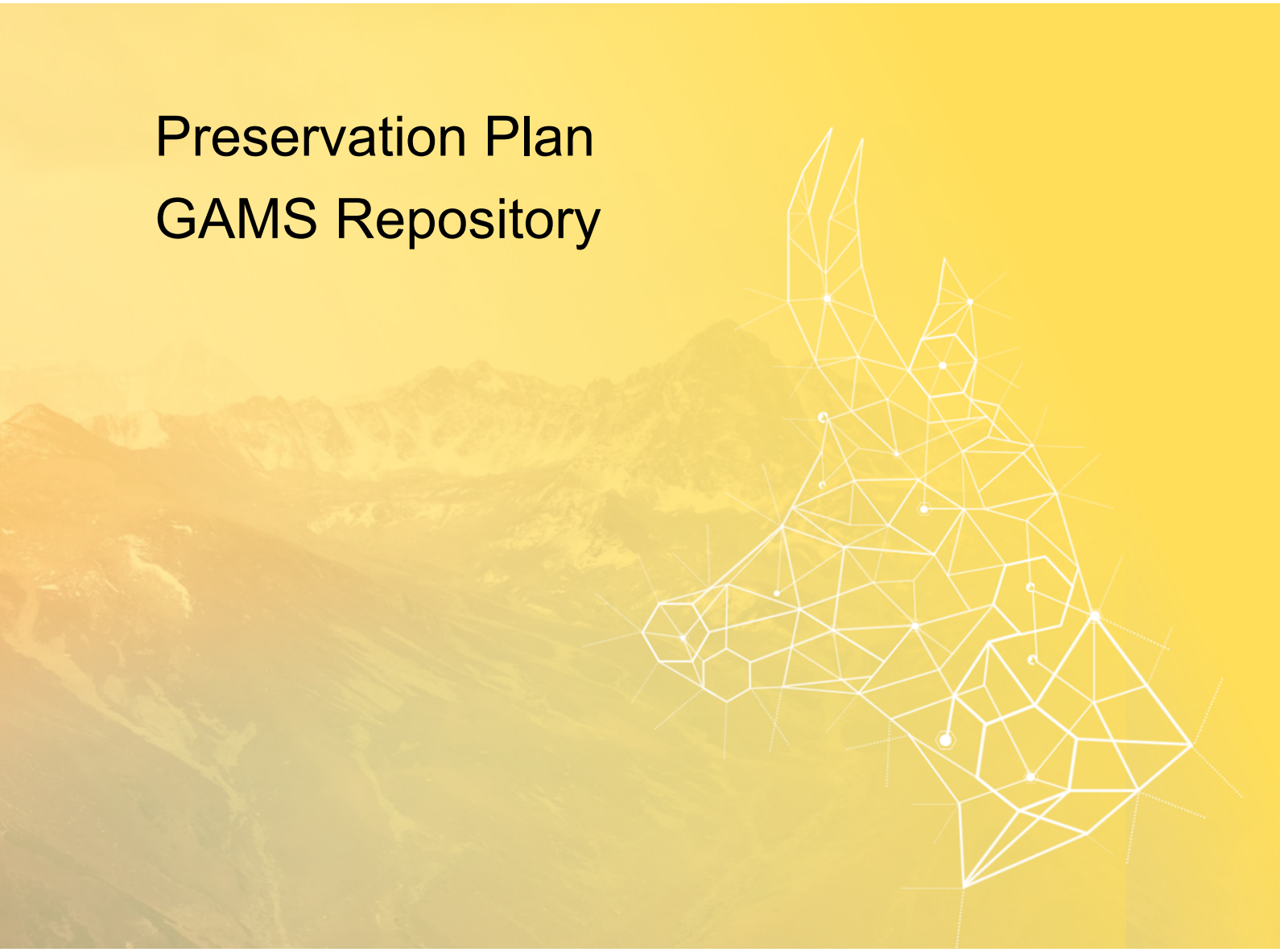
Preservation Plan
GAMS Repository



Published by the Centre for Information Modelling at the University of Graz in March 2023

Citation Recommendation:
Centre for Information Modelling, University of Graz (2023): Preservation Plan GAMS Repository. v. 1. URL: https://gams.uni-graz.at/o:gams.preservationplan, PID: https://hdl.handle.net/11471/521.1.3.

