

Analysis of Issues When Creating OAIS-based Archival Information Packages: Through an Exploratory Survey Using an Online Questionnaire

Boyoung Kim^{1,*}, Satoru Nakamura^{1,†}, Yasuyuki Minamiyama^{2,†} and Hidenori Watanave¹

1 The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, Japan

2 National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

Abstract

This study aims to analyze the issues encountered in the creation of Archival Information Packages (AIPs) based on the Open Archival Information System (OAIS) reference model, an international standard for digital information preservation, and to support efficient AIP creation. We conducted an exploratory survey using an online questionnaire and identified 23 issues. We further organized these 23 issues into practical and administrative categories and considered each category. Consequently, it became clear that there are issues related to the concrete definition of AIPs, documentation, metadata, file structure and format, collaboration, and access rights. Our results can be used as considerations for the preparation of the AIP, which is expected to facilitate the organization of abstract concepts within the AIP.

Keywords

OAIS, Information packages, AIP, Digital preservation, Archives

1. Introduction


Archiving digital information involves more than simply creating backups. Digital information is represented by a bitstream that is unreadable to the human eye without using reproductive equipment. Content stored on media such as hard disk drives, USB memory devices, and CDs can be extracted from the media and easily modified or duplicated. To preserve such digital information, which completely differs from paper, the information written in bitstreams must be associated with a specific format, hardware, software, or other information to ensure comprehensibility. Maintaining these relationships makes it possible to prevent digital information from being altered or lost, and the archived digital information can be used with trust.

Several models have been proposed to support the preservation of digital information. The Open Archival Information (OAIS) reference model provides conceptual models of the

* Corresponding author.

† These authors contributed equally.

✉ kim-boyoun688@g.ecc.u-tokyo.ac.jp (B. Kim); nakamura@hi.u-tokyo.ac.jp (S. Nakamura); minamiyama@nii.ac.jp (Y. Minamiyama); hwtntv@iii.u-tokyo.ac.jp (H. Watanave)

 0000-0002-0013-8144 (B. Kim); 0000-0001-8245-7925 (S. Nakamura); 0000-0002-7280-3342 (Y. Minamiyama); 0000-0001-6281-4606 (H. Watanave)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

environment, functions, and information required by archival institutions and systems that preserve digital information and serve as a standard for the reliable preservation of digital information. The information models defined in the OAIS reference model include Submission Information Packages (SIPs), Archival Information Packages (AIPs), and Dissemination Information Packages (DIPs). The person in charge of digital preservation creates SIPs using the information accepted from the producer and the AIPs for preservation. DIPs are created using AIPs. In other words, the AIP is the subject of long-term preservation; thus, its creation is an essential issue in digital preservation.

However, creating an AIP presents several challenges. For example, no guidelines specifically define the concept of an AIP [1]. Because the classification of elements included in an AIP is general, subjectivity cannot be eliminated in mapping metadata [2].

Furthermore, the AIP captures digital information, determines the elements necessary for preservation, and enters the initial stages of building an OAIS-compliant archive. Therefore, difficulties in AIP creation produce significant barriers to introducing the OAIS.

2. Subject and approach

This study aims to analyze the possible issues encountered during the creation of AIPs and to support more efficient AIP creation. The issues and methods for creating AIPs have been discussed in various studies. For example, the U.S. Library of Congress examined technical issues related to AIP design and the digital preservation of audiovisual items. It broadly defined the content and structure of an AIP [3]. The European Archival Records and Knowledge Preservation (E-ARK) project also surveyed existing key abstract concepts using a questionnaire and provided recommendations for creating an AIP based on them [4]. The results of both surveys modeled abstract AIP concepts and contributed to the creation of AIPs. However, both studies focused on the content and structure of the AIP itself and did not address more general concepts of AIP creation.

Therefore, this study conducted a questionnaire survey to analyze the challenges in creating an AIP from a broad perspective. Specifically, we surveyed people with experience building and operating OAIS-compliant archive systems to determine what they considered the most challenging issues in creating an AIP. The results were grouped to comprehensively analyze the challenges in creating an AIP. The results of this study will make it easier for archival institutions to interpret the elements of AIP, which are highly abstract, at a practical level. Consequently, barriers to implementing the OAIS reference model are expected to be reduced.

