Qualitative and Quantitative Methods in Libraries (QQML) 9,1: 117-130, 2020

ARCLib – development of open source system for long-term preservation for library digital collections

Martin Lhotak
Library of the Czech Academy of Sciences

Abstract. Article 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. The solution is connected to Fedora commons repository and Archivematica, as far as it counts with creation of submission information packages (SIP) using these software solutions.
Keywords. Long term preservation, LTP, digital archiving, open source, OAIS

1. Introduction

The ARCLib project responds to needs of memory institutions and especially libraries to ensure long-term preservation of digital documents. The project includes the preparation of methodological materials and technical solutions, all of which will be freely available - methodologies in the form of open access publications and developed software tools as open source software.

Masaryk University, the National Library of the Czech Republic and the Moravian Library in Brno cooperate on the project led by the Library of the Academy of Sciences of the Czech Republic. TheARCLib project is under the designation DG16P02R044 solved in the period 2016-2020 with financial support from the Ministry of Culture of the Czech Republic within the national applied research funding program NAKI II.

Received: 13.9.2019 Accepted: 20.2.2020 ISSN 2241-1925

2. Main goals of the project

The main project aim is development of a LTP (Long-term preservation) solution, based on freely available open source applications and systems. An important part of the project is in creating best practice documentation for the long-term preservation of library digital collections. Best practice will be based on international long-term preservation standards (OAIS ISO 14721 and ISO 16363) and will reflect upon current systems used in Czech libraries for the generation, management and publication of digital data. Another best practice and working solution will be prepared for storing the data and bit-level preservation. The functionality of the whole solution will be validated in a pilot implementation at one of the project partners‘ institutions. The open source long-term preservation system Archivematica is considered as one of main building blocks. Archivematica is an established system with a worldwide community. The new ARCLib solution will enable small and midsize libraries to obtain and implement an OAIS compatible long-term preservation solution in an easy and affordable way. It will be a budget alternative to commercial solutions which often only big institutions on a national level can afford. The new system will have out-of-the box interoperability with the current LTP system implemented at the National Library of the Czech Republic. Archival packages are going to be compatible for both solutions. This will provide an extra level of preservation. ARCLib can be an alternative future option for the National Library system replacement which will have an available continuation in the ISO 16363 sense. ARCLib is an open solution with possibility to add new tools and services in the future. This way it can be implemented not only in libraries, but also in other memory institutions.

3. Current situation analysis

Since the second half of the 1990s, the Czech Republic has been digitizing and making available library documents in digital form. From the beginning, the most active were the National Library, the Moravian Library, the Academy of Sciences Library and some university libraries (e.g. at Masaryk University) [1]. In recent years, this activity has also increased significantly in regional libraries thanks to subsidies from European funds. Long-term protection of digital documents is directly addressed only in the National Digital Library (NDK) [2] project, in which the National Library of the Czech Republic and the Moravian Library in Brno use a commercial system for long-term archiving. No other archive solution is currently implemented in other libraries. At the Library of the Academy of Sciences of the Czech Republic, a solution for the production of ProArc digital documents is being developed, which will simultaneously provide certain archiving functions (creation of a SIP package for insertion into the archiving system). Libraries in the Czech Republic have standards for archiving ISO 14721 and ISO 16363, but there is no clear implementation methodology with regard to the Czech environment and the systems and formats used in it. At the same time, comprehensive archiving solutions such as the open

source Kramerius system are not readily available. The existence of a standard NDK format, which is followed in all major digitization projects and in smaller projects, is a great advantage in terms of the initial conditions of long-term archiving. The standard NDK package is based on currently used and recommended metadata formats (METS, PREMIS, MODS, Dublin Core, MIX and ALTO XML), which are based primarily on the standards recommended by the Congress Library and are respected internationally. The basic prerequisites for long-term archiving are ensured by the use of this standard, as well as the possibility of data sharing between individual digital repositories leads to improved conditions for the protection of digital documents. The NDK standard is designed for digitized documents and it is necessary to work on its extension for e-born (digital born) documents.