Table of Contents

1. Mission, Scope and Objectives

a. Introduction

GAMS (Geisteswissenschaftliches Asset Management System - Humanities' Asset Management System, henceforth the repository) is an OAIS compliant asset management system for the management, publication and long-term archiving of digital resources from the humanities. It enables scholars, researchers and students to manage and publish resources from projects with permanent identification and enriched with metadata.

Design and development of GAMS are carried out by the Centre for Information Modelling, an institute of the University of Graz (henceforth "the institute"), taking into account the specific requirements of humanities research.

b. Mission and Objectives of the Repository

The institute has an explicit mission in maintaining and developing the GAMS infrastructure now and in the future (cf. https://informationsmodellierung.uni-graz.at/en/institute/our-mission). The repository's objectives include keeping research data and results from the Humanities available over the long-term, thus serving the mission statement of long-term preservation. The statement also clarifies the collection scope (mainly research projects from the humanities) and designated community (primarily scholars and researchers from the humanities and related sectors like cultural heritage institutions) (cf. also https://gams.uni-graz.at/context:gams?mode=about&locale=en#collectionpolicy).

Preservation and availability of research data is guaranteed for at least 10 years in accordance with the research data policy of the University of Graz.

The institute and repository perceives itself at the intersection of research, infrastructure and teaching. Research is carried out and published using the repository and other infrastructure at the institute and experiences from that research and necessary skills are further disseminated via teaching in the institute's MA course programme. Furthermore, the repository commits to Open Access whenever possible (cf. https://gams.uni-graz.at/context:gams?mode=about&locale=en#openaccess).

c. Scope and Purpose of this Document

The present document represents the preservation plan of the repository and lines out the principles to which all preservation planning and infrastructure development of the repository conform. It applies to all digital holdings collected by the repository as the primary custodian with the objective of long-term preservation and availability.

d. Structure of the document and definitions

The structure of the document follows relevant guidance in the field of trusted digital repositories, namely the Attributes and Responsibilities Report (Research Libraries Group 2002), the OAIS Reference Model (The Consultative Committee for Space Data Systems 2012) and the FAIR Data Principles (Wilkinson et al. 2016). The definition and use of terms including acronyms and abbreviations follows those documents.

2. Attributes of the repository

a. OAIS Compliance

i. Workflow

Before starting out on how the repository applies the principles of the OAIS reference model, an explanation on the workflow and data life cycle is needed. In contrast to many other repositories, data isusually deposited only in combination with a joint research proposal of the depositors and the repository. The data forming the basis of research will then be collected and created in a cooperative effort. Such cooperation projects entail a person employed at the institute as a data curator and research data manager for this particular project. These employees usually have qualifications in a humanities' discipline as well as in digital methods and research data management and accompany the whole data life cycle from data capture scenarios, quality control and enrichment and publication via the repository including a partly project specific graphical user interface. This approach guarantees continuously high data quality and extensive communication with depositors throughout the research and deposition process.

ii. Pre-Processing

During the pre-processing phase, the project's data manager will set up a data creation and validation workflow and develop a suitable data model. This includes not only choice of formats and metadata standards but also enrichment with thesauri and authority files and validation and quality control mechanisms.

