Sustainable Semantics for Sustainable Research Data

Steffen Hennicke[1,*], Pascal Belouin[1], Hassan El-Hajj[1,2], Matthew Fielding[3], Robert Casties[1] and Kim Pham[1]

[1]*Max Planck Institute for the History of Science, Boltzmannstr. 22, Berlin, 14195, Germany*

[2]*BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, 10587, Germany*

[3]*Takin.solutions, 36 Koprivshtitsa Str. Plovdiv 4002 Bulgaria*

Abstract

In view of the steadily growing volume of digital output from Humanities research projects in recent decades, the question of the long-term and sustainable preservation of this research data is becoming increasingly urgent. To meet this challenge, we are establishing the Central Knowledge Graph (CKG) as a key element of our documentation and publication strategy for research data. In this paper, we present two of the cornerstones of this strategy: The newly developed Project Description Layer Model (PDLM) provides the means to document the required contextual metadata about research projects and their digital outputs; the Zellij Semantic Documentation Protocol systematically documents the modeling patterns used to create CIDOC CRM representations of project data in a transparent and reusable way.

Keywords

CIDOC CRM, knowledge graph, semantic modelling, semantic documentation, research data management

1. Introduction

The Max Planck Institute for the History of Science (MPIWG) can look back on a long tradition of digital scholarship. Since its foundation in the 1990s, the MPIWG has been able to build up an extensive portfolio of digital offerings, including extensive digital libraries such as ECHO or Digital Libraries Connected (DLC), and research databases created by individual research projects, such as the Islamic Scientific Manuscripts Initiative (ISMI), Sphaera, or Commoning Biomedicine.[1] An increasingly pressing problem, however, is the question of how to deal with the decay of the usability and accessibility of digital offerings and the data they contain after a project has ended.

To address these challenges, we are working on an institutional research data management strategy that both adequately documents the digital output of our research projects and preserves

---

[1]ECHO: https://echo.mpiwg-berlin.mpg.de/home, DLC: https://dlc.mpg.de/index/, ISMI: https://ismi.mpiwg-berlin.mpg.de, Sphaera: http://db.sphaera.mpiwg-berlin.mpg.de/resource/Start, ComBio: https://combio.mpiwg-berlin.mpg.de (all 01.03.2024)

it in a form that makes valuable data available and reusable in the long term. One of the cornerstones of our preservation strategy is a graph database, the Central Knowledge Graph (CKG), where we document and publish data from our research projects as Linked Open Data with CIDOC CRM[2] as the common target model. The CIDOC CRM plays a significant role in enabling FAIR [1] data representation, in particular by providing a semantically well-defined vocabulary for describing cultural heritage data. Our principle approach provides for research data to be gracefully degraded by mapping and converting it to a CIDOC CRM based Linked Data representation.

However, when it comes to the long-term reusability of the data published in the CKG and the sustainability of our preservation strategy, we have identified two additional semantic challenges: The first is that we need to provide enough context about the published instance data in the CKG so that researchers can confidently assess its provenance and relevancy; the second is that we need to document the semantics used in the data modeling at a schema level and in a way so that common modeling patterns become transparent and easily reusable.

For the purpose of documenting research projects and their digital outputs, we have developed the Project Description Layer Model[3] (PDLM). The PDLM is a semantic model based on CIDOC CRM and the Parthenos Entities Model (PEM) [2] for describing the context of research projects and the provenance of their digital outputs. We consider these contextual metadata about projects and their digital outputs just as important as the research data itself, and therefore decided to store and record these contextual data together with the research data in the CKG and in the same semantic target model, the CIDOC CRM. This way, the CKG consists of two conceptual layers, the project description layer realized by the PDLM, and the project data layer which holds CIDOC CRM representations of the original research data.[4]

In the same vein, the comprehensive semantic documentation of modeling patterns used in the creation of project data is a key component of our long-term institutional research data preservation and publication strategy. On the one hand, researchers working with data from the CKG need to be able to clearly understand the origins and context of the data, but also the semantics at the schema level, such as their ontological scope and intended meaning. On the other hand, with regard to mapping and converting research data to the CIDOC CRM, the ability to reuse existing and proven modeling patterns is a prerequisite for efficient and semantically aligned data. To that extent, semantic documentation is required as reference for confident reuse of existing data and for efficient creation of new data for the CKG.