The National Library of the Czech Republic is currently using a tailor-made commercial solution for its long-term archiving project. This is mainly for documents digitized by the National Library and does not provide a general solution for the whole community. Many other documents digitized in regional and specialized professional libraries do not enter the archiving solution of the National Library and responsibility for their long-term archiving remains with individual libraries. All libraries that store digital and digitized data should have a freely available software solution that enables them to protect the data in accordance with the requirements of the OAIS standard. The archival solution in the National Library is currently also limited by several organizational and technical factors. Despite the fact that the current development is aimed at removing these restrictions, the procedure (and no suggestion of its solution) in case of termination of the NL NL repository as required by the ISO 16363 standard [3] is still not solved. From this perspective, an alternative solution would represent a way for the NL CR to fulfill the requirement of standards for the existence of an exit strategy.

Internationally, we encounter a different approach to long-term archiving of digital documents, using both commercial and freely available solutions. In some countries coordination at national level takes place, in others libraries work with academia to share knowledge and support research to address long-term archiving [4].

Of the freely available systems, Archivematica, the development of which is led by Canadian company Artefactual Systems Inc., is most often used for long-term archiving. Archivematica is developed in accordance with ISO standard OAIS (ČSN ISO 14721), but does not address all functional entities of this model [5]. It focuses only on critical archiving functions (transfer, reception, creation of SIP / AIP / DIP information packages). Further development is needed to meet all the parameters required by the standard. The entire archiving solution can be composed of parts that meet certain requirements of the standard, while providing all the functions as planned in the project, which will

result in an ARCLib archiving solution covering and interconnecting individual components.

4. Description of main goals

1) Development of LTP (Long Tem Preservation) open source solution ARCLib

The number of projects and applications generating large volumes of digital data is constantly growing in the libraries of the Czech Republic. Large-scale digitization projects are underway and more and more data is produced directly in digital form (born-digital). Long-term Digital Preservation (LTP) is essential to ensure long-term protection and accessibility for this data. It is necessary to provide both bit-level data protection (protection against physical loss, alteration or disaster of digital files and media) and logical protection (protection against adverse impacts of changes and obsolescence of information technologies and data formats on the availability and usability of digital information).

The issue of long-term protection of digital data (LTP - Long-Term Preservation or DP - Digital Preservation) was until recently the exclusive domain of large institutions such as national libraries or national archives, which had the necessary mandates, finances and expert resources. These institutions typically focused on building comprehensive tailor-made solutions based primarily on commercial systems. Advances in digital protection theory and practice, coupled with the growing need to address long-term digital data archiving in smaller institutions, have led to the recognition that scarce resources can begin to create custom solutions using freely available software.

The new ARCLib archiving solution will meet the requirements derived from the OAIS functional and information model, it will protect information content in AIP packages with all OAIS metadata, and will have tools to support all OAIS functional entities, including the “planning” preservation planning. To do so, the system user community will jointly maintain the knowledge base needed to make informed decisions in the long-term retention of information content in the developed system - a database of formats, rules and services, migration paths, tools - and perform OAIS retention planning functions.

ARCLib will be compatible with the commercial solution of the National Library of the Czech Republic and will allow the transfer of archive AIP packages between instances of the newly developed system to each other and to the LTP system in the NL, and vice versa. From the point of view of the OAIS model, this is the possibility of creating a network of cooperating OAIS archives linked by the repository exchange package standard (eg as http://wiki.fcla.edu/TIPR); the DIP output package from one system should serve as the SIP input package for other systems (and the NK system), and vice versa. Interoperability that enables the exchange of archival data on both sides

with a commercial LTP solution in the National Library will significantly increase the degree of security of archived data in the Czech Republic. At the same time, the NL can fulfill the requirement of archival standards for the existence of exit strategy

The solution will be able to store data in the structures and formats already stored in the Czech libraries (NDK, older standards of digitization, audio documents, map collections, theses, academic publications, etc.).
The input for ARCLib archiving solutions will be data from all majorly used software solutions for production, access and storage of library digital documents in the Czech Republic. In particular, digital documents from systems:

- Kramerius - a system for accessing digital documents; used in most large libraries in the Czech Republic
- ProArc - digital document production system; used by the Library of the Academy of Sciences of the Czech Republic, Research Library in Hradec Králové and Municipal Library in Prague
- DSpace - a repository used mainly at universities as an access system for both digitized collections and emerging digital documents (university archives, institutional repositories of scientific publications and research data); used for example by Masaryk University, VŠB-Technical University of Ostrava, University of Pardubice, Tomas Bata University in Zlín, Czech Technical University in Prague and others

ARCLib is designed as an open solution that will allow additional systems to be added to the archiving process if necessary.

