Ontologies in the Digital Repository: Metadata Integration, Knowledge Management and Ontology-Driven Applications

Sean M. WINSLOW [a], Gerlinde SCHNEIDER [a], Roman BLEIER [a], Christian STEINER [a], Christopher POLLIN [a] and Georg VOGELER [a]
[a] *Centre for Information Modelling, University of Graz*

Abstract. This paper illustrates how ontologies form an important foundation for the preservation, management, publication, and retrieval of humanities research data in the University of Graz' institutional repository for the Humanities, the Geisteswissenschaftliches Asset Management System (GAMS). The GAMS is a Fedora Commons-based repository which hosts more than one hundred research projects from different scholarly domains that extensively employ semantic web technologies and ontologies in their implementation. Many projects expose their research data as RDF and use shared vocabularies for shared entities. Using representative examples from digital scholarly editions, digital collections, and language resources, we show how ontologies enrich the user experience at the levels of modelling, analysis, and integration. The use of shared ontologies by projects with similar data allows search and retrieval scenarios to function across equivalent datasets in different projects and the formalization of data through large datasets allows the comparison of larger bodies of information than were possible using traditional (analogue) humanities methods.

Keywords. GAMS, Digital Humanities, Digital Preservation, Ontologies

1. Introduction

Semantic modelling and the use of semantic web technologies have become important research practices in the humanities. Scholars create Simple Knowledge Organization System (SKOS) taxonomies to hierarchically classify their research items, build ontologies to describe conceptually / model their research domain, organize their metadata according to vocabularies available in RDF, publish their research assets as linked data, or even create them solely as RDF resources. This poses new challenges for digital repositories, which aim to preserve digital objects for the long term, but also to make them manageable and publish them in a useful and reusable way. Linked open data from cultural heritage archives has a variety of use cases [1], and many projects expose their research data as RDF and use shared vocabularies for shared entities. The University of Graz' institutional repository for the Humanities, the Geisteswissenschaftliches Asset Management System (GAMS) hosts more than one hundred research projects from different scholarly domains that extensively employ semantic web technologies and ontologies in their implementation. In accordance with the principles of Linked Open Data [2], we strive to make hosted content, often shaped by many years of hard work, accessible to the widest variety of uses and users. While ontologies in the GAMS provide undeniable benefits to data integration and linking, RDF in the GAMS is neither just a convenient interchange format nor only a method to create a common discovery service: it is an essential part of making scholarly work available for future use, while embedding expert decisions in the process:

> "Digital editions [. . .] that not only publish text but try to express their interpretation of the text in a 'content' layer and that publish this interpretation online with the help of Semantic Web technologies do what scholarly edition is meant to do: publishing the critical analysis of the document by a competent scholar." [3]

Ontologies are an important part of our data ecosystem, controlling essential components of the infrastructure, but are also used to enrich data with context based upon scholarly expertise, enable linkability, and even control presentation. In the following, we will introduce the GAMS repository and discuss the specific ways that its architecture has been extended to foster work with semantic models. Using representative examples from digital scholarly editions, digital collections, and language resources, we will discuss how ontologies in the GAMS support the research needs of various projects.

2. The GAMS Infrastructure

The GAMS is an OAIS compliant digital asset management system used for the administration, publication and long-term preservation of digital resources. It enables scholars, researchers, and students to manage and publish resources from projects with permanent identification and enriched with metadata [4]. The repository, based on the free/open source software Fedora Commons, is platform-independent, web service-based (SOAP, REST), and—as is standard practice for humanities data [5]—pursues a largely XML-based content strategy. Various content models are provided, designed to meet the specific requirements of research data from different domains (e.g. implementing LIDO[1] for museum studies or TEI[2] for textual scholarship). Services defined for these content models combine the various data streams into useful presentational content that can be disseminated for public use via a number of different output formats and APIs. In addition to Fedora's built-in administrative client, the GAMS infrastructure offers *Cirilo*, a java client interacting with Fedora's Management API (API-M). The client provides custom functionality for data curation and management by the human user such as mass ingest or replacement operations [6].

