TAO: a document- and person-centric ontology for storing rich metadata of manuscripts

Luiz do Valle Miranda[1,*], Jakub Gomułka[2], Krzysztof Kutt[1] and Grzegorz J. Nalepa[1]

[1]*Jagiellonian Human-Centered AI Lab, Mark Kac Center for Complex Systems Research, Institute of Applied Computer Science, Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University, prof. Stanisława Łojasiewicza 11, 30-348 Kraków, Poland*

[2]*Faculty of Humanities, AGH University of Krakow, Czarnowiejska 36, 30-054 Kraków, Poland*

Abstract

We present TAO (The Arcarium Ontology), an ontology that will serve as the backbone of the forthcoming Arcarium knowledge base for storing and retrieving rich metadata of manuscript collections housed at Jagiellonian University (Kraków, Poland). The main motivation for its creation is the ongoing digitization of the precious Autograph Collection from the so-called Berlin Collection, however, the resulting solution is intended to be as universal as possible to allow metadata acquisition in a wide variety of projects. Therefore, the proposed ontology, in addition to the document properties collected by experts analyzing the digitized collection, also includes concepts and relationships from the Europeana Data Model (EDM), CIDOC Conceptual Reference Model (CIDOC CRM), Records in Contexts Ontology (RiC-O), Encoded Archival Description (EAD) and Gemeinsame Normdate (GND), thus enabling integration with external knowledge bases and facilitating advanced research scenarios in the digital humanities. Although still a work in progress, yet the ontology has already been fed with the first real instances—the documents described in the digitization pilot—and discussed with an international panel of experts, which allowed its critical evaluation and indication of its usefulness for the Arcarium system under development.

Keywords

Ontology, OWL, Metadata, Cultural Heritage, Digital Humanities, Manuscripts, Autographs

1. Introduction

The Jagiellonian University (Kraków, Poland; JU) and the Jagiellonian Library (JL), which is part of it, have many cultural heritage artifacts in their collections, including numerous manuscript collections comprising many unique objects such as the oldest Polish documents, "De revolutionibus orbium coelestium" by Nicolaus Copernicus and a large collection of Mozart autographs [1]. Most of these objects have already been cataloged, and some have been additionally digitized, but in both cases users receive only a basic description in either the MARC 21

Proceedings of the Joint Ontology Workshops (JOWO) - Episode X: The Tukker Zomer of Ontology, and satellite events co-located with the 14th International Conference on Formal Ontology in Information Systems (FOIS 2024), July 15-19, 2024, Enschede, The Netherlands.

*Corresponding author.

luiz.dovallemiranda@doctoral.uj.edu.pl (L. do Valle Miranda); jgomulka@agh.edu.pl (J. Gomułka); krzysztof.kutt@uj.edu.pl (K. Kutt); gjn@gjn.re (G. J. Nalepa)

https://krzysztof.kutt.pl/ (K. Kutt); https://gjn.re/ (G. J. Nalepa)

0000-0003-1838-5693 (L. do Valle Miranda); 0000-0002-9100-0334 (J. Gomułka); 0000-0001-5453-9763 (K. Kutt); 0000-0002-8182-4225 (G. J. Nalepa)

standard (interlibrary catalog; https://katalogi.uj.edu.pl/) or Dublin Core (Jagiellonian Digital Library; https://jbc.bj.uj.edu.pl/). However, if one would like to store richer metadata, the systems currently available at JU and JL do not provide such a possibility. This shortage is addressed by two of JU's current so-called flagship projects[1], which aim is to create dedicated infrastructure and establish international cooperation to improve the university's research ecosystem.

The first of these projects, *European Heritage in the Jagiellonian Library: Digital Authoring of the Berlin Collections* (DiHeLib; https://dihelib.id.uj.edu.pl/en_GB/), is oriented around the Autograph Collection (Autographa Sammlung) of the so-called "Berlin Collection". It consists mainly of various types of documents (mostly letters) associated with historical figures, scholars, and various European officials. The collection was kept in the former Prussian State Library until World War II, but is currently divided between the State Library of Berlin and the Jagiellonian Library (the history of the collection is described in [2]). The German part has been cataloged and can be searched through the Kalliope catalog (https://kalliope-verbund.info/en/). The Polish part has been kept in storage for a long time [2] and will only be cataloged and digitized in the ongoing DiHeLib project [3]. The description and digitization of such a large and valuable collection provides a sound test-bed for developing and experimenting with solutions to the infrastructural deficiencies identified earlier. To date, among the contributions of the DiHeLib project, we should point out the development of a digitization workflow extending the existing JL's processes to include the acquisition of rich metadata [4], and the proposal of a preliminary idea for the Arcarium system (https://dihelib.id.uj.edu.pl/en_GB/arcarium), which will complement the JU's ecosystem by offering the ability to store and search rich metadata and to conduct specialized research in the area of digital humanities.