In this paper, we discuss how to address these two challenges. We present the Project Description Layer Model (PDLM), a CIDOC CRM compliant model, that we have developed to describe the context of the project data stored in the CKG in Section 2. We then present our approach to the sustainable semantic documentation of the modeling patterns used in the CKG, the Zellij Semantic Documentation Protocol, taking the PDLM as a leading example in Section 3. Finally, we highlight our current data transformation, testing, and serving strategies applied to what we call the Legacy Project at the MPIWG in Section 4.

---

[2]https://cidoc-crm.org (06.02.2024)

[3]https://github.com/mpiwg-research-it/drih (07.03.2024)

[4]In this paper, we are focusing on the project description layer. However, the principles outlined here apply to the project data layer just the same.

2. Project Description Layer Model

The Project Description Layer Model (PDLM) is a semantic model based on CIDOC CRM for describing research projects and the provenance of their digital outputs. The documentation of digital objects serves the purposes of our institutional research data management strategy where we keep track of active and archived digital research outputs of our projects. The documentation of research projects, on the other hand, serves to appropriately contextualize the digital objects so that researchers can assess the relevancy and provenance of data in the CKG, while also creating a record of the digital research conducted at the MPIWG.

The PDLM is heavily based on the Parthenos Entities Model (PEM), an ontology that was developed in the context of the Parthenos project[5] (2015 - 2019) to conceptually integrate digital services and e-infrastructures from the Humanities into a larger research infrastructure. The PEM provided well thought-out conceptualizations for a domain of interest largely congruent to the one we intended to represent. For this reason, we developed the PDLM from the conceptualizations provided by the PEM, such as the concept of a project or service, using a subset of the PEM's original set of entities types and relations.

Generally, the main entities types that are required to establish the necessary context and that we consider pivotal to our domain of interest are (1) digital objects, which include datasets and software, (2) activities, which include research and service projects, types of services provided by projects, and the creation and modification of digital objects, and (3) actors, which include persons, groups, and project teams that carry out activities such as projects or services.

2.1. Digital Objects

Digital objects are the digital outputs that research projects create and modify and that are curated and hosted as part of their activities. In the context of the PDLM, we distinguish between datasets and software. Datasets are "identifiable immaterial items that can be represented as sets of bit sequences and whose content contains propositions about the objective world" [3]. The concept of a dataset is rather inclusive where typical examples include complex aggregates such as databases and research websites, static-HTML archives, the CKG, or a repository on GitLab, but also individual data files such as image files, text documents or structured data files. Software, on the other hand, are specifically "software codes, computer programs, procedures and functions that are used to operate a system of digital objects" [4]. Typical examples are specific software applications such as Word, ResearchSpace, or X3ML[5], scripts for data conversion, algorithms for topic modeling, but also formal schemas such as CIDOC CRM or Dublin Core.

Furthermore, we distinguish between volatile and persistent digital objects, which allows us to track those digital scholarly products that are under scholarly investigation and may potentially change at any time, and those digital scholarly products that are stable and final outcomes of scholarly investigation. When assessing available data through the CKG this distinction is crucial for users that may want to reuse these data. Furthermore, and in line with the requirements of institutional research data management strategy, the distinction allows

---

[5]http://www.parthenos-project.eu (29.02.2024)

us to specifically record those final, persistent snapshots of digital scholarly products that are being archived and published as research data.

2.2. Activities

Projects are activities that represent a "collaborative enterprise undertaken over a period of time (…) with the intention of effectuating some defined programme" [3]. While the PEM makes no further distinction with regard to the type of projects, we introduced two new sub-classes to projects, research projects and service projects, which we consider key concepts to the practical scope of our domain of interest. A research project is a scholarly undertaking of individual researchers or of project teams that are carried out at or with the participation of members of the MPIWG and that can create instances of digital objects. We record research projects that have ended in terms of their official project duration and we track research projects that are still active and have not yet reached their official end date. A service project, in contrast, acts in the primary role of a provider of a service that is used, but not offered, by a research project, for example, as the provider of a research data repository where a research project archives its research data. A service project is not a scholarly undertaking or its primary purpose is not to conduct research, though it may produce digital objects as part of its overall program.

