Preservation Policy

Preamble

ARCHE offers stable hosting of digital research data for the Austrian humanities community. A core function of ARCHE is digital preservation, which is essential for supporting current and future research activities. This document describes how born-digital and digitised data are sustainably preserved over time within ARCHE. By identifying key activities to sustain and add value to the hosted data, ARCHE's preservation policy intends to provide a comprehensive framework for decision making for future development of procedures and workflows, as well as provide essential documentation about archiving methods and processes (e.g., file formats, storage procedures, backup strategies, and migration procedures).

Scope of Preservation

The scope of this policy is limited to ARCHE's digital data collections. It applies to all data in all formats managed and stored for long-term preservation by ARCHE and for which ARCHE is the primary custodian. Internal administrative documents and correspondence are not in the scope of this policy.

Preservation Principles

ARCHE's implementation is based on the OAIS reference model (ISO 14721) for an Open Archival Information System (OAIS) and its related activities. An OAIS is an archive that has accepted the responsibility to preserve data and to make it available for others. As such, ARCHE meets all mandatory responsibilities described in the OAIS reference model, which will be described in the following subsections.

As its primary preservation strategy, ARCHE performs migration of formats as opposed to providing software emulation. ARCHE aims at preserving its digital data collections for re-use, while retaining authenticity and readability by mitigating risks such as deterioration, damage, loss, corruption or obsolescence of data storage, file formats, or dissemination means.

ARCHE recognises its responsibility to monitor any changes that may affect the sustainability and long-term accessibility of its data. Therefore, we continuously monitor technological and scientific developments and the implementation of such through our technical and organisational strategies.

Negotiation for Appropriate Information

«*Negotiate for and accept appropriate information from information Producers.*» [OAIS mandatory responsibilities]

Negotiations for appropriate information take place before and during the deposition process between the depositors and ARCHE. They are based on the Collection Policy and the Deposition Agreement, which detail the requirements for deposition of research data.

Key requirements include:

- Data must fit into the scope of the Collection Policy (consult Collection Policy)
- Any formats submitted should be formats suitable for long-term preservation (consult Formats for Long-Term Preservation)
- Sufficient metadata must be provided (consult ARCHE schema)
- Any legal issues are addressed (consult Deposition Agreement)

In order to help depositors meet these requirements, ARCHE provides information and consulting prior to the submission of data. This helps to ensure the quality of data and metadata, resolve any legal issues, and also reduce costs that may arise during the curation and ingest of data.

Obtain Control for Long-Term Preservation

«*Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.*» [OAIS mandatory responsibilities]

The Deposition Agreement ensures that ARCHE obtains sufficient control of the information for long-term preservation of digital research data. It clearly explains both the depositors' and the repository's rights and obligations, with regards to the deposition, curation, preservation, maintenance, and dissemination of data.

The depositor must sign the agreement and in doing so acknowledges that he or she has the right to deposit the data, allows ARCHE to disseminate the data according to chosen access modes, and has considered and taken care of any legal or ethical issues.

Determine the Target Audience

«*Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided, thereby defining its Knowledge Base.*» [OAIS mandatory responsibilities]

The designated community is specified in the Collection Policy.

Ensure Independent Understandability

«*Ensure that the information to be preserved is Independently Understandable to the Designated Community. In particular, the Designated Community should be able to understand the information without needing special resources such as the assistance of the experts who produced the information.*» [OAIS mandatory responsibilities]

The independent understandability of data is ensured by bespoke metadata associated with the data. ARCHE's metadata schema, which is formalised in OWL-format, makes the semantics of individual properties explicit, and provides a description of the data at project, collection, and individual resource levels. The schema was designed to be generic and flexible enough to preserve the documentation of heterogeneous resources from a variety of disciplines. International standards like the Dublin Core Metadata Initiative, DataCite, and Component Metadata by CLARIN were taken into account. The metadata set also includes properties to establish relationships to other publications, data sources, and related data.

Metadata properties can be mandatory, recommended, or optional. However, information on the top collection level is required to provide sufficient information to allow the collection to be understandable on its own.

With the help of automatically generated reports, designated data curators perform quality checks for consistency and accuracy of data. Since ARCHE's designated community is multidisciplinary, it is necessary to work in close cooperation with the depositors. A data curator inspects the metadata provided by the depositor to verify the consistency, understandability, and completeness of the data. If any issues arise, the data curator contacts the depositor for further actions. Insofar as it is possible, ARCHE will assign the curation of the deposited data to staff with a matching disciplinary background.

