GAMS – AN INFRASTRUCTURE FOR THE LONG-TERM PRESERVATION AND PUBLICATION OF RESEARCH DATA FROM THE HUMANITIES

by Johannes Stigler & Elisabeth Steiner

Abstract: *Research data repositories and data centres are becoming more and more important as infrastructures in academic research. The article introduces the Humanities' research data repository GAMS, starting with the system architecture to preservation policy and content policy. Challenges of data centres and repositories and the general and domain-specific approaches and solutions are outlined. Special emphasis lies on the sustainability and long-term perspective of such infrastructures, not only on the technical but above all on the organisational and financial level.*

Keywords: *research data; repository; preservation; sustainability*

GAMS – EINE INFRASTRUKTUR ZUR LANGZEITARCHIVIERUNG UND PUBLIKATION GEISTESWISSENSCHAFTLICHER FORSCHUNGSDATEN

Zusammenfassung: *Forschungsdatenrepositorien und Datenzentren gewinnen als Infrastruktur in der akademischen Forschung immer mehr an Bedeutung. Der Artikel stellt das Geisteswissenschaftliche Asset Management GAMS vor, beginnend bei der Systemarchitektur bis hin zur Archivierungsstrategie und den gesammelten Inhalten. Herausforderungen von Datenzentren und Repositorien werden zusammengefasst und allgemeine und domänenspezifische Lösungsansätze beschrieben. Besonderes Augenmerk liegt auf der Nachhaltigkeit und Langzeitperspektive von solchen Infrastrukturen, nicht nur was die technische Seite angeht, sondern vor allem auf organisatorischer und finanzieller Ebene.*

Schlüsselwörter: *Forschungsdaten; Repositorium; Archivierung; Nachhaltigkeit*

DOI: *http://doi.org/10.31263/voebm.v71i1.1992*

Content

1. Background and technical foundations

For 15 years the Humanities' Asset Management System (GAMS – Geisteswissenschaftliches Asset Management System)[1] has been providing long-term preservation of research data at the Faculty of Arts and Humanities at the University of Graz. According to the basic understanding of its operator, the Centre for Information Modelling at the University of Graz, the repository not only provides a technical solution, but a way of achieving sustainability in the handling of research data. The aim of GAMS is not only to provide long-term archiving and storage for digital content but to function as a platform for realizing standardized workflows in research projects in the Humanities. In cooperation with scholars from various domains the Centre has been working on questions of the digital representation of textual corpora, source material and other scientific content. Due to the increasing degree of digitization in research, modelling scholarly content has become more and more of an issue in the Humanities and related disciplines.

Design and development of GAMS have been carried out by the Centre for Information Modelling in cooperation with multiple partners within and outside the university, with regards to the specific requirements of humanistic research.

The Centre took over a technologically proprietary pool of research-supporting software projects from its predecessor. Since the maintenance of such a plurality of tools and environments had become more and more difficult and costly over time, a different approach was taken from 2003. All projects existing at that time were transferred to a single environment for long-term archiving and provision of scientific data and content. This new environment enables the Centre to conduct different research projects of various Humanities' domains with the same standardized infrastructure.

Based on standardized data models and annotation languages, sustainable solutions for semantic markup and enrichment of scientific content and sources create new concepts of long-term preservation and knowledge management.

The OAIS-compliant repository[2] enables scholars and students to publish digital resources in a citable manner including descriptive and technical metadata. Every object is referenced and thus accessible by a persistent identifier provided by the Handle system[3].

Currently, the repository contains about 80.000 mostly deeply annotated digital objects in 50 different, usually third-party funded research projects.