In particular, any legal or ethical concerns are addressed and solved in this workflow part. The repository provides guidance on these issues and supports necessary operations like anonymization or pseudonymization as well as choice of a suitable usage license (usually a Creative Commons license, preferably allowing open access).

iii. Ingest

The ingest process will typically be carried out by the designated data manager of the project, depositors do not self deposit. The data will be fitted to AIP specifications during pre-processing, suitable data formats and sufficient metadata will be included and checked once more at the ingest process with the ingest and management client. Accompanying material and administrative or legal documents not subject to long-term preservation can be kept at the internal GitLab instance or the university's group file share for later reference.

iv. Archival Storage

The archival storage of the repository has sufficient measures for storage, identification, retrieval and backup in place (cf. the sections below). There are different “templates” for AIPs available which provide a combination of suitable data formats and metadata elements for frequent scenarios (cf. the section on content models in the documentation: https://gams.uni-graz.at/archive/objects/o:gams.doku/methods/sdef:TEI/get?mode=&locale=en#cirilomodels).

Limiting the accepted data formats to a closed list of well-documented and long-term suitable data formats lies at the heart of the preservation approach. A list of accepted formats can be found here: https://gams.uni-graz.at/context:gams?mode=about&locale=en#formats. Primarily, various XML based formats are in use.

v. Data Management

Once ingested into the repository, resources are managed by Fedora’s internal database, the Mulgara triplestore. This includes all data management tasks like updating, querying, reporting, etc. and providing all necessary information for the other functional entities of the repository.

vi. Access

Usually the preserved research data is available via a project specific graphical user interface allowing users to browse and query the archived information. Every object has additionally a generic view with main metadata and links to all dissemination methods. Raw data in the form of links to the original XML files are also available. Depending on the project scenario other dissemination options like print versions, analyses or visualizations may be available. In all cases, the repository relies on the concept of single source publishing: all derived dissemination options are directly generated from the underlying data in the AIP. Data Re-Use is especially encouraged by following the FAIR principles (see below).

vii. Preservation Planning

Preservation planning includes monitoring and community watch by qualified staff, e.g. by taking part in national and international working groups and initiatives like DARIAH-EU, CLARIN or CLARIAH-AT. If threats to the infrastructure or data are discovered suitable measures will be implemented to counteract those threats. With regard to bitstream preservation, regular change of storage media in coordination with the university’s central IT department are in place. The primary strategy for logical preservation and data curation is

migration. This pertains also to the integral infrastructure parts, where technical progress and security concerns call for a constant update of all used software.

New community trends will be tracked and implemented only after the long-term viability of the new standards have been established.

viii. Administration

The administrative function of the repository is carried out by staff of the institute. This includes employees specializing in hard- and software, data curation and metadata management and legal matters.

b. Administrative responsibility

The repository's objective is to comply to the OAIS specification and the principles of trusted digital repositories as well as the FAIR principles. All policies and actions taken with regard to the archived data are in accordance with those standards. The repository communicates its regulations on the collection policy, preferred formats, commitment to open access and the like as transparently as possible through their web page. The present document forms part of this transparent approach.

c. Organizational viability

The repository is run by the Centre for Information Modelling, a permanently endowed institute of the Faculty of Humanities at the University of Graz. The mission statement of the institute reflects the purpose of long-term preservation of research data from the Humanities and to continuously maintain and further develop the dedicated repository. The institute employs a number of staff with permanent contracts to guarantee continuity of repository work in the fields of hard- and software, data curation and metadata management as well as legal consultation.

d. Financial sustainability

Hardware (e.g. storage) is covered by the university infrastructure, but does not constitute a major cost factor for humanities' applications and data. Due to the emphasis on open source software no costs for software incur. The major cost factor for the repository is thus human resources.
Funding and business plans for human resources rely on a mixture of permanent and short-term positions. The core personnel of the repository consists of permanent positions financed by the university. This is supplemented by short-term contracts financed usually by third party funding via joint project proposals in cooperation with depositors (see above); the individual project's data manager is thus funded by the respective project, while basic infrastructure work is funded by the university.

e. Technological and procedural suitability

The repository runs on a Kubernetes cluster on servers provided by the university IT department (uniIT). GAMS consists of a set of Linux based containers, which are scaled, managed and supervised by Kubernetes. A Prometheus/Grafana based monitoring solution allows us to keep track of metrics and logs.

