutions. ARCLib is designed as a solution for all types of memory institutions – museums, galleries and archives. As part of the project two methodical guidelines were created – Methodology for logical preservation of digital data and Methodology for bit preservation.

Keywords

Long term preservation, open source, ARCLib, OAIS

This article was created within the research programme „Programme of applied research and experimental development of national and cultural identity 2016 – 2020 (NAKI II)“, for the „ARCLib Project – Designing complex Solution for Long-term Preservation in Czech Libraries“, project identification DG16P02R044.

Introduction

Work on the ARCLib project has been ongoing since 2016. The project responds to the need for memory institutions and, in particular, libraries to ensure the long term preservation of digital documents. There are several commercial and open-source solutions on the market at present which cover the issue of LTP (Long Term Preservation) to a greater or lesser extent. Of course there is no open solution within the Czech environment that would comprehensively cover the needs of libraries (and other types of memory institutions) relating to the long term storage of digital objects and, at the same time, enjoy a broader community of users. These are mainly regional and specialised libraries that need to ensure the long term preservation of such documents, which libraries have been working with for some time now and which are gaining in importance. Such institutions, however, do not have the funds to be able to bring in the bigger team required to implement a more robust open source solution and, simultaneously, frequently want to maintain the flexibility which an open source provides. It is precisely for these institutions that the results of the ARCLib project are intended.

The principal planned outcome is a newly-created open source tool for the long term preservation of digital documents prepared on the basis of OAIS and experience of other tools in use in the Czech environment. In addition to its own software tool, methodical materials will be created and subsequently made public in the form of open access. The use of a software tool and methodologies will make it possible to protect, over the long term, digital data and institutions that cannot call on a large team of specialised workers.

The first year of the solution involved the preparation of detailed technical and procedural specifications of tender dossiers for the development of the system. InQool, the company which won the tender, took on the task of actual development. A prototype now exists and individual functional requirements are being tested on this. Information about the project is available at **https://arclib.cz/**.

The objectives and development of a project solution

The objective of the project is to create a comprehensive LTP solution known as ARCLib on an open source basis that uses freely-available tools and systems. One part of the project, and at the same time a significant product of the project, is the creation of a methodology for the logical, long-term preservation of digital data that takes into consideration international standards in this area (reference model OAIS – ČSN ISO 14721 and ČSN ISO 16363 standards) and systems used to create digital data at Czech libraries and makes these accessible.[1] A methodology and solutions for the physical storage of data and the assurance of bit-level preservation will be prepared at the same time.

The whole project is the result of the changing situation in the sphere of Czech libraries (see, for example, Hutař and Melichar, 2014). The long-term management and preservation of digital documents (digital born and digitised) is becoming less and less a specialised matter that only those interested devote their attention to. It is necessary to ensure both "bit-level preservation" of data (safeguarding against physical loss, alteration or crashes involving digital files and carriers) and, at the same time, logical protection (safeguarding against the negative

[1] HUTAŘ, J., A. MIRANDA, E. PAVLÁSKOVÁ, Z. VAŠEK and Z. HRUŠKA. *Metodika logické ochrany digitálních dat.* 2018. Available from: **http://hdl.handle.net/11104/0282107**

impacts of changes and the ageing of information technologies and data formats on the availability and usability of digital information).

The ever-increasing spectrum of libraries and other institutions in the Czech Republic now has to preserve digital documents. Nonetheless, long term preservation and logical preservation in line with the concept of OAIS remain a costly business. First of all there is the need to acquire a software tool that enables storing and management of a large number of documents. At present, libraries are offered the option of acquiring a commercial solution or of using the progressive development of the Archivematica open source tool or other freely-available tools, although these invariably require extensive adaptation to meet the needs of a specific institution and the development of new parts. Commercial solutions are at present frequently based on a cloud basis, which is contrary to the standard policy of memory institutions in this area. A different path was chosen for the ARCLib project.[2] The aim is to develop an open-source LTP solution that is able to provide the required functionalities to ensure long term preservation within the environment of Czech libraries (whilst respecting all the international standards which are common in the community). While this is not a closed tool, the preservation possibilities are restricted to a pre-set group of data types (in light of the open-source nature of the tool, however, it is possible to broaden the set of data types as part of onward development). The aim of the new tool is to provide support for data preservation within the standards which are currently used at Czech libraries. These are primarily the standards in place at Národní digitální knihovna (National Digital Library), the Kramerius digital library, the ProArc production system and the repositories of DSpace. The ARCLib solution makes it possible to ensure the long-term protection of digital data at libraries of varying size and will be a freely-available alternative to accompany commercial solutions, the application of which in Central Europe is more common at large institutions such as national libraries and national archives.