2) Creation of methodology for long-term logical protection of digital data for the Czech environment with regard to international standards (especially the OAIS reference model - ISO 14721 and ISO 16363)
The methodology for logical protection of digital data presents procedures for long-term storage of this data, especially from library collections using the ARCLib software. It defines the necessary steps to fulfill the objective of long-term preservation of digital data at the level of logical protection in accordance with the procedures recommended by international standards ISO 14721 and ISO 16363. The methodology is based on these standards and represents their application on a specific system of long-term preservation. At this point it should be emphasized that such a document of a methodical nature with applied procedures in the Czech environment is still missing.

The logical protection of digital data is still a relatively new concept, although its importance is radically growing. It brings different and conceptually different procedures than usual in digital data storage. Qualified personnel using internationally proven and documented procedures, ideally combined with

diversified technologies, remain at the core of long-term protection. The purpose of this methodology is not only to provide guidance on how to manage digital data within a specific ARCLib software solution, but also to describe general procedures for performing logical protection steps and defining the requirements for repository personnel. The key features of logical protection systems include credibility, which can be ensured by compliance with standards and transparent system operation (in terms of professional service and documented processes and software). This is verified by a number of certification tools. Therefore, the proposed methodology also includes recommended procedures on how to certify in a credible way.

The aim of the methodology is to provide guidance to users of the ARCLib software solution on how to apply the above-mentioned procedures, how to manage data in the system and assess risks. It also documents specific functions of the solution, describes the way of data storage, their identification and structure. It is this documentation that is necessary to recognize the credibility of ARCLib repositories. The methodology is based on internationally defined requirements for logical protection systems of digital documents and converts them into specific procedures available within the developed solution. Using the instructions of this methodology in all the mentioned areas (operation of the solution, personnel and financial security and preparation for certification) is a necessary condition for the use of the ARCLib software solution. Although logical protection procedures are generally described, in each archiving solution their application differs based on the different nature of the software system.

The rules and procedures of the methodology were verified by its authors both in connection with other systems (e.g. the LTP system in the National Library of the Czech Republic, the digital repository of Charles University and other implemented systems that the authors met during their careers), in developing ARCLib solutions.

The methodology was certified by the Ministry of Culture of the Czech Republic in 2017 and it is freely available at http://www.nusl.cz/ntk/nusl-371612.

3) Creating a methodology and designing a solution for physical storage of large amounts of data and providing bit-level protection for long-term archiving needs

The methodology certified by the Ministry of Culture describes solution for physical data storage and bit-level protection within the ARCLib system for long-term archiving of digital data and documents. Part of this methodology is a description of basic storage requirements that can be used for long-term data storage along with bit-level protection.

Methodology considers and eliminates the risks that threaten data storage (hardware failure, unintentional operator error, intentional attack by an operator or other entity, natural disasters, armed conflicts, legislative restrictions on the use of data stored in a specific territory, etc.) minimizing the damage caused by such events - storing identical copies of data in multiple geographically separated locations on different types of repositories managed by different groups of people, while ensuring regular checks on data availability and integrity. An integral part of the data management policy is its regular revisions and adjustments according to changed circumstances over time. The described technical solution complies with the policy defined requirements for exit strategy (export all data in a suitable form for transfer to other / newer systems).
The proposed solution is well scalable (small and very large data volumes, system development with respect to the number of participants involved), with good throughput (technically solved e.g. by hierarchical data storage with fast online access of frequently used smaller data volume versus offline storage large amounts of data (which means high latency in data processing) and allows more independent solutions to meet the specific needs of each institution. In addition, the solution addresses some of the basic limitations that are specific to many types of storage - for example, the difficulties and limitations of storing too many small files, etc.

The methodology was certified by the Ministry of Culture of the Czech Republic in 2018 and it is freely available at http://www.nusl.cz/ntk/nusl-393240.