The vision of GAMS is to ensure sustainable availability and flexible (re-) use of digitally annotated and enriched scientific content. This is achieved through a largely XML-based content strategy based on domain specific data models. XML-based data formats such as TEI or LIDO provide means for flexible, metadata-enriched forms of storage of textual data. The primary content of documents is enriched with additional descriptive elements based on modelling standards. These standardizations provide a basis for the semantization and, consequently, the automated processing and analysis of specialist knowledge. Special emphasis lies on incorporating domain-specific ontologies and vocabularies. The separation of content and its presentation as a fundamental feature of XML-based formats implies a high degree of flexibility when dealing with the analysis and transformation of the original (textual) data in different presentation forms. On the other hand, this also calls for standardized workflows in the processing of such data.

This approach has created a pool of re-usable data objects from the Humanities over the past 15 years. In addition, automatic extraction of semantic relations of the ingested material implement further possibilities of textual analysis and content representation for the designated community of the repository.

One of the unique features of GAMS is that digital resources are not managed at the file level, but as complex digital objects: the digital representation of a digitized medieval manuscript consists of descriptive metadata, a plurality of facsimile photographs, a TEI-based[4] full-text transcription of the text and so on. All these data streams are stored within an object; every data stream can be seen as an attribute of the object in the sense of object-oriented programming. Wolf[5] (2004) mentions object orientation as one of the most important paradigms of software design. In combination with modular software concepts with high granularity as well as systematisation and standardization of object- and component oriented software development, this leads to a high degree of re-usability of software technologies. Object-orientation is characterised by: a) The definition of classes with associated features (attributes) and interaction interfaces (methods) as essential structural elements and b) the creation of class hierarchies by the usage of inheritance with the advantage of polymorphism.

In GAMS these principles are not only present on the level of system development but also structure the application logic on the user level: The design of content models (object classes) construct complex objects and object class hierarchies[6]. Content models not only describe the content structure of an object class (the datastreams) and possible relations to other objects (e.g. container objects) but also bind via WSDL[7] disseminators (methods) to a model. These can be XSL-transformations creating various output formats of the datastreams (HTML, PDF, etc.), the presentation of book-like content in special viewers, or e.g. methods to transform images to other sizes, formats or color models.

As an Austrian contribution to the European project DARIAH[8] the Java-based client and an Open Annotation[9] compliant annotation framework for Fedora based archives have been further developed into an archive-in-a-box solution; the package is available from http://gams.uni-graz.at/archive-framework. In an international and European context, GAMS not only contributes its (meta)data to various aggregation services, but acts as knowledge hub for new repositories as well as provider of expertise for re-use of the infrastructure. The source code of all components of this infrastructure can be obtained from Github (https://github.com/acdh/cirilo) and extensive documentation of client and policies can be accessed at http://gams.uni-graz.at/docs. The infrastructure is for instance re-used at the Austrian National Library in Vienna and at the Petöfi Literary Museum in Budapest, in both cases as an environment for the hosting of digital editions.

Implementing the aforementioned software design features and respecting the principles of long-term preservation, the GAMS project only uses open source software. Its core technologies are Fedora Commons (currently version 3)[10] for storage and management of digital objects, Apache Lucene and Solr for full text search, Blazegraph Triplestore as graph database, Postgresql Database Server as relational database, Apache Cocoon as main platform for web services used as object disseminators, and Loris IIIF Image Server to provide access to images via the IIIF Image API[11].

2017 marked the beginning of a major migration process to ensure accessibility of content for the next 15 years. An internal project takes on the task to migrate the core system of the repository. Special attention is being paid to the conservation and re-implementation of the interfaces already in use to guarantee that all surface components will remain working without adaptations. The new version of GAMS will be based on Fedora 4. All components will be available as Docker images and will be ready for clustering via Docker Swarm. A new version of the client for data creation and data curation will offer advanced functionality. The effective migration of the whole system to the new version is planned for 2019.

2. Research focus and content policy

The designated community of GAMS consists of scholars and students of the Humanities and related disciplines, as well as cultural heritage institutions (libraries, archives, museums).

Ingested data usually is highly structured and analyzed and annotated in various ways. There is a strong focus on text in all of its forms and historic shapes, establishing the field of digital editions as a main research focus of the Centre and thus the repository.