Meanwhile, restriction to such major institutions is not solely based on the cost of acquiring the software tool involved. Indeed such costs are actually falling. An essential prerequisite for putting one's own policy of long term preservation into place is having sufficiently broad human resources, whereby the software tool is merely a necessary tool within a comprehensive LTP solution. The process of planning and subsequent operation of a trustworthy digital repository brings with it personnel and financial costs - for a clear description see, for example, Rosenthal (2009). Large teams remain the domain of large institutions and we cannot expect, even within the medium term, that smaller organisations such as regional libraries will have a comprehensive team of experts that would be able to cover the full range of issues involved, from the operation of hardware, through the management of content from the perspective of logical protection of stored data, to specialists that plan future steps in relation to long term preservation. Neither should we forget regular audits of the systems required for LTP, the evaluation of these and the preparation of documents for evaluation. Future users of the ARCLib solution will also have to fulfil the demands placed on evaluation. The implementation team is aware that regional and specialised libraries will not have the large team described at their disposal, but will need to ensure long term preservation all the same. As far as commercial products are concerned, they will be able to draw on the support of the supplier, but will also have to respect the policies issued by national institutions. As part of the ARCLib project, standard technical and user documentation is now accompanied by two

[2] Inspiration taken from, for example, the POWRR – Preserving Digital Objects With Restricted Resources project, conceived in a similar way - **http://commons.lib.niu.edu/handle/10843/13610**.

methodologies that provide users with sufficient knowledge of how to use the system in the right way and execute operation to ensure logical preservation with its assistance.

The methodology for the logical preservation of digital data was created first and was this year certified by the Ministry of Culture of the Czech Republic. The methodology describes the whole concept of the proposed LTP solution and explains the individual functions which make up the whole and, based on these, presents users with detailed instructions on how to use the procedures which the tool makes possible in ensuring the long term preservation of digital documents. The methodology describes in detail the structure of an archival information package and fundamental metadata sections, and the information which the system itself generates is explained here (for example, about validations, the method of version control, etc.). Instructions are also found in the methodology on how to assess the risks of stored data, how to prepare an institution that uses ARCLib for basic certification, etc. The project implementers also suppose that, once the system has been expanded, a certain community of users will develop and collectively maintain the knowledge base required for qualified decisions in the long term preservation of information content within the developed system - this should involve decision-making and recommendations of how to approach the database of formats, rules and services provided, decision-making on format migrations and chosen tools - and execute the functions required by the OAIS standard in the sphere of preservation planning. The second of the planned methodologies was also created this year, i.e. the methodology for bit-level preservation of digital data, which on the contrary focused on ways of safeguarding the "physical" preservation and cohesion of stored data using the methods described in the ARCLib tool. This methodology was submitted for certification in September 2018. It is envisaged that both of these methodologies will be regularly updated in the future to take into account changes in the tool itself and the procedures recommended within the international community.

The aim of the project is to develop a software tool and extremely detailed recommendations on how to use it so that it is also possible to use these methodological documents to carry out the basic tasks involved in long term preservation at smaller institutions with a limited number of workers. However, the scope of both methodologies goes beyond demarcation for system users alone. The general sections are the first attempt at normatively summarising recommendations and good practice for the processes of long term preservation of digital data at libraries in the Czech Republic. They can be used to plan activities for other LTP systems or by digital data producers and such as they are established with regard to the need for their long-term protection in the future. The universal applicability of the methodologies is also based on the involvement of all relevant participants that share in the digitisation and management of digital documents at libraries. Knihovna Akademie věd ČR, v.v.i. (Library of the Czech Academy of Sciences), Národní knihovna ČR (National Library of the Czech Republic), Moravská zemská knihovna (Moravian Library) and Masarykova univerzita (Masaryk University) are all participating in the project. The involvement of the National Library of the Czech Republic guarantees interoperability with the standards of the National Digital Library. Such interoperability will not only be at a general level: it will also be ensured that the AIP created according to National Digital Library regulations can also be preserved in ARCLib. Cooperation between both archiving solutions will significantly increase the level of protection of stored digital data. The openness of the solution, after minor modifications and developments of the system, particularly in the sphere of data schema makes it possible to engage other memory institutions or other producers of digital born documents.

The ARCLib tool is designed for the management and preservation of digitised and digital born documents. Procedures for the processing of both types have been prepared. The prerequisite is that access for a specific producer and its data format is invariably regulated. It is conditional on adherence to the master format for the storage of metadata, which in the case of ARCLib is the METS standard. ARCLib is not a production system and is not capable of creating submission information packages: it merely preserves them and creates archive alternatives from them.