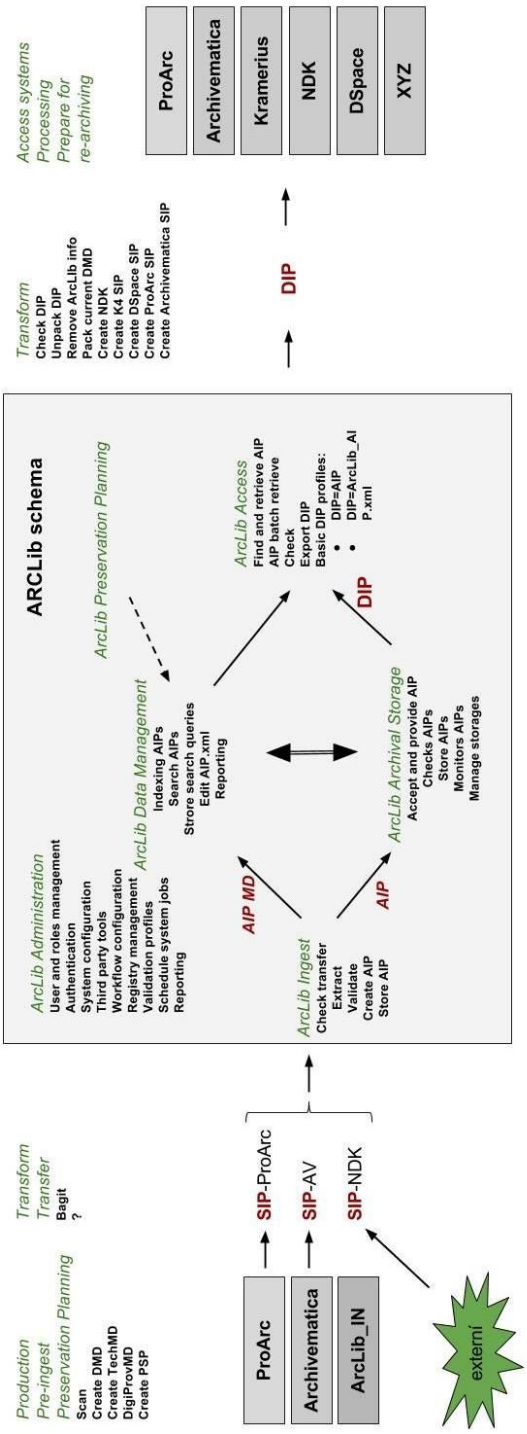
4) Practical verification in the form of pilot

ARCLib archiving solution will be deployed in the form of pilot and verified in the Library of the Academy of Sciences of the Czech Republic in 2020. For example, data from institutes of the Academy of Sciences of the Czech Republic will be stored and archived there.