The purpose of this paper is to present The Arcarium Ontology (TAO) – an OWL model which will form the basis of the planned Arcarium system. While developing TAO, we attempted to maintain a balance between a model allowing for a very detailed description of the "Berlin Collection" under digitization, and a highly flexible solution, which can be easily extended to other use cases, such as documents related to the history of the Jagiellonian University explored in the second flagship project: *Cultural Heritage Exploration and Retrieval with Intelligent Systems at Jagiellonian University* (CHExRISH).

The remainder of the paper is structured as follows. Sect. 2 summarizes the most relevant related works. Sect. 3 highlights the approach for TAO development and the requirements it was intended to address. The TAO ontology is outlined in Sect. 4, while examples of its use can be found in Sect. 5. Sect. 6 concludes the paper.

2. Related works

A significant step in the development of an ontology to meet the needs of a given institution is the definition of target platforms for integration and the extent of shared vocabularies. In the context of the development of TAO, four ontologies preliminarily appear in such a position are Europeana Data Model (EDM) [5], CIDOC Conceptual Reference Mode (CRM), and the Records in Contexts Ontology (RiC-O) and the Gemeinsam Normdatei (GND) ontology.

[1]See https://id.uj.edu.pl/en_GB/projekty-flagowe for more details.

Europeana is a digital cultural platform integrating resources from more than 3000 institutions concerning European cultural heritage. Europeana's metadata representation is based on the native schema Europeana Data Model (EDM) [5]. It is a flexible framework within the GLAM (galleries, libraries, archives, and museums) realm that embraces various community standards such as LIDO for museums [6], EAD for archives [7], or METS for digital libraries [8]. While providing a structured way to present metadata, EDM also facilitates enhancement through external sources. It is based on linked data principles and is structured around a knowledge graph, which is provided as a `.owl` file without specific examples. In EDM, there are three key classes [5]: the cultural heritage object itself (`edm:ProvidedCHO`), its digital representation (`edm:WebResource`), and the aggregation that connects the two (`ore:Aggregation`). An illustration of how EDM represents the Mona Lisa painting can be found in Fig. 1.

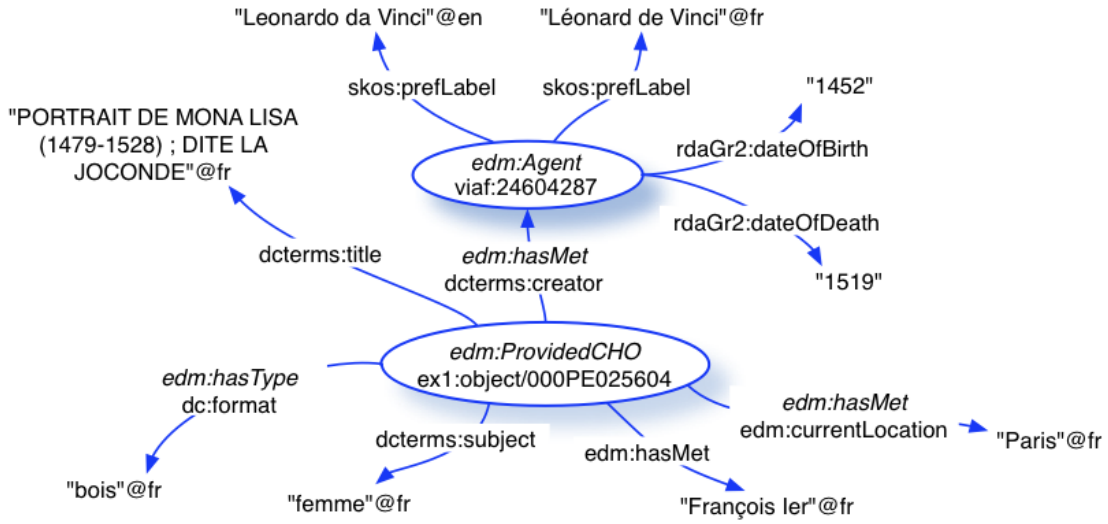

Figure 1: Mona Lisa – an object-centric description enriched with an Agent contextual entity [5].

Ontologies that focus on organizing knowledge around objects or entities and their attributes are called object-centric ontologies. As seen in Fig. 1, one way to represent cultural heritage information according to EDM is precisely in an object-centric manner. However, EDM also allows for an event-centric approach, that is, a method of organizing information or describing objects by focusing on the events in which those objects have been involved. The most prominent example of an event-centric ontology is CIDOC CRM.