Document and Follow Policies and Procedures

«*Follow documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, including the demise of the Archive, ensuring that it is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.*» [OAIS mandatory responsibilities]

All the actions relevant to prepare sustainable preservation of deposited data are specified in our documentation and policies. All key steps are explicated in ARCHE's Deposition Process.

The Collection Policy helps depositors in deciding if ARCHE is the right place for their data. The relevant steps for the preparation of data by depositors are documented in the section "Deposition Process – Before Submission". The submission is completed by signing the Deposition Agreement.

After submission, automated checks are executed with dedicated software tools and any errors are reported back to the depositors with request for action or confirmation of proposed measures. Only after successful completion of automated checks, can manual curation take place. A general idea of curation activities is given in the "Deposition Process – Curation". Detailed curation procedures followed by ARCHE curators are internally documented.

The curation workflow respects the OAIS reference model by leaving the Submission Information Package (SIP) submitted by the depositor untouched in the process of preparing the Archival Information Package (AIP). The workflow was developed by considering the practices and recommendations from the Data Curation Network (DCN), the Micro-services by Archivematica, PAIMAS, and IANUS.

The overall curation workflow is divided into seven consecutive phases:

1. **Information and Contact:** Starts when first contact between the depositor and the archive is established and does not always lead to an actual deposition of files. It mainly involves consulting.
2. **Submission:** The depositor hands in the collection to be archived. The submitted information represents the SIP.
3. **Accession:** Starts after conclusion of the submission phase. This phase involves several tasks and concludes when the content is under full control of ARCHE.
4. **Curation (of data and especially metadata):** The curator in charge analyses the directory structures, file relationships and naming conventions used, checks for any quality or usability issues in the data as well as in the provided documentation and metadata. Any unaddressed legal issues, especially regarding copyright violations and sensitive data should be detected. The designated curator informs the depositor about any issues found. Fixes are applied in close collaboration.
5. **Transformation of SIP into AIP:** After the SIP was checked and curated, it can be transformed into an AIP. Transformation can include file renaming, rearranging of the folder structure, deletion, and conversion of file formats. Any changes applied are documented in the curation log.

6. **AIP Check:** The AIP is ingested into a staging instance of the ARCHE system where it is checked by the depositor and a second curator for any inconsistencies or undocumented transformations. This phase is repeated until no issues are found.
7. **Ingest to System:** The AIP is ingested into the production instance of the ARCHE system and automated checks for integrity are applied.

In order to avoid data loss, documented procedures for backup and recovery are in place. The physical safety and security of the preserved data is ensured by ARZ, the computing centre of the Austrian Academy of Sciences (OeAW). Any potential corruption of data is monitored with regular fixity checks.

In case of the demise of the archive, data and accompanying metadata can be transferred to another suitable archive. Dissemination services for downloading collections and respective metadata in ttl-format enable the export of the archive's content. Currently, records about format migrations and other alterations to the original SIP are kept outside of ARCHE. The provision of such provenance information in ARCHE currently is under development.

Disseminate Preserved Information

«*Make the preserved information available to the Designated Community and enable the information to be disseminated as copies of, or as traceable to, the original submitted Data Objects with evidence supporting its Authenticity.*» [OAIS mandatory responsibilities]

All data ingested in ARCHE are described via accompanying metadata. Metadata can be browsed on a dedicated web interface and accessed via a set of machine-readable endpoints, the dissemination services. This growing set of services delivers the data and metadata in custom DIPs, depending on the data type and format requested.

Three access modes can be applied on collection or individual resource level:

- Public content implies free access to the general public without any restriction. The classification of a resource as public content does not mean that the resources may be used for any purpose. The permissible types of use are further detailed by the licence accompanying every resource.
- Academic content requires registration as an academic user in order to access the resource. This is accomplished by authentication with the home identity provider by means of the Identity Federation.
- Restricted content is only available through special authorisation rules that are detailed in the accompanying metadata record. Usually, a dedicated user account will be required to access data.

The access mode is agreed upon in the Deposition Agreement and it is clearly displayed for every resource to the users.