Description of the system

ARCLib is a system for the logical and bit-level preservation of digital data that has been designed in line with the requirements inferred from the ČSN ISO 14721 (OAIS) standard. It uses tools that are already in existence, such as ProArc and Archivematica, to the maximum possible extent, mainly for the creation of SIP packages. It validates the prepared SIPs, converts them to archival packages (AIP) and preserves them in accordance with OAIS. When identifying individual modules, this paper is based on the actual naming of modules established during the definition of functional requirements and is still maintained. *Figure 1* illustrates the clearly-arranged schema of the ARCLib system and its modules and its relationship to the outside world. The modules of the system correspond to the modules specified in the ČSN ISO 14721 (OAIS) standard: however, they are adapted to meet the character of the system, i.e. to the fact that this is a dark archive that is not intended for end users and the fact that the system envisages the input of SIP already having been processed in other systems.

ARCLib does not have means of displaying archived data (image servers, browsers, etc.). ARCLib is used by the managers of archive digital data and data is used, following export, by dissemination systems and, where appropriate, by other digital library systems (DAM systems). The updating of AIP and the generation of new versions of AIP proceeds in large part through data editing in external systems (ProArc, DSpace) and subsequent re-ingest in ARCLib.

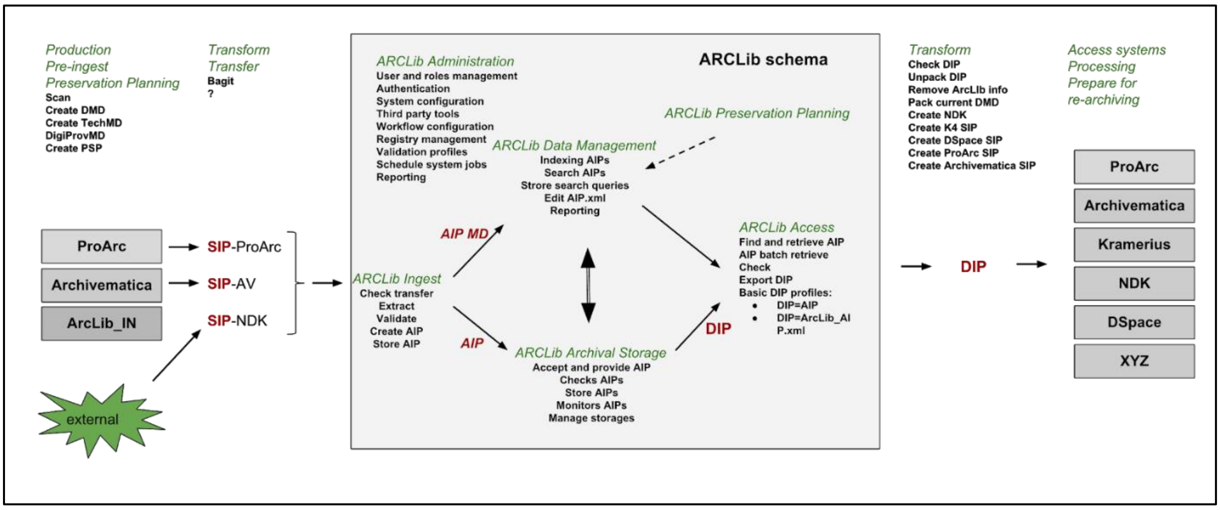


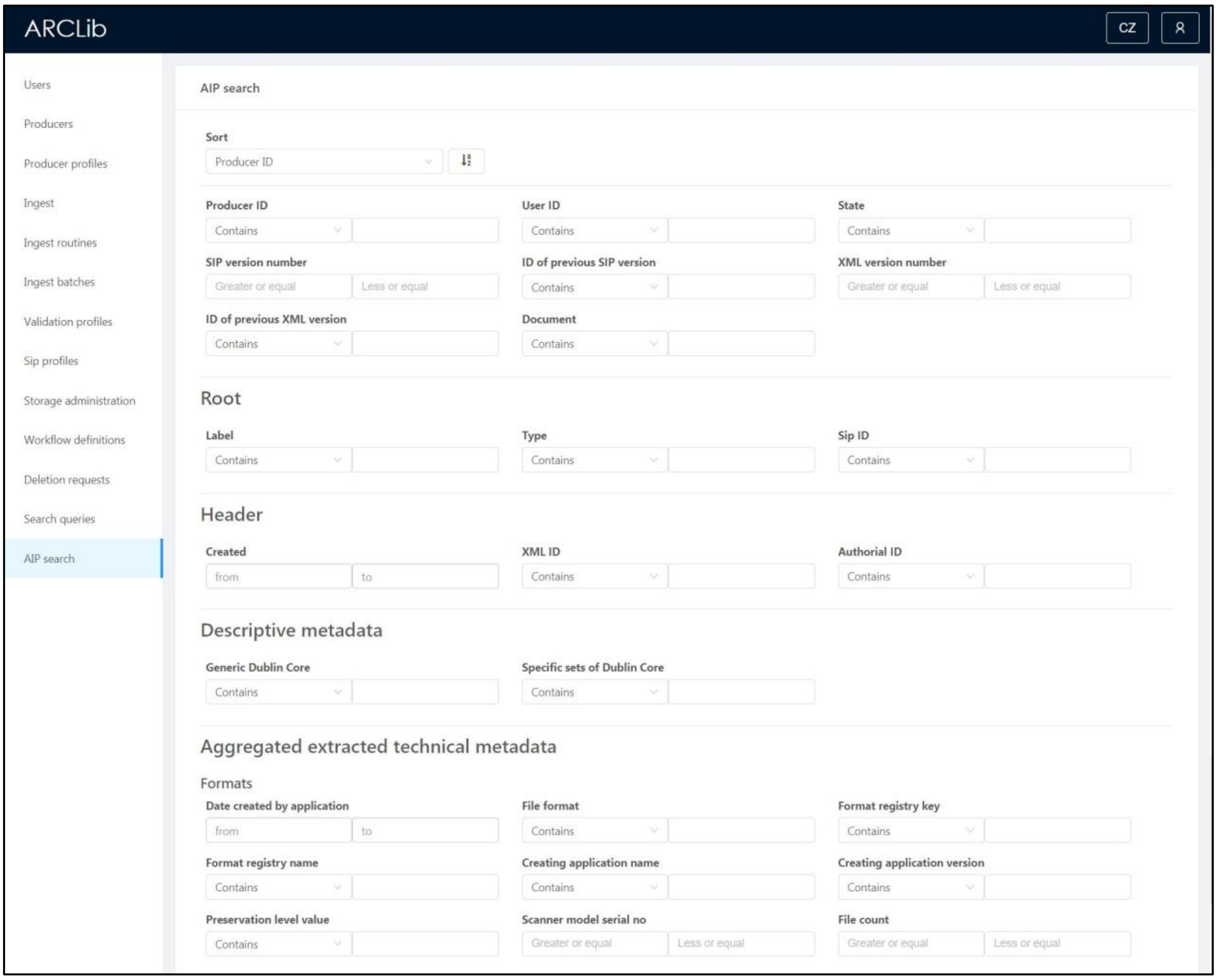
Figure 1: ARCLib schema

The ARCLib Ingest module

The ARCLib system envisages the input of data in the form of fully-fledged SIP created according to the pre-set standard employed by the institutions or external system. The functions of the module are primarily as follows: validation of input SIP according to validation templates provided by the producer, the extraction of metadata from SIP and the creation of new metadata. Metadata