The repository's content ranges from rather modern sources like collections of 18th century journal texts (http://gams.uni-graz.at/mws), 19th century letter editions (http://gams.uni-graz.at/rollett) to older material like medieval accounting books (http://gams.uni-graz.at/srbas). Image centered projects are also available, like a collection of historical photographs from the Balkans (http://gams.uni-graz.at/vase) or a cooperation project with the Graz Museum publishing historical picture postcards of Graz (http://gams.uni-graz.at/gm). GAMS also partly functions as an institutional repository for the University of Graz, containing most of the digitally available artifact collections. This includes for instance the coin collection (http://gams.uni-graz.at/numis), the archaeological collections (http://gams.uni-graz.at/arch) the museum of criminology (http://gams.uni-graz.at/km) or an ethnographic collection (http://gams.uni-graz.at/ges). Another application is integrating resources of various digitized collections to create a regional portal of cultural heritage (http://www.kulturerbe-stmk.at).

Regardless of the source type, Humanities' data is characterized by its fuzzy and often uncertain nature; furthermore, it is subject to continuous and repetitive interpretation, creating an iterative research process. These characteristics have to be supported and represented in an adequate manner in a repository specializing in this field.

GAMS makes use of recognized international standards like TEI, LIDO[12], SKOS[13], EDM[14] or Dublin Core[15]. Evolvements in these community driven standards are adopted and the infrastructure is continuously further developed and improved with respect to new developments in the Digital Humanities.

According to the FAIR data principles data should not only be stored and archived but should also be findable, accessible, interoperable and re-usable[16]. Thus, the repository provides access to the data in the most FAIR way, exposing sufficient metadata not only in human readable (via various representational forms and search interfaces) but also machine readable (OAI-PMH[17] interface) form. This also includes delivering metadata to domain specific aggregation services like Europeana[18], CorrespSearch[19], Pelagios[20], or Nomisma[21].

Data can only be deposited to the repository as part of a cooperation research project; cooperation can be established within the University of Graz, but also with other universities or cultural heritage institutions on a national or international level. This excludes a variety of challenges data repositories are facing when receiving data from depositors in unknown formats not suitable for archiving or dissemination. Each project is accompanied by a metadata manager from the Centre to assist with the workflow, data modelling, deposition and publication process from the beginning to the end of the project. This approach guarantees the creation of high quality data prepared for publication and long-term archiving. Usually, materials are highly curated and enriched with cross-links and references to authority files and thesauri, which facilitates data discovery, interoperability and re-use.

Issues like IPR (Intellectual Property Rights) and licensing (possible limitation of access) are discussed and determined as early as in the course of project planning.