Besides an access mode, data are also accompanied by a licence clearly stating how the resources can be reused. An open licence is preferred whenever possible. Available licences are listed in a dedicated vocabulary.

Metadata in ARCHE are public and licensed under CC0. This allows for wide dissemination, for example through external aggregators, which requires an OAI-PMH endpoint. The endpoint delivers metadata in a variety of formats and also groups metadata in thematic sets.

Formats for Long-Term Preservation

Due to the fact that ARCHE's primary preservation strategy consists in performing migration of formats, only a defined set of data formats are accepted for archiving. The data formats have to be suitable for long-term preservation, which implies a few key prerequisites. Suitable formats should be widely in use, compliant with open and non-proprietary standards (if possible), and when applicable, be acknowledged by the respective research communities. Whenever a choice for encoding is possible UTF-8 without the byte order mark (BOM) should be chosen.

A list of suitable formats grouped by file types is provided below. It is based on the "IT-Empfehlungen" by IANUS, which in turn take into account several international standards and recommendations, like e.g., the Guides to Good Practice by the Archaeology Data Service or the Sustainability of Digital Formats web site provided by the Library of Congress. This list of formats is subject to change over time, because new formats will be developed and others will become obsolete.

The list helps in informing depositors and curators to decide upon which formats to deposit and accept. Preferred formats are suitable for long-term preservation. Accepted formats require conversion, which might be provided by ARCHE upon request. If file conversions become necessary, potential loss of information should be minimised. If lossless conversion into an open or recommended format cannot be achieved the original files will be kept together with the converted versions.

Furthermore, files should not be password protected, encrypted or compressed in a lossy way. If files depend on references to other files, fonts or other external data, these objects should be deposited as well, or at least described in e.g., a plain text README file.

Exceptions are possible, but have to be carefully checked and decided on a case-by-case basis. One such exception can be raw data, as e.g., files produced by a measuring instrument.

See list with Formats for Long-Term Preservation.

Roles and Responsibilities

Preservation activities at ARCHE are carried out by a dedicated team situated within the larger organisational structure of the ACDH at the Austrian Academy of Sciences (OeAW). The OeAW provides funding for staff and resources (see Funding and Sustainability below) and its management provides the general strategic direction. The strategy is further refined by the ACDH's management, which also influences ARCHE's policies. Policy development and implementation is steered by the working group heads within the core unit DH Research & Infrastructure of the ACDH.

Team members of ARCHE can be roughly divided into curation and technical staff with a coordinating member who manages team communication, meetings and oversees activities of both parties in respect to their compatibility. The team jointly provides documentation and user information about ARCHE.

Curation staff are the contact point for depositors. Curators receive data deposits and check data for consistency, accuracy, and suitability for preservation and re-use. They also help in enriching metadata to increase findability of data and convert the SIP into the AIP for ingestion into the system. On a more abstract level curation staff is also involved in developing the ontology and required vocabularies, as well as in formulating the policies and legal texts. For any legal matters the legal department of the OeAW can be consulted.

Technical staff is responsible for developing and running ARCHE's underlying technology. For this task, it can rely on the provision of the technical infrastructure by OeAW's computing centre (ARZ), which leaves capacities for back- and frontend development. Technical staff does not only provide the core system components, but also develops additional features such as e.g., a file checker or a metadata editor.

Funding and Sustainability

Funding for core staff is provided from the global budget of the Austrian Academy of Sciences (OeAW). Maintenance and provision of server hardware is provided by the computing centre of the OeAW. Both secure basic operation and monitoring of the underlying systems as well as general management tasks such as sustainability for the operation of ARCHE.

Curation and ingest of data from institutional depositors are made possible by charging fees for the work required. Data deposits by individual depositors, as well as further technical development of the ARCHE systems, are covered directly by funds of the ACDH, which are financed by the global budget of the Academy, as well as a mix of national and international third-party financed projects, either conducted directly by the institute or by a number of external cooperation partners.

Addressing Risks to Digital Continuity

ARCHE actively work to ensure digital continuity of archived data, i.e., its usability over time. Any changes of organisation, management processes or technology are monitored over time. Required actions will be planned and applied in a timely manner.

Digital preservation at ARCHE is especially focused on mitigating risks connected to the rapid pace of technological development and the growing rate of digital data. Relevant risks and the measures to control them are listed below. The risk priority number determines the level of risk based on its probability and the potential impact it may have. The probability and the impact are indicated with a score of 1 (lowest) to 5 (highest) and their product represents the risk priority number.