CIDOC CRM is a conceptual framework used in the cultural heritage sector to organize and describe information about objects, events, and concepts. CIDOC CRM is a comprehensive, adaptable, and flexible model that has been widely used by various institutions, including museums, libraries, and archives. The central concept in CIDOC CRM is event, which provides a robust foundation for representing cultural contexts through human activities. This approach offers expressive capabilities by capturing dynamic and detailed records, allowing for the representation of cultural heritage objects via the processes that it has undergone or witnessed. By simplifying metadata processing and improving visualization, CIDOC CRM enhances precision in documenting cultural heritage. An illustration of how a sculpture is represented in CIDOC

CRM can be seen in Fig. 2.

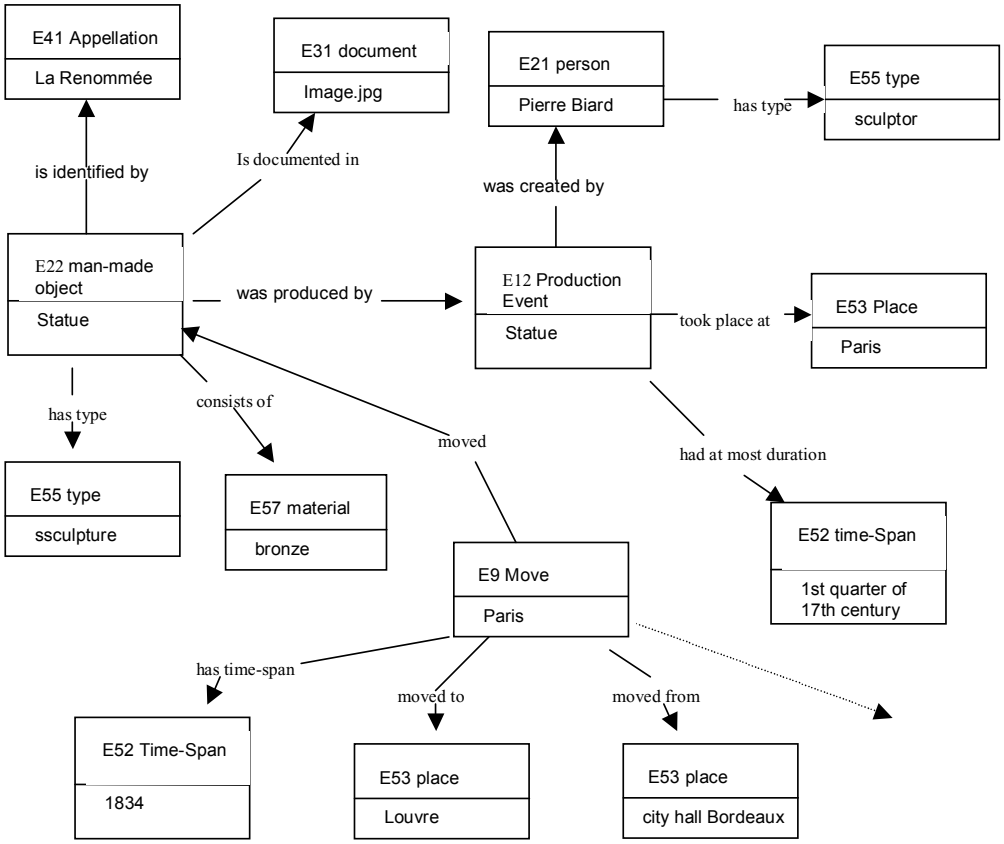

Figure 2: Example of a sculpture modeling in CIDOC-CRM [9].

Another significant ontology worth mentioning is RiC-O (Records in Contexts Ontology) [10], developed by the International Council on Archives (ICA) and implemented at the French National Archives (ANF). RiC-O attempts to integrate as a knowledge graph different widely used XML-based resource representations, including the Encoded Archival Description (EAD). The core of the conceptual model behind RiC-O are Records, Agents, Activities, as well as an Instantiation entity. Besides these core classes, it presents an extensive list of object properties that capture a wide range of interesting relations concerning manuscripts, including letters. The RiC-O representation of the relationships between agents and documents, including properties such as “isAddresseeOf”, “hasSender”, “hasProvenance”, successfully exemplifies the formalization of such relations in an object-centric way. An example of a provenance representation of a cultural object according to RiC-O is present in Fig. 3.