Services are the second type of activity we document. They are "declared offers by some instance of E39 Actor of their willingness and ability to execute an activity or series of activities at the request of another instance of E39 Actor for the specific benefit of the latter" and "include all auxiliary abilities of the same actor to execute the respective activities" [3]. The service model offered by the PEM defines curation and hosting and the provision of e-services as three high-level classes, which have nine specialized sub-classes. After some initial modeling tests, we found that the original conceptualization of services in the PEM was not sufficient for our purposes. We therefore decided to extend the original ontological structure by also defining the two high-level service classes for digital hosting (PE5) and digital curating (PE10) as sub-classes of e-service (PE8). These two classes cover the two essential questions to our domain of interest: who holds the data or software, which is the actor who provides the digital hosting service, and who works with the data or software, which is the actor who provides the digital curating service.

Lastly, we include as activities "digital machine events" [4] (DME) that represent the creation context of digital objects, i.e. activities of creation or modification of digital objects, such as the generation of a static (persistent) version of a (volatile) research website, or the mapping, conversion, and ingestion of a CIDOC CRM representation of an original project dataset. By making such activities explicit, we can document for a particular digital object which researchers or projects supported or participated in its creation, when the creation took place, which data was used in the creation or, in the case of derivative digital objects, from which project the digital object conceptually originates.

2.3. Actors

As the third main category of documented entity types, actors carry out activities and are divided into project teams, groups and individuals. Project teams generally represent groups of actors,

typically human individuals, that join together with the shared will to support and maintain a specific project and its aims. As such, project teams are unique and bound to the existence of a particular project: they typically come into existence with inception of the related project and end when the project ends. By contrast, groups represent all other gatherings of actors that exhibit more lasting organizational features and whose existence is not bound to one particular project. Generally, we distinguish between internal groups, such as departments of the Institute, research groups, or service units, and external groups, such as the Max-Planck-Society, or the *Deutsche Forschungsgemeinschaft*[6] (DFG). Persons are human individuals that, in the context of the PDLM, must be the member of at least one group or project team in order to establish a minimal context for that person through its group affiliation. As a member of a project team, a person is considered to have participated, at some point, in the project maintained by that project team.

With the current version of the PDLM, we have created a core model for documenting the context of projects and the provenance of their digital outputs. The metadata recorded by the PDLM is considered essential research data and is as much part of the CKG as the CIDOC CRM representations of project data. The ontological model of the PDLM has been developed and documented using the Zellij Semantic Documentation Protocol, which constitutes the second pillar of our approach to sustainable research data.

3. Zellij Semantic Documentation Protocol

As noted above, the locus of our semantic documentation rests upon a series of core entities: Persons, Project Teams and Groups, Volatile and Persistent Datasets and Softwares, Service and Research Projects, along with Digital Machine Events for tracking the creation and/or modification of digital objects. Once such a basic list of entities has been proposed, a standard approach to their documentation within the domain must be determined in order to provide a non-arbitrary list of the properties required to describe those entities. Typically, source databases form the foundation for a bottom-up formulation of the model, the function of which is then to: a) deduce the basic properties of interest regarding those entity classes and b) propose a standardized semantic representation for the entities and the set of properties as they are to be applied to them. As such, this method closely followed the basic strategy of formal ontology development [6] in general, with regards to faithfulness. It differs, however, in that it does not seek to exhaustively categorize every possible entity within the domain for its own sake. It rather aims to isolate and provide a generalized set of properties for those entities that are explicitly addressed in the documentation, while remaining open to reuse and extension as required.

To document the semantic patterns determined in this process we used an in-house semantic pattern documentation protocol called Zellij, developed at Takin.solutions.[7] The purpose of this protocol is to provide a stable and sustainable repository for the semantic patterns deployed in the model, in a manner that facilitates their subsequent reuse and continued development

---

[6]https://www.dfg.de/en (15.04.2024)

[7]Cf.   presentation   on   the   Zellij   protocol   at   https://www.cidoc-crm.org/Resources/ zellÄńj-a-semantic-pattern-development-and-documentation-system (15.04.2024).

over time, both within a single organization and across partner institutions, and thus that also speaks to a variety of users with diverse technical capacities. This is achieved by breaking the full knowledge graph up into modular pieces, which allows the semantic patterns to be created, modified, and reused across the domain in question, as well as to be inspected *in situ*, where they serve to exposit particular entities and their potential relations to each other.