information is stored in a record in ARCLib AIP XML format, which is based on the METS and PREMIS international standards. The original metadata record of SIP is always preserved within the package. SIP processing proceeds according to the profile that is specific for the concerned data type of the relevant producer.

The ARCLib Ingest module is able to process the following structures of input data:

- ProArc National Digital Library monographs;
- ProArc National Digital Library periodicals;
- ProArc native monographs and periodicals;
- ProArc audio documents;
- National Digital Library periodicals and monographs;
- Archivematica DSpace;
- Archivematica General;
- National Digital Library electronic documents.



AIP search

Sort: Producer ID

Producer ID: Contains | User ID: Contains | State: Contains

SIP version number: Greater or equal, Less or equal | ID of previous SIP version: Contains | XML version number: Greater or equal, Less or equal

ID of previous XML version: Contains | Document: Contains

Root

Label: Contains | Type: Contains | Sip ID: Contains

Header

Created: from, to | XML ID: Contains | Authorial ID: Contains

Descriptive metadata

Generic Dublin Core: Contains | Specific sets of Dublin Core: Contains

Aggregated extracted technical metadata

Formats

Date created by application: from, to | File format: Contains | Format registry key: Contains

Format registry name: Contains | Creating application name: Contains | Creating application version: Contains

Preservation level value: Contains | Scanner model serial no: Greater or equal, Less or equal | File count: Greater or equal, Less or equal

Figure 2: The ARCLib Data Management module

ARCLib Data Management module

This module is primarily designed for the management of stored AIP. It contains information about the AIP stored in the system and makes it possible to search for and index such information. The module also makes it possible to browse the content of AIP and edit metadata, and provides the option of creating a new version. Tools for reporting are also part of this module.