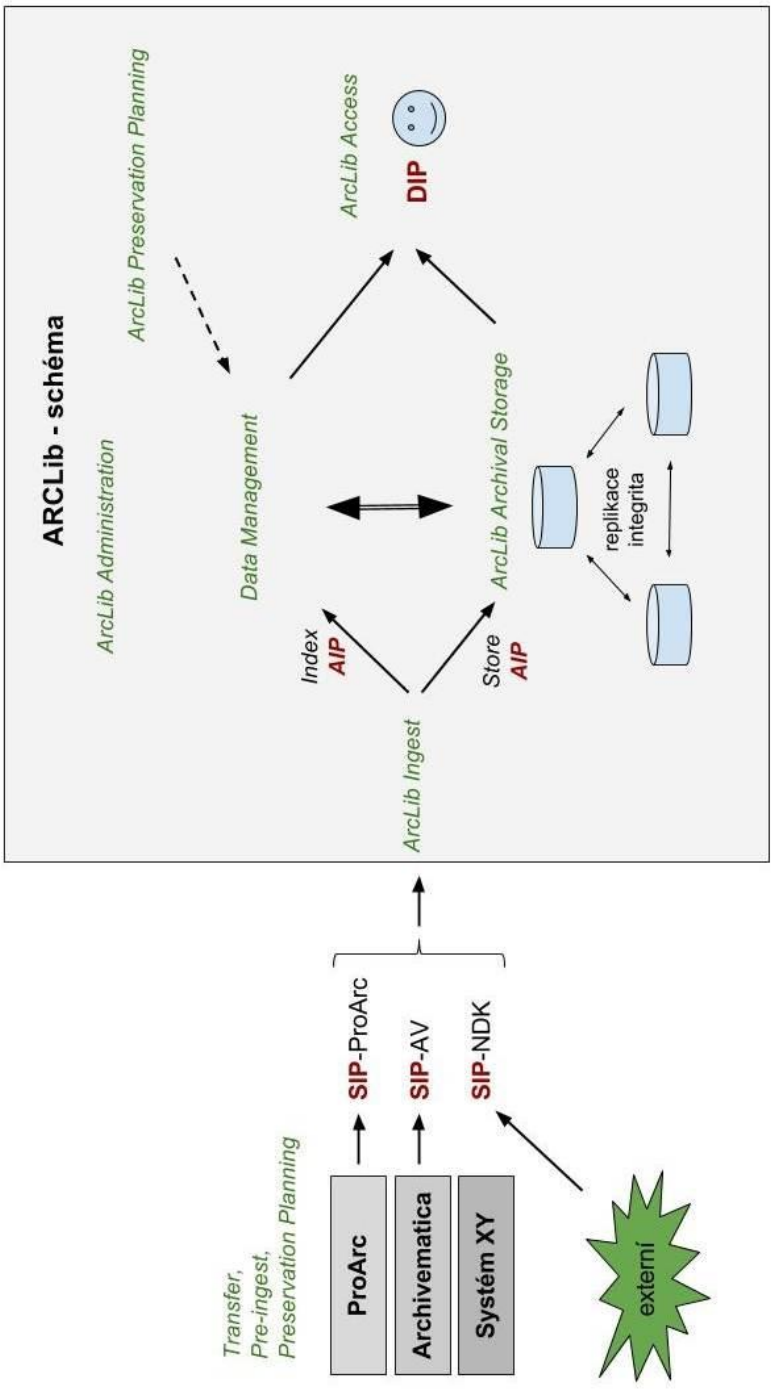
5. Project progress to date

The initial activity was to design the overall architecture of the ARCLib system, which will offer a comprehensive solution, but at the same time it will also provide the necessary flexibility by involving several different subsystems used or prepared in the Czech Republic. During the preparatory work on the ARCLib umbrella solution, the project team prepared a description of the application and functional requirements, which were the basis for the factual part of the tender documentation for the selection procedure for the supplier of programming work.

Description of the ARCLib solution based on the analysis of the project team:
FIG. No. 1 - ARCLib scheme


Production
Pre-ingest
Preservation Planning
Scan
Create DMD
Create TechMD
DigiProvMD
Create PSP
Transform
Transfer
Bagit
?
ProArc
Archivematica
ArcLib_IN
SIP-ProArc
SIP-AV
SIP-NDK
externí
ARCLib schema
ArcLib Administration
User and roles management
Authentication
System configuration
Third party tools
Workflow configuration
Registry management
Validation profiles
Schedule system jobs
Reporting
ArcLib Data Management
Indexing AIPs
Search AIPs
Strore search queries
Edit AIP.xml
Reporting
ArcLib Preservation Planning
ArcLib Ingest
Check transfer
Extract
Validate
Create AIP
Store AIP
AIP MD
AIP
ArcLib Archival Storage
Accept and provide AIP
Checks AIPs
Store AIPs
Monitors AIPs
Manage storages
DIP
ArcLib Access
Find and retrieve AIP
AIP batch retrieve
Check
Export DIP
Basic DIP profiles:
DIP=AIP
DIP=ArcLib_AIP.xml
DIP
Transform
Check DIP
Unpack DIP
Remove ArcLib info
Pack current DMD
Create NDK
Create K4 SIP
Create DSpace SIP
Create ProArc SIP
Create Archivematica SIP
Access systems
Processing
Prepare for
re-archiving
ProArc
Archivematica
Krameriius
NDK
DSpace
XYZ

or alternative Fig. No. 1 (with less details)


ARCLib - schéma
ArcLib Administration
ArcLib Preservation Planning
Data Management
ArcLib Access
DIP
Index
AIP
Store
AIP
ArcLib Archival Storage
replikace
integrita
ArcLib Ingest
Transfer,
Pre-ingest,
Preservation Planning
ProArc
Archivematica
Systém XY
SIP-ProArc
SIP-AV
SIP-NDK
externí

In the first phase of development, ARCLib focused primarily on the creation of ARCLib Ingest, ARCLib Data management and ARCLib Archival storage modules. Other functions must be implemented in other parts of OAIS, especially in ARCLib Administration, ARCLib Access. In particular, ARCLib Administration has a number of functions related to other functional entities.

ARCLib is a dark archive. It is not a repository designed to make documents available to end users. It does not have the means to display archived data (image servers, browsers, etc.). ARCLib users are administrators of archive digital data; after export, the data are used by access systems or other digital library systems (DAM systems). AIP update and creation of new versions of AIP are mostly done by editing data in external systems (ProArc, Dspace) and then re-Ingesting to ARCLib.

ARCLib AIP is an archive package management system, not an end user system or descriptive metadata management system. This is the basis for the design of the AIP structure and functions. In addition to SIP (in BagIt structure) ARCLib stores and maintains one XML with metadata - ARCLib AIP XML. ARCLib will allow versioning and deletion of AIP.