GAMS pursues a strategy that focuses on the implementation of individual projects in cooperation with partners. Ideally, a project is developed as a technical collaboration with ontologies and linked open data in mind, from the definition of the research questions to the analysis and dissemination of the data. Figure 1 provides an overview over the workflows and data life cycle of the repository; The first two stages, project planning and data creation, are done outside the GAMS repository and are a collaborative work between the metadata managers at the Centre for Information Modelling and the

[1] http://www.lido-schema.org/

[2] http://www.tei-c.org/Guidelines/P5/

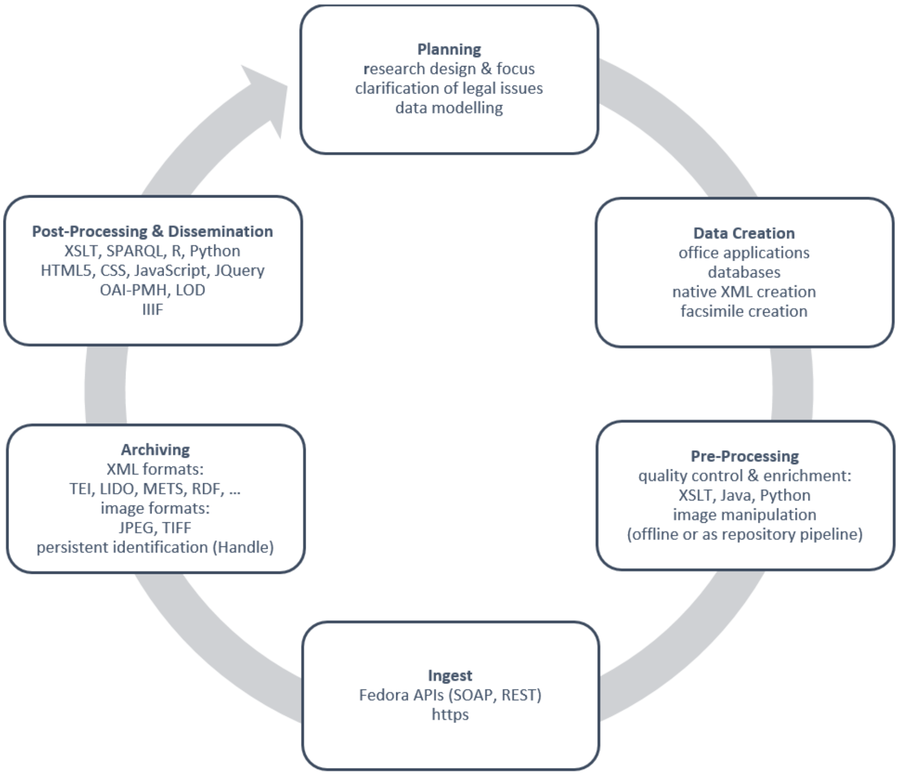

Figure 1. GAMS Workflow and Data Life Cycle.

humanities scholars. The third stage, pre-processing, which involves “quality control and enrichment with semantic information, thesauri and references to authority files,” is vital for the remaining stages of the life cycle [4]. Pre-processing is frequently done offline, but this stage can be added to a repository pipeline and thus it would be executed during data ingest. During ingest, data is uploaded to different triplestores (see next section). In the final two stages, Archiving and Post-Processing & Dissemination, ontologies are fundamental for the representation of scholarly knowledge (in RDF) and are the basis for data discovery, retrieval, and search scenarios (via SPARQL).