3. Overview of AIP

This section provides an overview of the AIP defined by the OAIS [5]. An Information Package (Figure 1) is a conceptual container that encompasses content Information (CI) and Preservation Description Information (PDI). They were wrapped and identified using Packaging Information (PI) and discovered using Descriptive Information (DI). The DI contains a portion of the package's external information, including basic information such as the title and ID and detailed cataloging information. The CI is the actual target of preservation and consists of content data objects and representative information to understand it. The PDI consists of Reference,

Provenance, Context, Fixity, and Access to the correct information. Table 1 lists information related to the elements of these packages.

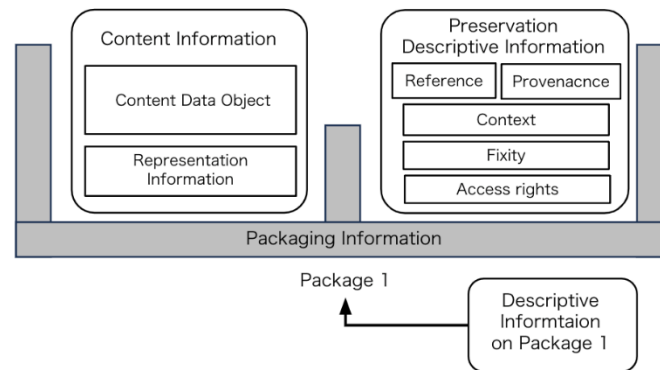


Figure 1: Concepts and relationships of information packages.

Table 1

Elements of Information Packages

Elements of Information Packages		Related Information
Content Information (CI)	Content Data Object	(Target data for preservation)
	Representation Information	Information that maps Content Data Objects to more meaningful concepts (e.g., software and hardware information for interpreting content information, file format information, semantic information such as terminology and data dictionaries, etc.)
Preservation Descriptive Information (PDI)	Reference Information	Identification
	Context Information	Information on the relevance of the CI to other information outside the Information Package (reasons for creation, records of other available CI)
	Provenance Information	History of CI (e.g., provenance, sources, logs, audit trail, etc.)
	Fixity Information	Information to prevent alteration (e.g., checksums, digital signatures, etc.)
	Access Rights Information	Information on access rights, including storage, distribution, and use of CI (e.g., OAIS permissions for storage operations, provision of licenses for distribution, information on management of rights information, and information on access control)
Packaging Information (PI)		Information that combines, identifies, and associates CI and PDI in a practical and logical manner

4. Survey

We conducted an online survey on the difficulties encountered while creating the AIPs. Few studies have addressed AIP creation from a practical perspective, and AIP creators' attributes remain unclear. We designed an exploratory questionnaire to obtain common insights into creating an exploratory questionnaire. All but two questions were open-ended and structured, similar to the interview survey. Although the nature of the questionnaire design resulted in a lower response rate, we determined that gathering a broader range of opinions would be beneficial.

The survey was conducted between March 6 and March 15, 2024, by posting to the "DIGITAL-PRESERVATION" of email discussion lists for the UK Education and Research communities, targeting a wide range of people interested in digital preservation. The survey consisted of eight questions and was sent to 2,388 e-mails. Anonymous valid responses were obtained from 18 people with experience building and operating preservation systems based on the OAIS.

The survey had eight questions. We analyzed questions 1-5. The results are presented in the next section.

List of questions:

1. What is your profession? (Archivist, Curator, Librarian, Researcher)
2. Please indicate the location (city or country) of the institution where you performed the work related to implementing or utilizing the OAIS.
3. How long have you been involved with OAIS?
4. What tools did you use to create an Archival Information Package (AIP)?
5. What was the most difficult part of creating AIPs and why?
6. What was the most challenging aspect of maintaining OAIS compliance?
7. If you answered "Others" above, please give specific details.
8. What preservation workflow do you think is the most important factor for creating AIPs? And why?

5. Results

This section presents the results of responses to survey questions 1-5.

5.1. Professions of respondents

Valid responses concerning the professions of the respondents were obtained from 18 respondents. Respondents held more than one position, with 23 positions identified (Figure 2); archivists (30%), librarians (22%), and engineers (13%) accounted for more than half of the respondents. There was a wide range of other positions, indicating the diversity of occupations related to digital preservation. For archivists and librarians, job titles were subdivided into digital archivists and system librarians.