Finally, given the number of German language texts to be integrated into the Arcarium, it is necessary to acknowledge the Gemeinsame Normdatei ontology during the creation of TAO. The GND ontology serves as a structured framework within the Gemeinsame Normdatei (GND), also known as the Integrated Authority File. The GND is an international authority file that is used to organize personal names, subject headings, and corporate bodies from library catalogs,

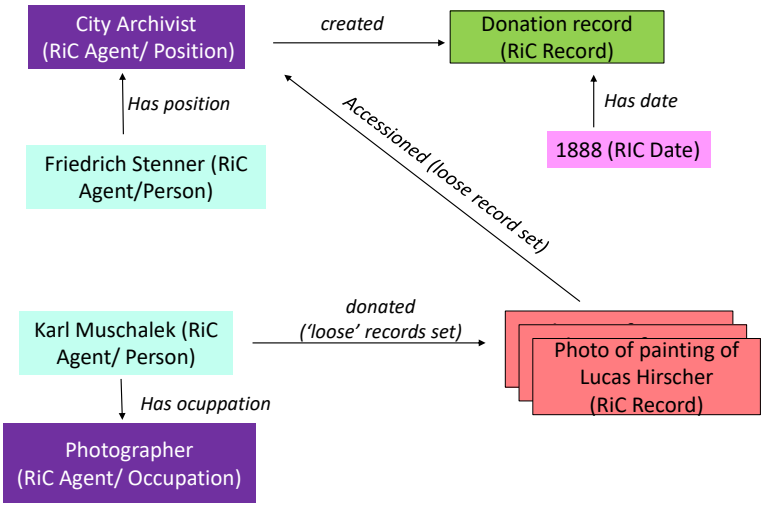

Figure 3: Provenance Context of the Photo of a painting of Lucas Hirscher according to the RiC-O model [10].

archives, and museums. Managed by the German National Library (DNB) in collaboration with regional library networks and other partners, the GND aims to provide a standardized approach to the organization and identification of these entities. The GND ontology is a key component of the GND specification, offering a hierarchical structure of high-level entities and subclasses that are relevant for library classification. In addition, it provides a method to ensure the unambiguous identification of individual elements within the authority file. This ontology is designed not only for traditional library documentation but also for knowledge representation within the semantic web. Fig. 4 shows a triplet representation of the 1977 novel "The Professor of Desire" by Philip Roth.

```
−<rdf:RDF>
  −<rdf:Description rdf:about="https://d-nb.info/gnd/7864443-4">
      <rdf:type rdf:resource="https://d-nb.info/standards/elementset/gnd#Work"/>
    −<wdrs:describedby>
      −<rdf:Description rdf:about="https://d-nb.info/gnd/7864443-4/about">
          <dcterms:license rdf:resource="http://creativecommons.org/publicdomain/zero/1.0/"/>
          <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2018-10-16T16:00:26.000</dcterms:modified>
          <gndo:descriptionLevel rdf:resource="https://d-nb.info/standards/vocab/gnd/description-level#1"/>
        </rdf:Description>
      </wdrs:describedby>
      <gndo:gndIdentifier rdf:datatype="http://www.w3.org/2001/XMLSchema#string">7864443-4</gndo:gndIdentifier>
      <owl:sameAs rdf:resource="http://viaf.org/viaf/245476526"/>
      <gndo:oldAuthorityNumber rdf:datatype="http://www.w3.org/2001/XMLSchema#string">(DE-588c)7864443-4</gndo:oldAuthorityNumber>
      <gndo:variantNameForTheWork rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Professor der Begierde</gndo:variantNameForTheWork>
      <gndo:preferredNameForTheWork rdf:datatype="http://www.w3.org/2001/XMLSchema#string">The professor of desire</gndo:preferredNameForTheWork>
      <gndo:firstAuthor rdf:resource="https://d-nb.info/gnd/118803433"/>
      <gndo:gndSubjectCategory rdf:resource="https://d-nb.info/standards/vocab/gnd/gnd-sc#12.2p"/>
      <gndo:geographicAreaCode rdf:resource="https://d-nb.info/standards/vocab/gnd/geographic-area-code#XD-US"/>
      <gndo:languageCode rdf:resource="http://id.loc.gov/vocabulary/iso639-2/eng"/>
      <gndo:biographicalOrHistoricalInformation xml:lang="de">Roman, 1977 erschienen</gndo:biographicalOrHistoricalInformation>
      <gndo:dateOfPublication rdf:datatype="http://www.w3.org/2001/XMLSchema#string">1977</gndo:dateOfPublication>
    </rdf:Description>
  </rdf:RDF>
```

Figure 4: RDF representation of a novel using the GND schema (adapted from http://d-nb.info/gnd/7864443-4/about/rdf).