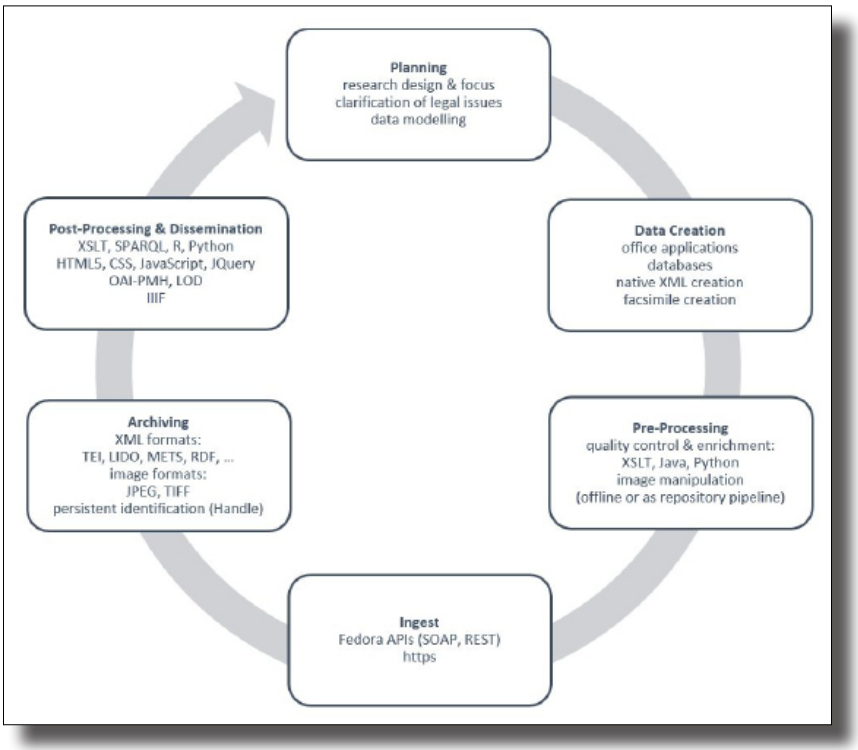

Fig. 1: Workflows and data lifecycle in the repository

Figure 1 summarizes the workflows and policies outlined previously. Project planning and data creation are carried out in close cooperation between the metadata managers representing the repository and the humanities' scholars. These steps lie outside the repository. The pre-processing step is a very important one: quality control and enrichment with semantic information, thesauri and references to authority files form the basis for all further steps. This can be either performed offline or be integrated into a repository pipeline before ingest. The figure also illustrates the limited list of preferred formats and technologies. The cycle can start anew when data is added or edited.

3. Challenges for research data repositories

In addition to the challenges posed by the nature of Humanities' research data, data centres face common challenges regardless of their scientific content and primary discipline. These include mainly organisational and financial issues.

The main tasks of a scientific data centre have been identified as a) deposit/collect, b) search/access and c) visualize/present; to a varying degree maybe also d) process/analyze[22]. Although these tasks have been defined by the working group for Humanities' data centres, this holds more or less true for all disciplines, although the visualize/present function may be less present or restricted due to legal restrictions in some domains. To fulfill these tasks, it is necessary to provide an adequate technical, organisational and financial environment for the repository. In contrast to many other infrastructures, GAMS covers all tasks within one infrastructure, using a limited list of preferred formats and technologies to ensure maintainability over a long period of time.

Contrary to popular belief the technical challenge usually does not prove the most difficult; given there is enough organisational commitment and financial support (in the form of skilled staff) this can be addressed adequately. This is also reflected in certification processes for Trusted Digital Repositories (TDRs). GAMS has been a TDR in compliance with the guidelines of the Data Seal of Approval[23] from 2014 to 2017, currently it is undergoing the process of renewal of certification with its successor CoreTrustSeal[24]. Technical requirements of course form a part of the reviewed guidelines, but emphasis is on the institutional anchoring and support of the repository and ensuring sufficient organisational sustainability and governance to fulfill its mission.

Sustainability often poses challenges in the academic context, given that permanently employed staff is often very hard to come by in university administrations. Repositories often have to cope with a relatively small number of permanent staff supplemented by project-funded short term support. Moreover, the profile of qualified staff such as research technologists, data curators and repository managers often does not fit in traditional categories of either administration related staff or scientific staff at universities: they sure need to be highly qualified and keep up with the latest developments in respective technical and scholarly fields, but often are not researchers aiming for higher academic qualification in the traditional understanding of university careers. In contradiction to common practice in university administrations, research data management is also a core task of the university itself and should not be outsourced to third party service providers. Research data repositories should be anchored in the university itself, providing an important part of research infrastructure equivalent to libraries.

GAMS is operated by the Centre for Information Modelling at the University of Graz, being a part of the Faculty of Humanities. The nature of a Centre is its temporal limitation: it has to be newly founded every five years. This sometimes poses challenges with regard to long-term preservation, since this task formally exceeds the life span of the institution handling it. Recently, the repository also experiences a gap between the amount of permanently employed staff and third party project funded staff, which may lead to a restriction of accepted projects in the future.