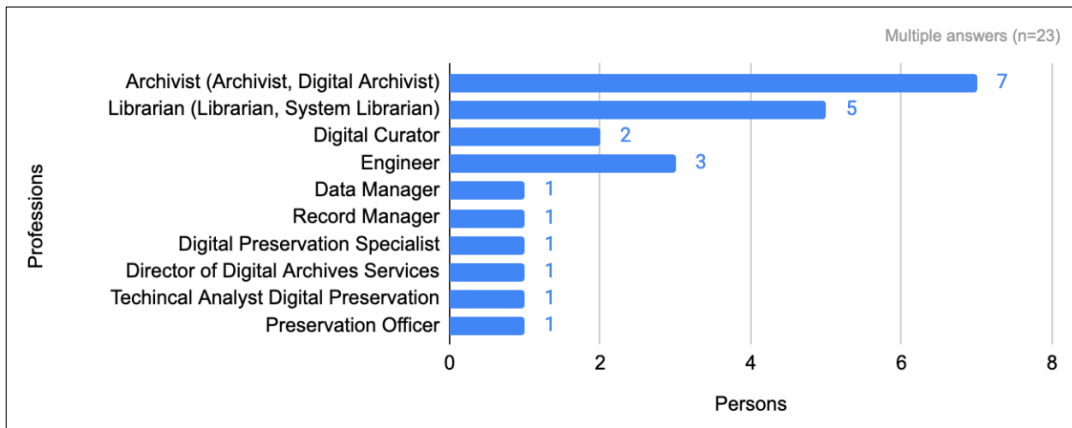


Figure 2: Professions of respondents (n=23).

5.2. Location of institutions

Valid responses concerning the locations of the institutions the respondents served were obtained from 18 participants. There are 21 archival institutions. Figure 3 shows the distribution of these institutions by country. Information identifying institutions was not collected; therefore, duplication is possible.

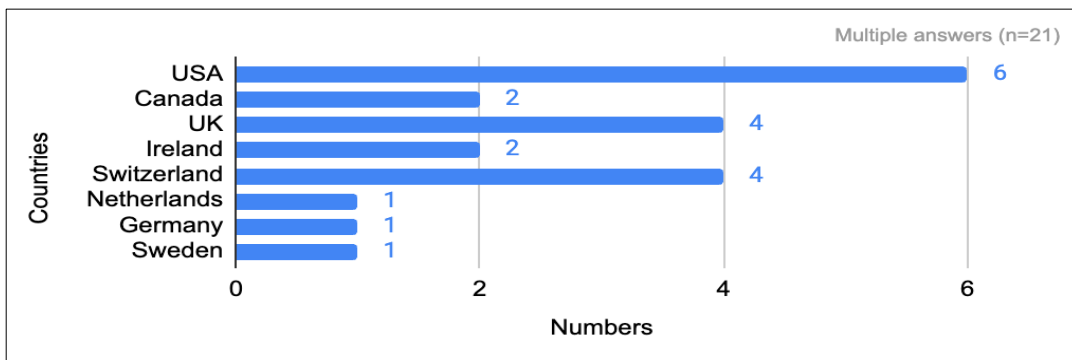


Figure 3: Locations of institutions (n=21).

5.3. Period of experience

Valid responses concerning the duration of the OAIS experience were obtained from 18 participants. Seventeen respondents (94%) had been involved in the OAIS for more than three years, sufficient to acquire basic knowledge of the OAIS. Therefore, it can be concluded that their responses were based on professional experiences.

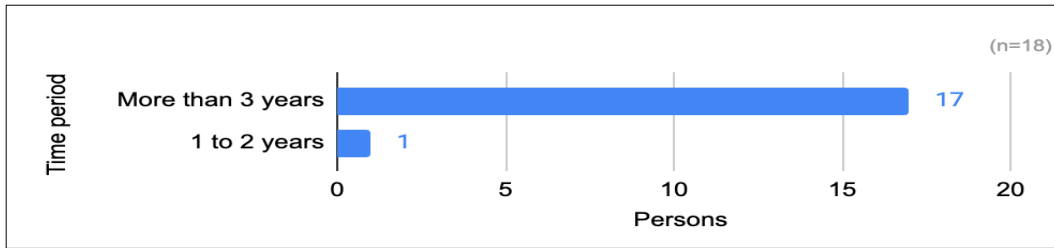


Figure 4: Period of OAIS experience (n=18).

5.4. Tools for AIPs

Valid responses concerning the AIP tools were obtained from 17 participants. As multiple tools are used by a single institution, they were divided according to the extent to which they can be split, resulting in the extraction of 33 tools. For example, three data sets were extracted from the answers to DROID, BagIt, and homegrown tools. Table 2 presents the results of further dividing the extracted tools into three types: in-house methods, open-source, and products/services.