The core component of the repository is a Fedora Commons instance accompanied by a set of services for dissemination and search purposes like data transformation services, a fulltext search engine and a IIIF server.

Data is stored in an exclusive Ceph storage cluster currently consisting of 24 nodes. Relevant data is backed every night to a backup server and an additional second off site backup.

For dissemination purposes, the repository used standardized formats like LIDO, METS and TEI. As a more user friendly access path GAMS also provides a web interface, which is based on HTML5, CSS, and JavaScript, all recommended W3C standards and validated upon publication of a project. Infrastructure development is largely guided by demands from preservation planning, i.e. obsolescence of software solutions and formats as well as continuous upgrade to newest software versions.

f. System security

All work is subject to the publicly available IT security policy of the University (in German only): https://online.uni-graz.at/kfu_online/wbMitteilungsblaetter.display?pNr=75196.

Physical and non-http(s) access to data is restricted to system administrators. HTTP-based write access to the repository is password protected and restricted to computers in the institute's subnet. HTTP-based read access is not restricted by default as we promote a creative commons strategy. In rare cases, when depositors require access restrictions to parts of their content, we are able to restrict access using http basic authentication on configurable URL patterns.

Access to the repository from the internet has to pass two hardware firewalls. Only ports 443 and 80 are open. As the system is running in a Kubernetes cluster, there is another barrier: All services are only connected to a virtual network using a private IP range which is inaccessible from outside this network. Only specific HTTP requests are routed via an ingress-controller.

Each chunk of data is stored redundantly in the Ceph cluster on at least three nodes (and therefore disks). The Ceph cluster containing the repository data runs in a dedicated private network not accessible from non-private networks. All volumes containing relevant data are backed up nightly to a backup server and a second offsite backup. Recent backups are available on a daily basis, older backups on weekly and monthly basis spanning a period of 7 years.

Backup and restore processes are documented and restoring is trained twice a year.

All services are monitored 24/7 with a reaction time of typically only a few minutes during working hours to some hours during weekends and nights. Software security and authorisation management is carried out by the system administrators and security officers of the repository. Responsible staff is adequately qualified and continuously trained to be prepared for any security related incidents.

g. Procedural accountability

Repository practices and preservation strategies will be planned and implemented in accordance with the legal framework of the repository and existing policies of the repository and the governing institution. The repository's responsibilities are assumed by following the below listed legal regulations.
Adherence to these principles is checked upon publication of projects and deposition of data in the repository. Accountability is underlined by obtaining certification with bodies like the Core Trust Seal or CLARIN.
Feedback from the community is gathered through membership in international organizations and working groups as well as attendance at subject-related conferences and exchange with fellow repositories. This serves also the purpose of community watch and monitoring new developments.

- Bundesgesetz zum Schutz natürlicher Personen bei der Verarbeitung personenbezogener Daten (Austrian Data Protection Act), https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=bundesnormen&Gesetzesnummer=10001597
- Bundesgesetz über das Urheberrecht an Werken der Literatur und der Kunst und über verwandte Schutzrechte (Austrian Copyright Act), https://www.ris.bka.gv.at/GeltendeFassung.wxe?Abfrage=Bundesnormen&Gesetzesnummer=10001848
- Directive (EU) 2019/790 of the European Parliament and of the Councilon copyright and related rights in the Digital Single Market, https://eur-lex.europa.eu/eli/dir/2019/790/oj
- Regulation (EU) 2016/679 of the European Parliament and of the Council on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, https://eur-lex.europa.eu/eli/reg/2016/679/oj
- Forschungsdatenmanagement-Policy der Universität Graz (Research Data Policy, University of Graz), https://static.uni-graz.at/fileadmin/ub/doc/Publikationsservices/fdm/PS_Forschungsdatenmanagement-Policy_der_Universitaet_Graz.pdf

- Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk
- Austrian Science Fund: Open Access to Research Data, https://www.fwf.ac.at/en/research-funding/open-access-policy/open-access-to-research-data