The backbone of this protocol is a triptych of relational databases, currently provided by Airtable[8], which, by modularizing the essential elements of the semantic model, facilitates the reuse and redeployment of the semantic patterns that have been defined. Presently, this triptych is made up of three, interrelated bases (cf. Figure 1:a): 1. The Field Base, 2. the Collection Base, and 3. the Model Base, each of which comes with a suite of metadata specifications to support their functionality. The Field Base, for example, serves as the 'library' of unique semantic patterns that are to be deployed across the model in various constellations, and as such serve as the basis of the desired sustainability. The Collection Base groups some of these fields together, insofar as they are intended to capture common, collective, and uniform semantic build outs from a given node anywhere in the model; timespans on event nodes are an example of such common, collective and uniform semantics that differ little (if at all) across their deployment within the model. In The Model Base, fields and collections are joined with the core entities of interest, which we call 'Reference Entities', in order to create a series of modular 'Reference Models' that determine the scope of the semantic expressivity of the overall knowledge graph and represent it in a piecemeal manner.

Key to this protocol is the attribution of a unique identifier to each of the semantic patterns defined. Giving each semantic pattern a unique identifier allows for the reuse of previously defined fields in varied contexts. A field used to describe a given Reference Entity can be transported to another Reference Entity, so long as the ontological scope satisfies, and the identifier clearly indicates where a particular pattern has been reused throughout the knowledge graph. The ontological consistency of the whole is thus reinforced and large areas of data can be accurately covered by a small subset of basic semantic pathways deployed in various constellations.

In the case of the PDLM, for example, we defined a number of core metadata patterns, which could be applied across all entity types uniformly. These include, e.g., semantics for attributing names, identifiers and identifier types, entity descriptions, and digital reference fields pointing to URIs. With this, potentially heterogeneous semantic patterns for documenting the desired fields are standardized and consistently applied across the complete knowledge graph. This standardization process applies also to smaller subsets of Reference Models in accordance with the ontological scope of the Reference Entities determined at the outset. For example, service projects and research projects are both subtypes of CRM E7 Activity, which allows us to apply to them semantic patterns related to their temporality and to link them to the various actors that carry them out, along with the roles those actors play there, etc., via the inheritance of E7 Activity properties. The unique articulation of these patterns in the Field Base ensures standardized deployment across the relevant Reference Models, enhancing the coherence of the model as a whole and the validity of search results.

Employing such a basic documentation protocol to the semantic models themselves provides

---

[8]https://airtable.com (15.04.2024)