2.1. GAMS Extensions for Ontology Management

Fedora natively employs a Mulgara triplestore for internal management and control of object metadata and the assignment of individual objects to collections. A Blazegraph triplestore[3] extends the infrastructure for the storage and aggregated querying of structured data extracted from content data from different GAMS projects [6]. A specialized Marmotta triplestore[4] is currently being implemented to expand the repository's capacity

[3] https://www.blazegraph.com/
[4] http://marmotta.apache.org/

for the storage and use of geospatial information and allow GeoSPARQL-based queries.[5] We consider linked graphs the most convenient metamodel for structured data in the Humanities; accordingly, the GAMS offers a dedicated Ontology content model for the storage and management of data which can be structured as graphs. This content model uses "Ontologies" as a broad term covering both terminology (T-Box) and assertion (A-Box) documents. Ontologies serialised as Turtle or RDF/XML can be ingested into the repository and are stored as Fedora objects, where they are available as primary data streams. These objects are assigned a persistent identifier, can be versioned, and can be accessed via the web interface of the repository. Upon ingest, the data of each object is uploaded to the Blazegraph triplestore as a named graph. This guarantees consistency for data maintenance, as the data is persistently available in the repository as a separate object, as well as for aggregated query and retrieval across objects. Predefined searches are defined by a QUERY content model which functions as a "stored view," implementing predefined queries with parameters against the internal SPARQL-endpoint of Blazegraph and returning results as SPARQL-XML, JSON, or (via XSLT transformation) HTML.

The use of domain-specific data models generates meaningful, project-relevant RDF which semantically enriches the data, increasing the amount of information and the depth of domain knowledge encoded in the resulting object [7]. Project-specific SKOS documents both create new resources, as in the case of *Madg^{w}as* (described below) and extend existing SKOS resources, as in the case of the "Thesaurus for time periods and styles"[6] developed in the context of the project *Repository of Styrian Cultural Heritage*, an extension of the Getty Arts & Architecture Thesaurus which semantically links and expands upon that resource.

3. Use Cases

Here are a number of sample projects, highlighting different approaches to the use of ontologies in the GAMS:

3.1. Digital Scholarly Editions

While edited texts are not represented as graphs [8] in the GAMS (TEI-XML being preferred), several projects automatically transform TEI content and represent the information conveyed by the texts as RDF data:

DEPCHA. The Digital Edition Publishing Cooperative for Historical Accounts publishes editions of historical accounts using the Bookkeeping Ontology.[7] This conceptual model, derived from the REA accounting model [9] and the CIDOC CRM, has been developed in an iterative process involving historians, software developers, and digital humanists. It formalizes the interpretation of a transaction in a historical source as a combination of transfers of measurable objects, like monetary value, commodities and services, from one accounts record to another. It enables description of the various elements of transactions, such as debit-credit and the regularization of differing measures, goods, and currencies. Mapping different sources to a common schema lays the foundation for fur-

[5] `http://www.opengeospatial.org/standards/geosparql`

[6] `http://gams.uni-graz.at/o:pth`

[7] `gams.uni-graz.at/o:depcha.bookkeeping`

ther formal processing of the now-contextualized data, including direct linking to other GAMS projects, notably the “Rechnungsbücher der Stadt Basel.”

CANTUS NETWORK. Libri Ordinarii of the Salzburg metropolitan province project[8] analyses the liturgy and music of the medieval ecclesiastical province of Salzburg through liturgical ‘prompt books’ (*Libri Ordinarii*), which include a short form of more or less the entire rite of a diocese or a monastery. RDF assertions are extracted from TEI transcriptions describing hierarchies of time (feasts, canonical hours), entities (people, places, things), and metadata describing the manuscripts. A new RDF representation of Hermann Gotefends’ reference work *Zeitrechnung des deutschen Mittelalters und der Neuzeit* will facilitate the standardization and comparison of different historical calendrical systems as well as the linking of hierarchies of times from this project to others (e.g. the Corpus Kalendarium [10] or the cantusdatabase[9]).