4. The future role of research data repositories

Research data repositories and data centres are becoming more and more important as infrastructures in academic research. They must be recognized as an integral part of research infrastructure and treated as such, i.e. repositories and data centres must be consolidated and perceived as a central responsibility of their academic institution. This entails financial and organisational commitment to building in house expertise and solutions by help of permanently employed and sufficiently skilled staff.

An important part of this consolidation is to foster interdisciplinary abilities and communication, not only between the IT and the respective domain (in our case the Humanities) but also across different departments at the institution and between academic research institutions and related institutions (e.g. cultural heritage institutions).

Sustainable development of software as well as high data quality and a limited list of preferred formats must be at the heart of every planning from the very beginning, otherwise the infrastructure tends to become less and less maintainable over time.

Ass.-Prof. Dr. Johannes Stigler
ORCID: http://orcid.org/0000-0003-0803-1496
University of Graz, Centre for Information Modelling – Austrian Centre for Digital Humanities
E-Mail: johannes.stigler@uni-graz.at

Mag.ª Elisabeth Steiner, MA
University of Graz, Centre for Information Modelling – Austrian Centre for Digital Humanities
E-Mail: elisabeth.steiner@uni-graz.at

1 Geisteswissenschaftliches Asset Management System – GAMS. http://gams.uni-graz.at
2 The Consultative Committee for Space Data Systems 2012: *Reference Model for an Open Archival Information System (OAIS)*. https://public.ccsds.org/pubs/650x0m2.pdf
3 Handle. http://handle.net
4 Text Encoding Initiative – TEI. http://www.tei-c.org
5 C. Wolf. Systemarchitekturen. Aufbau texttechnologischer Anwendungen. In: H. Lobin & L. Lemnitzer (Hg.): Texttechnologie. Perspektiven und Anwendungen. Tübingen 2004, S. 166–192.
6 Cf. R. Green. University of Hull digital colour image object specification. 2006. Online unter: http://www.hull.ac.uk/esig/repomman/downloads/INT-D3-1-imageObject-v03.pdf; R. Green. University of Hull digital public document object specification. 2006. Online unter: http://www.hull.ac.uk/esig/repomman/downloads/INT-D3-3-documentObject-v01.pdf.
7 Web Services Description Language – WSDL. http://www.w3.org/TR/wsdl.html
8 DARIAH. http://dariah.eu
9 Open Annotation Data Model. http://www.openannotation.org/spec/core/
10 Fedora Commons. http://fedora-commons.org
11 International Image Interoperability Framework – Image API. http://iiif.io/api/image

12 Lightweight Information Describing Objects – LIDO. http://network.icom.museum/cidoc/working-groups/lido/lido-technical/specification/
13 Simple Knowledge Organization System – SKOS. https://www.w3.org/2004/02/skos/
14 Europeana Data Model – EDM. https://pro.europeana.eu/page/edm-documentation
15 Dublin Core. http://dublincore.org/
16 The FAIR data principles. https://www.force11.org/group/fairgroup/fairprinciples
17 Open Archives Initiative – Protocol for Metadata Harvesting – OAI-PMH. https://www.openarchives.org/pmh/
18 Europeana. http://europeana.eu
19 CorrespSearch. http://correspsearch.net
20 Pelagios. http://commons.pelagios.org/
21 Nomisma. http://nomisma.org/
22 Cf. Arbeitsgruppe Datenzentren im Verband DHd 2017: Geisteswissenschaftliche Datenzentren im deutschsprachigen Raum. Grundsatzpapier zur Sicherung der langfristigen Verfügbarkeit von Forschungsdaten. DOI: http://doi.org/10.5281/zenodo.1134760, pp. 12–14.
23 Data Seal of Approval. https://www.datasealofapproval.org
24 CoreTrustSeal. https://www.coretrustseal.org