3. Roles and Responsibilities

The repository is developed and maintained by the institute. The present document forms the foundation of work for all employees, permanent or short-term funded, occupied with curatorial or technical duties in the repository.
Access to crucial system security and storage backend is limited to repository-related permanent staff at the institute.(referred to as "core staff"). Every depositor is usually assigned a designated (meta)data manager and curator responsible among other duties for checking (meta)data quality and data formats as well as compliance with the FAIR principles. Access for those employees is controlled by the core staff and only granted after previous internal training. Administration in the sense of the OAIS is carried out by a core team of permanent staff members with a mix of technical, administrative and legal skills.

Producers can be any partners signing a depositor's agreement with the repository; usually other university institutes, research projects and cultural heritage institutions. The designated community (the OAIS consumers) consist of a similar target group: especially fellow researchers from the humanities and the cultural heritage sector as well as the interested public. Jurisdicial responsibility is assumed by the University of Graz as a legal entity, operational responsibility lies with the head of the institute.

4. Application of FAIR Principles

a. Findability

All digital resources in the repository are assigned a Handle (prefix 11471) as persistent identifier. The compound digital objects mandatorily include a Dublin Core Record (where possible rule based extracted from primary datastreams like TEI/XML or LIDO/XML) and usually other metadata prefixes like EDM via the OAI-Provider. There are internal recommendations and templates for the correct completion of metadata fields and qualified personnel is checking automatically and manually if sufficient metadata is provided during the deposition process. (Meta)data is findable via generic services like Google but also domain specific aggregation services like Europeana, Pelagios, Nomisma, CorrespSearch or the CLARIN Virtual Language Observatory to increase findability for the designated community.

b. Accessibility

All (meta)data are only accessible via https, so that authentication procedures can be established on this layer if needed. This however effects only a minority of the datasets, most resources are available open access without restrictions. Other (proprietary) protocols are not in use.

c. Interoperability

Datasets are structured and described using community standards, preferably in XML serialization. This includes for instance TEI, LIDO, METS/MODS or Dublin Core. Many datasets also make use of RDF, SKOS and OWL. Wherever possible controlled vocabularies and thesauri are used for quality control and enrichment, most prominently GND, VIAF, GeoNames, Pleiades, Getty Thesauri and Wikidata, and links to other relevant resources inside or outside the repository are included.

Facsimiles and other images are available via IIIF.

d. Re-Usability

(Meta)data always includes a suitable license, preferably a Creative Commons license. Provenance information is usually part of the information provided at deposition and recorded in the source data. Datasets are largely structured and described with openly available and well-documented community standards, most prominently the TEI for digital editions.

References

The Consultative Committee for Space Data Systems (2012): Reference Model for an Open Archival Information System (OAIS). URL: https://public.ccsds.org/pubs/650x0m2.pdf.

Geisteswissenschaftliches Asset Management System (GAMS). URL: https://gams.uni-graz.at, PID: https://hdl.handle.net/11471/521.1.

Research Libraries Group (2002): Trusted Digital Repositories: Attributes and Responsibilities. URL: http://www.oclc.org/research/activities/past/rlg/trustedrep/repositories.pdf.

Wilkinson, M. D. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 2016. URL: http://doi.org/10.1038/sdata.2016.18.

Consulted Documents

ARCHE (2020): Preservation Policy. URL: https://arche.acdh.oeaw.ac.at/browser/preservation-policy

Bischof, Christian (2022): Preservation Plan. V1.0. Vienna: The Austrian Social Science Data Archive. URL: https://aussda.at/fileadmin/user_upload/p_aussda/Documents/Preservation_Plan_v1_0.pdf

DANS/EASY: Preservation Plan. URL: https://dans.knaw.nl/en/preservationplan/

UK Data Archive (2022): Preservation Policy. Version: 13.00. URL: https://dam.data-archive.ac.uk/controlled/cd062-preservationpolicy.pdf