Table 2

Details of Types

Types	Tools (Number)
In-house methods	File system (1), IFIscripts (2), Customized software (1), In-house tools (1), DLCM technology (1), Customized solution (1), Customized Ruby script (1), Customized scripts (1), Custom scripts and workflows (1), Home-grown tools (1), Word doc for representation document (1)
Open source	Open source solution (1), BagIt (3), XSLT transformer (1), DROID (3), JHOVE (3), TAR (1), PREMIS (1), Netpbm (1), CSV (1)
Products/Service	Archivematica (2), Rosetta (1), vizRT (1), NibNova (1), TIND (1)

We do not know the details of the in-house methods because we have not investigated the functionality of each tool in detail. Open-source tools are used for the following purposes, which we have compiled by considering the common tasks required for creating AIP, open-source tools' functions, and respondents' comments.

- Packaging: BagIt, TAR
- File format identification: DROID
- File format identification, validation and characterisation: JHOVE
- Metadata standards: PREMIS
- Transform XML files: XSLT transformer
- Manipulation of graphic images: Netpbm
- Recording metadata: CSV

Products/services are ready-made products that support long-term preservation.

Figure 5 shows the distribution of tools by type. Open-source methods were most commonly used (45%), followed by in-house methods (36%). Archivematica is the most frequently used product/service type.

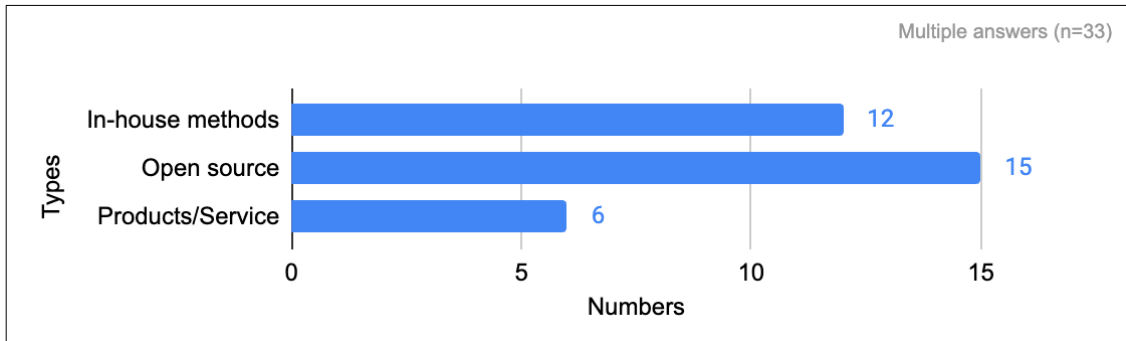


Figure 5: Distribution of tools by types(n=33).

5.5. Most challenging issues

Valid responses concerning the most difficult challenges of creating an AIP were obtained from 17 participants. If an answer contained more than one issue, it was split, and 23 issues were extracted. The results are shown in Figure 6, which compares practical and administrative issues. There were 18 practical issues (78%) and five administrative issues (22%). The details of the analysis method are described in Section 6.

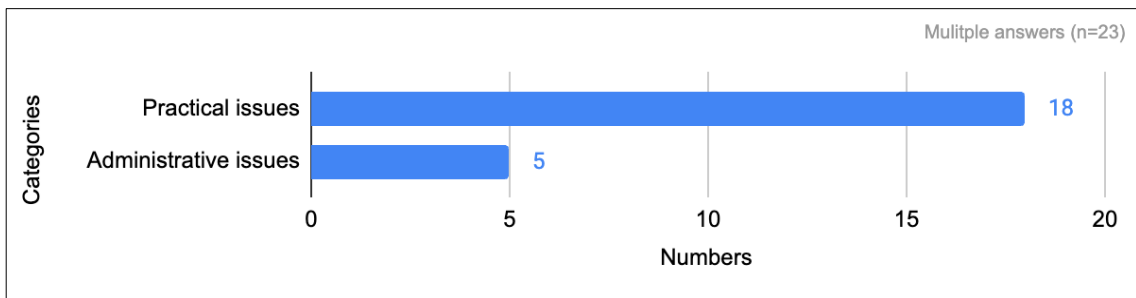


Figure 6: Distribution by issues (n=23).

6. Analysis of the most challenging issues

This section details the analysis of the most difficult issues in creating an AIP, as identified by experienced OAIS respondents, and is mentioned in Section 5.5.