Figure 2 shows the search interface of the system prototype currently in existence. Searching is adapted to the needs of digital data managers and is possible in relation to descriptive administrative and technical metadata. Searching is accessible via API.

The ARCLib Administration module

Administration includes a function for the configuration of workflow for the processing of Ingest, the relevant registers relating to this (the register of steps of the ingest workflow, the scripts used as part of ingest, the register of validation profiles, etc.). This module also includes the administration of the tools of third parties used within the system. It is also here that the administration of users and their roles and authentication settings is done.

The ARCLib Archival Storage module

This module is derived from the system and is approached as a separate application, which means that it can even be used outwith the ARCLib system. Functions for data management are available via REST interface. The module makes it possible to ingest and disseminate AIP, maintain information about the location of packages, check integrity, preserve operational metadata, update metadata, connect to a specific location of preservation technology, replicate data in a number of locations, back up, administer preservation technologies and media and report.

The ARCLib Access module

In light of the fact that ARCLib is conceived as a back-end application that is not intended for end users, the possibility of accessing data is restricted to the possibility of AIP export. The content of export DIP is equal to the content of AIP and is primarily intended for data producers. For this reason ARCLib does not contain any tools for forcing a policy which limits access to data.

The ARCLib Preservation Planning module

The functions of preservation planning, as they are perceived by the ČSN ISO 14721 standard, are shifted outside the system, in light of the character of the system - this primarily involves functions of an organisational and research-based nature (for example, monitoring a designated community or monitoring technology). Here the project primarily envisages the expert and methodological activity of the National Library of the Czech Republic. ARCLib itself comprises the basic tools for, in particular, work with formats.

The ARCLib system was launched in test regime on the infrastructure of the Library of the Czech Academy of Sciences in the autumn of 2018. This is the first prototype to contain all fundamental functionalities (following prototypes of individual parts). The individual elements of the system will be verified in this and, depending on the results, work will continue on modifying the system. The development process as a whole should culminate in verification of the functionality of the ARCLib tool in the form of semi-operation in the year 2020.

Conclusion

Once it has been completed, software output from the ARCLib will become a valuable tool for libraries and other institutions engaged in the long term preservation of digital data. Ideally, it will become an alternative to commercial solutions, as well as a variant which can be used by institutions that do not have the human resources or funds to operate a comprehensive solution to ensure LTP. The ARCLib system is also adapted to the needs of the Czech library environment. Nevertheless, the open-source character of the system will enable further development and potential expansion into other, related areas, particularly for other memory institutions.

The project will also produce two methodologies that focus on logical preservation and the preservation of bit-stream. These are review materials that are used both for work with the system and for general familiarisation with an issue.

References

ČSN ISO 14721, 2014. *Systémy pro přenos dat a informací z kosmického prostoru - Otevřený archivační informační systém - Referenční model.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví.

ČSN ISO 16363. *Systémy pro přenos dat a informací z kosmického prostoru - Audit a certifikace důvěryhodných digitálních úložišť.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

HUTAŘ, Jan and Marek MELICHAR. České paměťové instituce a digitální data - historický exkurz, současný stav a předpokládaný vývoj III. *Duha* [online]. **28**(2) [Accessed 25 September 2018]. ISSN 1804-4255. Available from: **http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1**

HUTAŘ, J., A. MIRANDA, E. PAVLÁSKOVÁ, Z. VAŠEK and Z. HRUŠKA, 2018. *Metodika logické ochrany digitálních dat* [online]. [Accessed 25 September 2018]. Available from: **http://hdl.handle.net/11104/0282107**

ROSENTHAL, Colin, Asger BLEKINGE-RASMUSSEN and Jan HUTAŘ, 2009. *Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER)* [online]. 1. vyd. Praha: Národní knihovna ČR. 65 p. ISBN 978-80-7050-569-4. Available from: **http://www.ndk.cz/platter-cz**

THOMAS, Lynne M., Jaime L. SCHUMACHER, Drew VANDECREEK, et al., 2014. *From Theory to Action: Good Enough Digital Preservation for Under-Resourced Cultural Heritage*

Institutions [online]. Washington (DC): Institute of Museum and Library Services [Accessed 26 September 2018]. Available from: **http://hdl.handle.net/10843/13610**