ARCLib Ingest assumes data entry at a more advanced stage of processing, i.e. data that already takes the form of a full-featured SIP package of a defined content standard created by systems such as ProArc or Archivematica packaged in a BagIt container.

ARCLib Data management contains information about AIP packages (or their parts) stored in permanent storage in the Archival storage module. It also provides an index and search interface. Ideally supplemented with reporting (above stored AIPs and reporting on processing performance). The DIP export event can be invoked from the data management search environment (now with DIP = AIP), ARCLib AIP XML information can be viewed individually or in bulk.

It is possible to search for descriptive metadata, administrative metadata, and technical metadata created in ARCLib (ie within the scope of ARCLib AIP XML content), and it is also possible to edit these metadata.
ARCLib administration module allows to configure the workflow for Ingest processing and controls the ARCLib system infrastructure. It contains user registers and their roles, their authentication settings and other administration processes are performed here.

ARCLib Archival Storage is a comprehensive bit-level protection service that enables replication to multiple geographic locations and multiple storage technologies. Archival Storage towards ARCLib provides Object Storage usable through a simple REST interface.

The philosophy of access for the ARCLib system assumes that users need to recover the embedded data in their original form. Thus, Access will allow you to export AIP as DIP, with the content of AIP and DIP being 1: 1. Further processing for making available to end users or updating AIP content (converting to Kramerius, modifying metadata, altering the structure and content of AIP, repeating format validation or re-extracting technical metadata) is already underway in other systems (ProArc, Archivematica). ARCLib is a back-end application and it is not intended for end users but just for archival administrators. Therefore, it does not enforce any policy restricting access to AIP data. Access Rights metadata is part of the supplied SIP, and it is checked at Ingest, but is not converted to ARCLib AIP XML.

ARCLib Preservation planning - much of the function of the preservation planning functional entity will be implemented outside the ARCLib information system. Definition and monitoring of the designated community and technology monitoring are activities of a mainly research and organizational nature and their performance is of interest to a number of communities. In the Czech Republic, the National Library of the Czech Republic plays a key role in standardization in libraries. Part of these activities is performed by the relevant departments of the NL CR, and users of the ARCLib system can follow their recommendations and issued standards.

Development of ARCLib including documentation, source code and link to running prototype is available at https://github.com/LIBCAS/ARCLib.

6. Conclusion

ARCLib is a system for logical and bit-level protection of digital data designed in accordance with requirements derived from ISO 14721 (OAIS). The main goal of ARCLib development is to create a solution that allows institutions to implement all OAIS functional modules and that respects the requirements of its information model. ARCLib makes the most of existing tools such as ProArc and Archivematica, especially for creating SIP packages. Prepared SIP packages are validated, converted to archive packages (AIP) and stored in accordance with OAIS.

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. Full version for production purposes will be available till the end of 2020.

This article was created within the research programme of the Czech Ministry of Culture „Programme of applied research and experimental development of national and cultural identity 2016 – 2020 (NAKI II)“, supporting the project„ ARCLib - Complex Solution for Long Term Archiving of (Library) Digital Collections “, identification DG16P02R044

References

[1] MELICHAR, Marek; HUTAŘ, Jan. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj I. Duha [online]. 2013, roč. 27, č. 4 [cit. 2019-08-20]. Dostupný z WWW: <http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj>. ISSN 1804-4255.

[2] LTP SAFE. LTP SAFE [online]. 2009 [cit. 2019-08-20]. Online: https://aipsafe.cz/pro-verejny-sektor/ltp/

[3] Systémy pro přenos dat a informací z kosmického prostoru – Audit a certifikace důvěryhodných digitálních úložišť: Space data and information transfer systems – Audit and certification of trustworthy digital repositories = Systèmes de transfert des informations et données spatiales – Audit et certification des référentiels numériques de confiance : ČSN ISO 16363 : schváleno v září 2011 ve Washingtonu, DC, USA. 1. vyd. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014, 53 s.

[4] MELICHAR, Marek; HUTAŘ, Jan. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj III.. Duha [online]. 2014, roč. 28, č. 2 [cit. 2019-08-20]. Dostupný z WWW: <http://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-1>. ISSN 1804-4255.

[5] UML Activity Diagrams. Archivematica [online]. 2011 [cit. 2019-08-20]. Online: https://www.archivematica.org/wiki/UML_Activity_Diagrams