The 23 challenges extracted from the responses were classified into two categories based on the following criteria:

- Practical issues: Issues in the pre-ingest and ingest phases of AIP creation that are directly related to AIP creation (Table 3)
- Administrative issues: Issues that are not part of the pre-ingest and ingest phases of AIP creation but impact AIP creation (Table 4).

Next, keywords were assigned to each issue to identify the similarities. The result showed that "Definition" was the most frequent keyword, followed by "Documentation," with "Metadata," "Structure," "File format," "Collaboration," and "Access rights" in third place (Figure 7).

Table 3

Practical Issues

No.	Most challenging issues	Keywords
B1	Definition of AIP as physical or conceptual packages	Definition
B2	Defining the object model for AIPs	Definition
B3	Modeling information elements	Definition
B4	Separating data (no modification allowed) from metadata, which can be modified	Definition
B5	Determining the canonical source of truth	Definition
B6	Creating a flexible data model to support a variety of access creation	Definition
B7	Creating comprehensive metadata for diverse digital objects can be complex and time-consuming.	Metadata
B8	Deciding on the principle choice between descriptive info versus metadata in AIP	Metadata
B9	Ensuring authenticity and integrity of records and maintaining high-quality digital evidence through the chain of custody.	Documentation
B10	Justifying each element of the AIP and each step in its creation by citing standards and guidelines such as OAIS	Documentation
B11	A huge amount of work is needed to prepare the records for ingest	Documentation
B12	Understanding how to structure Bags prior to SIP creation	Structure
B13	How to organize the structure of files, including data, metadata, rights, and permissions	Structure
B14	File format obsolescence	File format
B15	Understanding how to implement normalization rules by collaborating with stakeholders in the preservation department	File format, Collaboration
B16	Batch creation	Automation
B17	Completing AIP creation when the objects are damaged	Damaged objects
B18	The ingest and repository software mostly defines the AIP model, leaving little room for changes or modifications.	Modification

Table 4

Administrative issues

No.	Most challenging issues	Keywords
C1	Encouraging IT providers to implement AIPs in their systems	Collaboration
C2	Educate, advise, evangelize, and convince a user community to archive to preserve their data.	Outreach
C3	Store information on disclosure review, which requires a combination of automated and manual review	Access rights
C4	Legal and ethical Considerations	Access rights
C5	Understanding OAIS standard	Standard

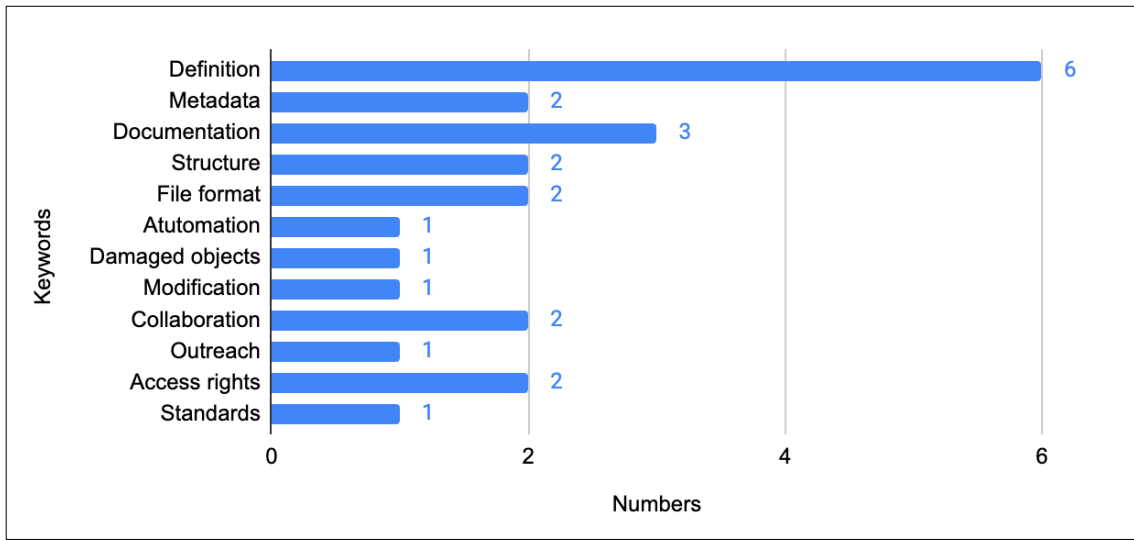


Figure 7: Distribution by keywords.

7. Considerations

This section provides an overview of the issues in creating an AIP based on survey analysis results.