CoReMa. Cooking Recipes of the Middle Ages: Corpus, Analysis, Visualization[10] documents the transmission of recipes in France and the German-speaking countries. Currently, this means that more than 80 manuscripts and about 8000 recipes have been marked up for the analysis of their origin, their relation, and their migration through Europe. The project is mainly concerned with food ingredients (i.e. animals, plants and fungi) which are already described in sophisticated detail by established ontologies[11] (including general knowledge bases like Wikidata and DBPedia) and used in other projects to present [11][12] and analyze [13] cooking data. The markup focuses on the occurrence of ingredients, preparation instructions, and tools in the texts. Not only does the project interact with the rich range of other ontologies available, but initial markup was effected through the use of (semi)automated alignment of Wikidata concepts based upon a list of medieval plant names[12] with the Reconciliation Service API[13] provided by OpenRefine.[14] With the editions marked up with links to the relevant concepts, they can be analyzed to reveal similarities in eating habits, the migration of text, and the influence of neighbouring countries on cuisine.

RTA 1576. Der Regensburger Reichstag von 1576[15] focuses on the Imperial Diet, the main forum for the Emperor and the Estates of the Holy Roman Empire to decide issues of pressing political concern. Modern historical scholarship highlights the role of communication practice beyond the political decisions [14]. The project focuses on the analysis of communication and negotiation documented in the records of the Imperial Diet and represents this in a formal ontology. Individuals are modelled as actors in a communication situation, i.e., as agents addressing others or those being addressed. Communication itself is a temporal entity in the sense of the CIDOC CRM (E2), which has a geographic place and time when it happens, a subject, and a duration. The situation can be classified as formal, i.e. following mostly implicit procedural rules established by

[8] https://gams.uni-graz.at/context:cantus

[9] http://cantus.uwaterloo.ca/home/

[10] https://corema.hypotheses.org/

[11] For an overview of ontologies covering these topics see http://www.ontobee.org/, http://aims.fao.org/,https://ndb.nal.usda.gov/ndb/, https://agclass.nal.usda.gov/about.shtml, http://zbw.eu/stw/version/latest/thsys/70498/about.de.html all of which are connected to the Linked Open Data Cloud by providing its data in one or more serializations of OWL and/or RDF(S).

[12] http://medieval-plants.org

[13] https://github.com/OpenRefine/OpenRefine/wiki/Reconciliation-Service-API

[14] http://openrefine.org/

[15] https://reichstagsakten-1576.uni-graz.at/

practice, or informal and it can be face-to-face or absent, etc. By explicitly modelling these aspects, communication situations are made comparable and may be studied conceptually, independent of, but linked to, their particulars. Additionally, it allows to build a conceptual index as entry-point for human reading of the texts.

3.2. Digital Collections

Digital collections in the GAMS use RDF as a metadata format, data model representation, and for data content (in the form of controlled vocabularies). This ensures consistent description of the data [1].

Madg^{w}as is an in-development project where ontologies drive the workflow and presentation. A catalogue of Ethiopian binding decoration patterns which aims to markup one instance of each binding tool pattern on each manuscript in the project, a major product of the project is a new (SKOS) ontology describing the different forms of tools. The resulting ontology is used in the process of generating the knowledge; when adding new material, the project uses an annotation tool which directly references the current version of the ontology, forcing decisions about whether it fits the data, as presented, at the time of markup. Patterns which do not fit the ontology are integrated during periodic refactoring, and the fit iteratively becomes better over the life of the project. Since the project takes a representative sample of each tool from each manuscript and makes it available as Linked Open Data, the collection of tool impressions as a whole can be re-organized by other scholars' ontologies, while still preserving the links to relevant catalogue and secondary information.

CEI2TEI, part of the Charters Open Research Data initiative,[16] is modernizing the model and long-term preservation environment of the *Illuminierte Urkunden* (Illuminated Charters) collection from monasterium.net[17] as a test case for the next stage in archiving for the remaining approx 600,000 Monasterium charters.[18] It has adapted and updated the Charters Encoding Initiative schema[19] to be a compatible extension of the Text Encoding Initiative P5. It uses ontologies as part of the design process for the TEI extension, to control vocabularies, and to enrich the TEI with glosses on art-historical and diplomatics concepts used in the description of charters. Many TEI elements accept pointers in their attributes (in the form of URIs); accordingly, the project provides a SKOS ontology to provide URIs to support a project-specific interest in authenticating features.[20] The flexibility of graph-based object management in Fedora also allows the overlapping of collections in a more flexible arrangement than the strictly hierarchical (eXist-db) XML database used in the source portal.