an efficient means by which to integrate new data into extant models or generate new models as necessary, through the deployment of previously defined fields in new data constellations. The Reference Models themselves, which many people are inclined to consider the most challenging part of semantic modeling, actually have a rather small set of defining parameters that distinguish them from one another other, as the bulk of the work comes from deciding which fields to populate the model with in order to represent the entity in question, which necessitates a high degree of reuse and redeployment. In this way, the complete knowledge graph can be built up out of distinct, modular pieces, allowing each to be easily inspected, reused or extended as required by this or future projects.
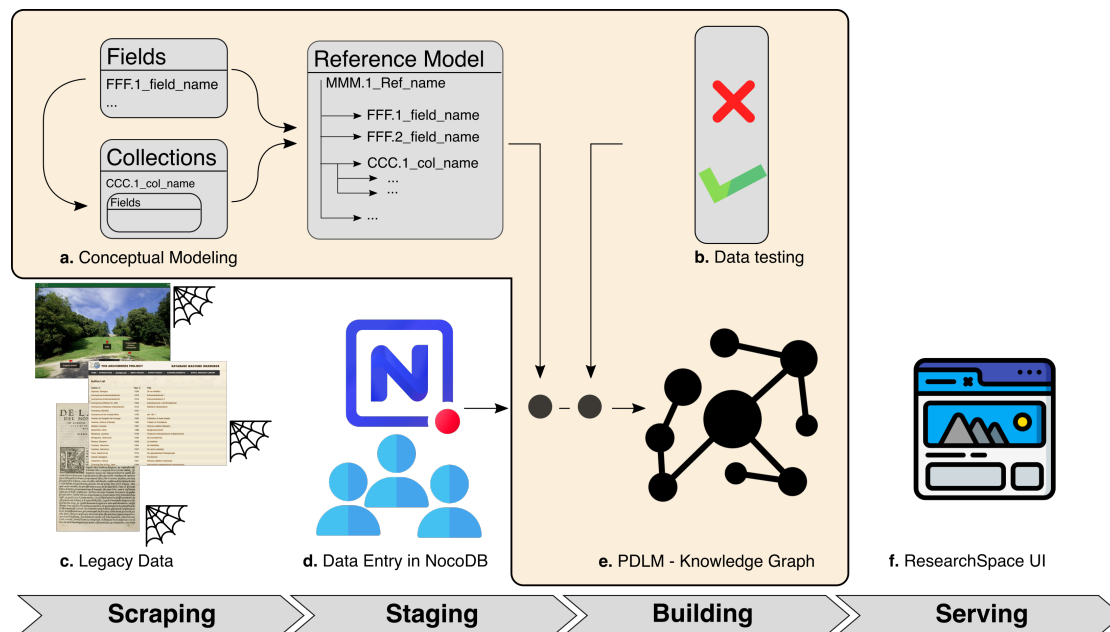
Figure 1: Pipeline showing the different steps undertaken from data conceptualization to serving. **(a)** Data modeling and conceptualization using Airtable and Zellij; **(b)** Unit tests used to ensure that the data is compliant with the PDLM scheme; **(c)** Heterogeneous Legacy data including websites, databases, image and text collections; **(d)** Data entry in NocoDB, **(e)** Moving the data from NocoDB to a Knowledge Graph structure after the data has been modeled **(a)** and passed all relevant tests **(b)**; **(f)** Serving the Knowledge Graph data to the MPIWG research community using the ResearchSpace platform as a UI.

4. Use Case: Legacy Research Projects

We are currently testing the first version of the PDLM and our preservation strategy in a use case centered on the Institute's digital legacy. These legacy projects and their data are an important resource for the MPIWG and the history of science due to their wealth of information[7]. To name just two examples: The recently completed *Geschichte der Max-Planck-Gesellschaft*[9] (GMPG) project collected extensive reference data on the history of the Max-Planck-Society,

---

[9]https://gmpg.mpiwg-berlin.mpg.de/en/ (01.03.2024)

which is a unique resource in this respect, or the research data on Immanuel Kant collected in various projects becomes relevant again in view of the *Kant Year 2024*, the anniversary year to mark his 300th birthday. Many of these older projects and their data, however, are hardly usable for new studies since their technical stack has become heavily outdated and no longer maintainable mainly due to the nature of research funding which rarely provides support beyond the lifetime of a project.

The aim of our overall preservation strategy is to enable and promote the reuse of research data. To this end, we aim to convert the digital output of legacy projects into a sustainable and standardized form that preserves as much of the original functionality and presentation as possible. In addition, we map and convert selected data into a CIDOC CRM representation published in the CKG.[10] As shown in Figure 1, we undertake a multi-step approach which starts by scraping and crawling the Web-based components of legacy projects, followed by a data-staging phase where the data is checked and modeled before passing it through a rigorous PDLM compliant testing phase and finally serving it to the clients as linked data through the Digital Research Infrastructure for the Humanities (DRIH) front end based on the open source platform ResearchSpace[11].

One of the major technical hurdles we faced in our efforts to preserve these legacy projects is their heterogeneity, with some of these projects built as static HTML pages, others built with Python-based web frameworks, as well as often deprecated collection management systems (see Figure 1:c). Due to this heterogeneity, we decided to transform all these projects to their simplest form, static HTML, and store those for long term preservation. In some cases, where turning a project into a static form is not feasible, we also attempt to extract structured data. Any available object data, such as images or audiovisual files, are stored alongside the archived static versions of the original legacy project.