7.1. Importance of definitions and documentation

According to the survey results, the most difficult tasks in creating an AIP are those related to definition and documentation operations. A definition is the process of determining the elements of information that should be preserved. For example, there is information on authenticity and reliability. In the case of archival materials, this information relates to the material's creator, provenance, and history. Documentation is a record of all work performed for preservation. For example, records of migration and normalization are relevant. These records are essential for maintaining the authenticity of the digital objects.

The fact that these are the most frequently cited challenges in creating an AIP indicates that preserving trustworthy digital information is an important and challenging task. In other words, prioritizing these two challenges is the key to creating a high-quality AIP.

7.2. Two perspectives in defining AIP

The survey results provided two important perspectives regarding the definition of AIP. One is whether an AIP is a physical or conceptual package. For example, if metadata exist in a distributed system and are used frequently, it may be challenging to physically generate them as a single package. This is expected to occur in many systems.

The second was the division of data according to whether modifications were allowed. In this case, descriptive information was obtained. Information in a collection may be updated or appended when it is arranged or used. Rather than including this information in the AIP, we may need to consider how to maintain a link to it so that it can be referenced to the latest information. These two perspectives provide valid criteria for defining AIP.

7.3. Issues related to descriptive information

One issue related to metadata is the selection criterion for descriptive metadata. Descriptive metadata are necessary information obtained from outside the AIP and used for discovering and retrieving the AIP. These may be created and updated to manage and provide access to the collections. The same applies to C3 Access rights and C4 Legal and ethical Considerations in Table 4. These are closely related to the two aspects of the definition of AIP described in Section 7.2. To help create AIPs, guidance and examples are needed that can be referred to when deciding which information is used or updated outside the AIP and is eligible for preservation.

7.4. Issues related to file format

Issues related to file-format degradation were highlighted from two perspectives. First, from a technical standpoint, ongoing migration is required to address file deterioration. The second perspective is from a collaborative perspective. An AIP includes records of the migration process according to normalization rules and data in converted formats. However, the understanding and interpretation of file format risks vary, making it difficult to reach a consensus among the parties involved. Regarding issues related to file formats, technical aspects tend to be emphasized, but the survey revealed that issues related to coordinating opinions within an organization are also important.

8. Conclusion

In this study, we analyzed the challenges encountered while creating AIPs based on the OAIS reference model, an international standard for the long-term preservation of the authenticity and integrity of digital information. We conducted an exploratory survey using an online questionnaire to identify the difficulties encountered when creating an AIP. We identified 23 challenges from a survey of people involved in the construction and operation of the OAIS. We further organized the 23 issues into practical and administrative categories and discussed each.

Consequently, it became clear that there are issues related to the embodiment and definition of the AIP concept, documentation, metadata, file structure and format, collaboration, and access rights. Previous studies have not focused on the challenges of creating an AIP from such a comprehensive perspective. By utilizing these issues as items for consideration in the initial

stages of AIP creation at archival institutions, considering the introduction of the OAIS, it is expected that they will be able to efficiently organize abstract concepts.

In the future, the results of this research will be verified and elaborated upon using concrete examples for generalization.

Acknowledgements

We would like to thank the "DIGITAL-PRESERVATION" mailing lists member who participated in this study.

References

- [1] D. Nicholson, M. Dobрева, Beyond OAIS: towards a reliable and consistent digital preservation implementation framework, in: Proceedings of the 16th International Conference on Digital Signal Processing (DSP'09), IEEE Press, 2009, pp.104–111.
- [2] H. Beedham, M. Palmer, R. Ruusalepp, Assessment of UKDA and TNA Compliance with OAIS and METS Standards, UK Data Archive, 2004. URL: https://dam.data-archive.ac.uk/reports/research/OAISMETS_report.pdf
- [3] Library of Congress, Library of Congress Archival Information Package (AIP) Design Study, 2001. URL: https://www.loc.gov/rr/mopic/avprot/AIP-Study_v19.pdf.
- [4] J. Rörden, R. Piret, S. Rainer, M. Thaller, C. Billenness, D. Anderson, J. Anderson, Archival Information Package (AIP) Formats and Restrictions, Zenodo, 2018. URL: <https://doi.org/10.5281/zenodo.1172649>.
- [5] CCSDS, Reference Model for an Open Archival Information System (OAIS) Recommended Practice, 2012. 650.0-M-2. URL: <https://public.ccsds.org/Pubs/650x0m2.pdf>.
- [6] Archivematica: Transfer, URL: <https://www.archivematica.org/en/docs/archivematica-1.15/user-manual/transfer/transfer/#>.