3.3. Language Resources

The management of linguistic resources in the GAMS infrastructure is relatively new, and is being developed to model and react to practices in the field of study. From the

[16] `http://gams.uni-graz.at/context:cord`

[17] `http://monasterium.net/mom/illuminierteurkunden/collection`

[18] `http://monasterium.net`

[19] `http://www.cei.lmu.de`

[20] A repository documenting the proposed extension can be found at `https://github.com/GVogeler/CEI2TEI`.

beginning, semantic technologies have been at the core of the strategy for interoperability and interconnectivity with other resources. Experience from the workflows described above has been applied and compatibility with well-established projects in the field—such as the Linguistic Linked Open Data Cloud [15]—is an essential objective.

AAIF. Open Access Database for Adjective-Adverb Interfaces in Romance[21] is a project focusing on the production of reusable linguistically annotated data which serves as a prototype for corpus linguistics projects in the GAMS ecosystem. Investigating the complex relationships of adverbs and adjectives in Romance languages, an ontology was developed to describe the various morphosyntactic, syntactic, and semantic attributes of this linguistic phenomenon. The ontology controls the annotation model and serves as a common conceptual model within the research group [16]. A linking model to the Ontologies of Linguistic Annotation (OLiA) provides interoperability with popular annotation models used for linguistic annotation [17]. The annotated data is converted into the CoNLL-RDF format [18] in order to facilitate interoperability with, e.g. lexicographic resources. These exchange formats make the data available for reuse in other corpus linguistics projects.

4. Conclusion

Long-term digital repositories solve the archiving need of making materials accessible in a stable format over the long haul, but archiving, divorced of context and interpretation, does not reach the full potential of the digital services and methods used [19]. Ontologies, and the use of semantic web technologies to expand the reach and interoperability of the data being preserved, are important elements in ensuring that projects are not only preserved for the future, but continue to be useful and accessible contributions to research. Accordingly, digital repositories, which provide crucial infrastructure for the deployment of linked resources, are increasingly a necessity for humanities work. The experience of the GAMS shows the potential but also some of the difficulties of ontology-driven work in the humanities. While some projects deal with cleanly sorted data that is already well-described by science (e.g. *CoReMa* and ingredients), others are faced with having to make computer-readable assertions about messy human situations (e.g. *RTA 1576* and communications actions). Ontologies have to be appropriate to the requirements of the specific project, but also universal enough in their design to accommodate re-use and linking of data. Humanities data requires domain-specific interpretation that is often highly disciplinary and frequently ambiguous, but the ability to find relevant ways to model data across documents and projects makes the formerly incomparable comparable and facilitates tackling larger projects, larger datasets, and more productive reusability and portability of data than has historically been the norm.

References

[1] K. Diwisch, et al., Managing Cultural Assets: Challenges for Implementing Typical Cultural Heritage Archive's Usage Scenarios, in: *Semantic Applications: Methodology, Technology, Corporate Use*, edited by T. Hoppe, B. Humm, and A. Reibold, 2018, 219–230, https://doi.org/10.1007/978-3-662-55433-3_15.

[21] https://gams.uni-graz.at/context:aaif

[2] T. Berners-Lee, Linked Data, in: *W3C Design issues*, 2006/2009, https://www.w3.org/DesignIssues/LinkedData.html.

[3] G. Vogeler, The Content of Accounts and Registers in Their Digital Edition: XML/TEI, Spreadsheets, and Semantic Web Technologies, in: *Konzeptionelle Überlegungen zur Edition von Rechnungen und Amtsbüchern des späten Mittelalters*, edited by J. Sarnowsky, Göttingen, V&R unipress, 2016, 13–41.

[4] J.Stigler, and E. Steiner, GAMS — Eine Infrastruktur zur Langzeitarchivierung und Publikation geisteswissenschaftlicher Forschungsdaten, in: *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen Und Bibliothekare* 71:1, 2018, 207–216, https://doi.org/10.31263/voebm.v71i1.1992.