We also focused on extracting relevant information for the PDLM such as copyrights, institutional affiliations, and research topics. These project metadata are entered by a dedicated team of student assistants into NocoDB[12], a flexible and user-friendly open-source relational database (see Figure 1:d). To manage, curate, and transform this project metadata into triple data compliant with the PDLM, we designed a pipeline which starts with a Python script that retrieves the data stored in NocoDB via its API. Making extensive use of the RDFLib Python library, this script generates detailed compliance reports by running a dynamic test suite, which validates the generated triples against a set of SPARQL queries based on the PDLM rules stored in Zellij. It also produces a number of RDF data files in various formats for easy inspection. Finally, this script can remove PDLM-related triples from a specified ResearchSpace instance before uploading the newly created triples. Our goal when building this pipeline was to focus on code reusability and extensibility. Thus, a large part of the code responsible for generating PDLM-compliant patterns has been modularized in a self-contained python library, which we aim to release as an open-source software package in the near future.[13]

---

[10] In our current use case, we are solely focusing on the conversion of data into a sustainable form and the corresponding documentation of the projects and their digital outputs with the PDLM; the mapping and conversion of the project data into the CIDOC CRM will only be the next step.

[11] https://researchspace.org (01.03.2024)

[12] https://www.nocodb.com (07.03.2024)

[13] https://github.com/mpiwg-research-it/drih (07.03.2024)

The final stage of our pipeline is to provide a clear and modern user interface for researchers to search and explore the metadata about our research projects and their digital outputs, captured using a unified schema, the PDLM, and directing them to where digital objects are now accessible, be they archived representations, still active instances or CIDOC CRM representations within the CKG. In the current proof-of-concept version, users can query and navigate the metadata for our legacy projects. Based on their feedback and the experience gained from the current use case of the legacy projects, we will revise the PDLM and further expand the functionality of the DRIH platform.

5. Conclusion

We consider the sustainable documentation of semantics one of the most important challenges and prerequisites when it comes to the management of research data at the institutional level that supports transparency and reuse of research data in the long term. In this paper, we have reported on our ongoing efforts to address these challenges by developing the Project Description Layer Model (PDLM) for the documentation of contextual information about research projects and their digital outputs and by applying the Zellij Semantic Documentation Protocol to the documentation of semantic modeling patterns.

Whilst we are currently mainly working through legacy projects as part of building and testing a proof-of-concept implementation of our Central Knowledge Graph (CKG), we are also planning to implement strategies that will enable us to work towards mapping and converting project data to CIDOC CRM from the very beginning of a project. Key to this strategy is the elaboration and documentation of common modeling patterns with the Zellij Semantic Documentation Protocol. Building up a treasure trove of semantic modeling patterns in Zellij will ensure that future mapping and conversion efforts will gain in efficiency.

At the same time, with ingesting increasing quantities of research data from projects as CIDOC CRM representations into the CKG, we will have to build an additional abstraction layer on top of the project data, in the sense of Fundamental Categories and Relations [8], that serves as additional access layer. With Zellij, and our experiences gained from the development of the PDLM, we believe, we are well prepared for systematically and sustainably documenting the emerging modeling patterns.

References

[1] M. D. Wilkinson, M. Dumontier, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:`10.1038/sdata.2016.18`.

[2] G. Bruseker, M. Doerr, M. Theodoridou, Report on the Common Semantic Framework, D5.1, 2017.

[3] FORTH-ICS, Parthenos Entities: Research Infrastructure Model DRAFT, V3.1, 2017.

[4] M. Doerr, M. Theodoridou, S. Stead, Definition of the CRMdig. An Extension of CIDOC-CRM to Support Provenance Metadata (3.2.1), 2016.

[5] Y. Marketakis, N. Minadakis, H. Kondylakis, K. Konsolaki, G. Samaritakis, M. Theodoridou, G. Flouris, M. Doerr, X3ML Mapping Framework for Information Integration in Cultural Heritage and Beyond 18 (2017) 301–319. doi:10.1007/s00799-016-0179-1.

[6] A. Gangemi, V. Presutti, Ontology Desgin Patterns, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems, Springer, 2009. doi:10.1007/978-3-540-92673-3_10.

[7] I. Milligan, Lost in the infinite archive: The promise and pitfalls of web archives, International Journal of Humanities and Arts Computing 10 (2016) 78–94. doi:10.3366/ijhac.2016.0161.

[8] K. Tzompanaki, M. Doerr, Fundamental Categories and Relationships for Intuitive Querying CIDOC-CRM Based Repositories, 2012.