File Format Obsolescence

Common file formats today might be superseded and not be supported in the future, due to changes in hardware and especially software. At ARCHE this risk is intended to be avoided by only accepting a defined set of data formats for archiving. This list is carefully curated (consult Formats for Long-Term Preservation) and updated when necessary. A technology watch particularly monitors any developments of formats with a high risk (see NARA Digital Preservation Risk Matrix) of becoming obsolete. Should a format become obsolete a bespoke migration plan will be developed and executed.

Risk priority number - 10 | Probability - 2 | Impact - 5

Complex Digital Objects

Very new, highly specialised as well as complex file formats with multiple dependencies do not always allow to make accurate preservation decisions. Future issues might arise suddenly if these types of digital objects do not get widely adopted or if they lack consistent specification. This risk cannot be avoided, but mitigation

is possible by keeping the original as well as a migrated file of the object along with sufficient technical documentation.

Risk priority number - 6 | Probability - 2 | Impact – 3

File Corruption

During transfer of data from one storage medium to another, the bitstream might become corrupted. A system failure might lead to the same problem. If file corruption is discovered at a very late stage, recovery might not be possible anymore. This risk is mitigated by the use of adequate server storage, a dedicated backup strategy and regular integrity checks based on checksums.

Risk priority number - 9 | Probability - 3 | Impact - 3

Storage Media Failure

The only storage media ARCHE relies upon is a dedicated institutional network storage (in a redundant RAID-6 setup). These are closely monitored and replaced in case of failure. Backups are stored on a separate storage system replicated in two physical locations ensuring recovery even in case of a disaster event.

Risk priority number - 5 | Probability - 1 | Impact - 5

Human Error (accidental deletion or alteration)

Human mistakes do happen and can lead to incidents where data is accidentally altered or even deleted. This risk is mitigated by granting write access to the archiving system only to well-trained data curators. Furthermore, a curation instance of the system functions as a safety net before data is ingested to the production instance. Before ingestion to the production instance the "two-man rule" is employed—no data is ingested to production before it is approved by another curator.
Full history of metadata changes is preserved, and all data are regularly backed up, thus any erroneous changes to the metadata or deletions of the binary content can be rolled back.

Risk priority number - 9 | Probability - 3 | Impact - 3

Non-Compliance

Non-compliance with legal obligations like laws, internal or external policies and standards can result in a loss of trustworthiness. Manual curation by trained staff helps to avoid this risk.

Risk priority number - 4 | Probability - 2 | Impact - 2

Organisational Risk

A change in the priorities of the Austrian Academy of Sciences (OeAW), a lack of funding or a general change in the political climate might lead to the demise of the archive. In any of these cases data and metadata from ARCHE will be transferred to another repository or even divided up between other suitable repositories.

Risk priority number - 3 | Probability - 1 | Impact – 3

Relevant Standards and Frameworks

ARCHE follows a variety of external standards, frameworks, and best practices for data management in general and for digital preservation in particular—all of which helps to maximise the quality of the data and the interoperability of the repository service. The list will be reviewed and updated when required. The Technical Setup provides an insight on the technical implementation of ARCHE's systems.

- Component Metadata of CLARIN (CMDI)
- CoreTrustSeal
- Dublin Core Metadata Initiative (DCMI)
- FAIR Principles
- Handle System (via ePIC)
- Linked Data Platform
- OAI-PMH
- OAIS Reference Model
- OWL
- W3C Standards

Sources

This policy is based on several sources listed below, which is gratefully acknowledged. All indicated links were last accessed on 11.11.2020.