[5] E. McLellan, General Study 11 Final Report: Selecting Digital File Formats for Long-Term Preservation, version 1.1, *The InterPARES 2 Project*, 2007, http://www.interpares.org/display_file.cfm?doc=ip2_file_formats(complete).pdf.

[6] E. Steiner, and J. Stigler, Dokumentation, 2017, *GAMS: Geisteswissenschaftliches Asset Management System*,http://gams.uni-graz.at.

[7] C. Pollin, and G. Vogeler, Semantically Enriched Historical Data: Drawing on the Example of the Digital Edition of the 'Urfehdebücher Der Stadt Basel', in: *Workshop on Humanities in the Semantic Web. Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) co-located with 16th International Semantic Web Conference (ISWC 2017)*, edited by A. Adamou, E. Daga, and L. Isaksen, 2017, 27–32, http://ceur-ws.org/Vol-2014/paper-03.pdf.

[8] R. H. Dekker, and D. J. Birnbaum, It's more than just overlap: Text As Graph, in: *Proceedings of Balisage: The Markup Conference 2017*, Balisage Series on Markup Technologies 19, 2017, https://doi.org/10.4242/BalisageVol19.Dekker01.

[9] G. L. Geerts, and W. E. McCarthy, The Ontological Foundation of REA Enterprise Information Systems, in: *2000 AAA National Meeting*, Alabama, 2000, https://www.msu.edu/user/mccarth4/Alabama.pdf.

[10] A. Macks, Corpus Kalendarium, 2015–2019, http://www.cokldb.org/.

[11] R. Ribeiro, F. Batista, J. P. Pardal, N. J. Mamede, and H. S. Pinto. Cooking an Ontology, in: *Artificial Intelligence: Methodology, Systems, and Applications*, edited by J. Euzenat, and J. Domingue, 2006, 213–221.

[12] M. Sam, et al., An Ontology Design Pattern for Cooking Recipes — Classroom Created, 2014, http://ontologydesignpatterns.org/wiki/images/0/01/Paper_4.pdf

[13] G. Vadivu, and S. W. Hopper, Semantic Linking and Querying of Natural Food, Chemicals and Diseases, in: *International Journal of Computer Applications* 11:4, 2010, 35–38, https://doi.org/10.5120/1567-2093.

[14] M. Lanzinner, and A. Strohmeyer, eds., Der Reichstag 1486–1613: Kommunikation — Wahrnehmung — Öffentlichkeit, Göttingen, 2006.

[15] J. P. McCrae, et al., The Open Linguistics Working Group: Developing the Linguistic Linked Open Data Cloud. in: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016, 2435–2441, http://www.lrec-conf.org/proceedings/lrec2016/pdf/851_Paper.pdf.

[16] C. Pollin, G. Schneider, K. Gerhalter, and M. Hummel, Semantic Annotation in the Project "Open Access Database 'Adjective-Adverb Interfaces' in Romance", in: *Proceedings of the Workshop on Annotation in Digital Humanities*, CEUR Workshop Proceedings, edited by S. Kübler, and H. Zinsmeister, 2018, 41–46, http://ceur-ws.org/Vol-2155/pollin.pdf.

[17] C. Chiarcos, and M. Sukhareva, OLiA – Ontologies of Linguistic Annotation, in: *Semantic Web*, 6:4, 2015, 379–386, http://semantic-web-journal.net/system/files/swj518_0.pdf.

[18] C. Chiarcos, and C. Fäth, CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way, in: *Language, Data, and Knowledge. LDK 2017*, Lecture Notes in Computer Science, 10318, edited by J. Gracia, et al., 2017, 74–88, https://doi.org/10.1007/978-3-319-59888-8_6.

[19] G. Vogeler, The 'assertive edition': On the consequences of digital methods in scholarly editing for historians, in: *International Journal of Digital Humanities* 1:2, 2019, 309–322 https://doi.org/10.1007/s42803-019-00025-5.