Preservation Policies

- Archaeology Data Service (ADS): Collections Policy. https://archaeologydataservice.ac.uk/about/policies/collections-policy (Version 9.0)
- British Library: Digital Preservation Strategy. https://www.bl.uk/digital-preservation/strategy
- DANS: Preservation Policy, Version 1.2. https://dans.knaw.nl/nl/over/diensten/easy/toelichting-data-deponeren/DANSpreservationpolicyUK.pdf [No longer available.]
- DANS: Preservation Plan, Version 1. https://web.archive.org/web/20190505125409/https:/dans.knaw.nl/en/about/organisation-and-policy/policy-and-strategy/preservation-plan-data-archiving-and-networked-services-dans-1
- Data Preservation Coalition (DPC): Data Preservation Handbook. https://www.dpconline.org/handbook
- Open Preservation Foundation (2019-20): Digital Preservation Community Survey. Findings report. https://openpreservation.org/wp-content/uploads/public/resources/surveys/2020/OPFCommunitySurvey_Report_v03_200907.pdf
- RDA: Reagan Moore, Rainer Stotzka, Claudia Cacciari, Petr Benedikt (2015): Practical Policy. DOI: 10.15497/83E1B3F9-7E17-484A-A466-B3E5775121CC
- Scalable Preservation Elements (SCAPE): Barbara Sierman, Catherine Jones, Gry Elstrøm (2014) Catalogue of preservation policy elements. https://scape-project.eu/wp-content/uploads/2014/02/SCAPE_D13.2_KB_V1.0.pdf
- The National Archives: Preservation Policy 2018. http://www.nationalarchives.gov.uk/documents/preservation-policy-june-2018.pdf

- TIB: Franziska Schwab (2019) Export und Exit-Szenario. https://wiki.tib.eu/confluence/display/lza/Export+und+Exit-Szenario
- TIB: Preservation Policy der drei Zentralen Fachbibliotheken (Version 3, 2019) https://www.tib.eu/de/die-tib/policies/preservation-policy-der-drei-zentralen-fachbibliotheken
- TIB: Preservation Policy of the Technische Informationsbibliothek (TIB) – German National Library of Science and Technology (Version 1.4 2019). https://www.tib.eu/en/tib/policies/preservation-policy
- UK Data Archive: Preservation Policy 2019. https://dam.data-archive.ac.uk/controlled/cd062-preservationpolicy.pdf
- University of Edinburgh: Developing a Digital Preservation Policy. http://www.dpconline.org/component/docman/doc_download/1321-making-progress-hsbc-nov-2014-lee

Curation Workflow

- Archivematica: Micro-services. https://wiki.archivematica.org/Micro-services
- Consultative Committee for Space Data Systems: Producer-Archive Interface Methodology Abstract Standard (PAIMAS). Magenta Book (2004). https://public.ccsds.org/Pubs/651x0m1.pdf
- Data Curation Network (DCN): Lisa R. Johnston: Curating Research Data. Volume Two: A Handbook of Current Practice. http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/booksanddigitalresources/digital/9780838988633_crd_v2_OA.pdf
- IANUS: Archivierung bei IANUS. https://ianus-fdz.de/langzeitarchivierung
- UC Curation Center. https://escholarship.org/uc/item/5313h6k9

File Formats

- Archaeology Data Service (ADS): Guides to Good Practice https://guides.archaeologydataservice.ac.uk/g2gp/Main
- IANUS: IT-Empfehlungen für den nachhaltigen Umgang mit digitalen Daten in den Altertumswissenschaften. DOI: 10.13149/000.111000-a
- CLARIN: Standards recommendation. https://www.clarin.eu/sites/default/files/Standards%20for%20LRT-v6.pdf
- Library of Congress: Digital Preservation at the Library of Congress. https://www.loc.gov/preservation/digital/
- U.S. National Archives and Records Administration Digital Preservation Framework: https://github.com/usnationalarchives/digital-preservation
- U.S. National Archives and Records Administration Digital Preservation Framework: The NARA Risk and Prioritization Matrix. https://github.com/usnationalarchives/digital-preservation/tree/e746d3a6a6806f209b9c9272cc5c43d940cb8830/Digital_Preservation_Risk_Matrix

Risk Management

- Bodleian Libraries: Introduction to Digital Preservation: Risks to digital assets https://web.archive.org/web/20211009015508/https://libguides.bodleian.ox.ac.uk/digitalpreservation/risks
- Data Preservation Coalition (DPC): Digital Preservation Handbook. Preservation issues. https://www.dpconline.org/handbook/digital-preservation/preservation-issues
- The National Archives: Risk Assessment Handbook https://www.nationalarchives.gov.uk/documents/information-management/Risk-Assessment-Handbook.pdf
- ZBW: Risk Management and Preservation Planning in the Digital Archive of the ZBW. https://www.zbw.eu/en/about-us/key-activities/digital-preservation/risk-management