3. Approach

The DiHeLib project is focused on describing and digitizing the Autograph Collection of the "Berlin Collection", but the methods and tools that will emerge as a result should be more versatile to be used in subsequent projects and to offer the greatest possible capacity for processing the knowledge so collected. As a result, the TAO ontology as the basis of the Arcarium system under development must meet a wide variety of requirements identified in the DiHeLib project in collaboration with a group of domain experts (12 people, philologists and librarians, members of the DiHeLib project research group) and with library staff responsible for maintaining information systems. The most essential requirements include:

- Collection-specific needs:
  - R1 Compliance with the description format used by Helga Döhn in the catalog of the Autograph Collection [11] – in the DiHeLib project, this catalog will be expanded and verified with actual documents.
  - R2 Alignment with the method of manuscript description adopted in the JL for "Berlin Collection" based on nine main types of documents [12].
  - R3 Ability to integrate with the descriptions of the other half of the collection (stored in Berlin) contained in the Kalliope catalog, which uses the Encoded Archival Description (EAD) standard.
- Demands related to the need to integrate with existing infrastructure:
  - R4 Compatibility with MARC 21 and Dublin Core standards (cf. Sect. 1).
  - R5 The scanned documents will not be stored in the Arcarium system, but on a separate server. TAO should contain information about the location of these files.
- Requirements aimed at creating a universal research ecosystem:
  - R6 Storing identifiers from Wikidata (persons) and Geonames (places), to facilitate the retrieval of contextual information from external sources, incl. other knowledge bases that are linked to these two.
  - R7 Compliance with the two most popular models in the cultural heritage field, i.e., CIDOC CRM and EDM, which will allow for easy future expansion of TAO (using concepts from CIDOC and EDM) and integration with other knowledge bases aligned to these two.

The starting point for developing TAO was a list of properties compiled by librarians based on their own experience, knowledge of the collection, and good practices followed in JL. The list included both a detailed description of individual documents (letters, drawings, etc.) and a general description at the level of the entire unit[2]. The list of properties was used to develop

[2]The Autograph Collection was already organized in Berlin into units containing groups of documents associated with particular individuals. Each of these units was assigned a shelfmark by Berlin librarians consisting of the abbreviation SA (Sammlung Autographa)and the person's name, e.g. "SA, von Herder, Johann Gottfried". The entire collection consists of about 30,000 units.

The Knowledge Matrix (TKM), a set of Excel sheets and scripts serving as an input interface for metadata entry by experts working on manuscripts [4]. Since the property list covered the planned need to enter data into library systems in MARC 21/Dublin Core formats and included all the fields required by domain experts, fulfilling R1, R2 and R4 could be reduced to the need to include all TKM fields in the TAO ontology.

To encompass the totality of the metadata collected by the domain experts, it became necessary to conceive TAO not solely as an ontology outlining objects, but also as encompassing individuals and locations implicated. To make this possible, the input interfaces have been extended to allow additional data to be entered detailing individuals with connections to shared vocabularies such as GND, the CERL Thesaurus, Trecanni, and Wikidata, a roster of aliases for each individual, pertinent birth and death dates, and optionally, a brief description of the person's activities. Information about places was less comprehensive, involving only a link to Geonames and a list of aliases. The input interfaces were further extended to allow the manuscript digitizing team to provide the location of the digitized scan files. All of the described extensions to the interfaces made it possible to collect the data necessary to fulfill R6 and R5.

The main part of TAO's development was the iterative design of successive versions incorporating further elements of TKM or mappings to external ontologies. Key steps taken in selected iterations:

1. Comparison of various ontologies (including EDM and RiC-O) with TKM data format (for more details, see [13]) and identification of parts that can be reused.

2. Preparation of a set of use cases and competency questions (see Fig. 8; due to communication issues with domain experts, most of these were conceived by the authors of this paper).

3. Attempts to design TAO as a subset of EDM (TAO v. 0.1).

4. Design of TAO (v. 0.2) as (mostly) a TKM-driven ontology.

5. Merge of both approaches (TAO v. 0.3 = TAO v. 0.1 + TAO v. 0.2).

6. Populating TAO with sample instances (based on TKMs completed as part of the digitization pilot). In this step, the Owlready Python package was used for convenient manipulation of the data provenient from spreadsheets (cf. [14]).

7. Discussion on TAO with an international group of experts (during a dedicated workshop in the DiHeLib project) and improvement of the ontology based on the conclusions.

The ontology prepared in this way is compatible with the EDM, thus fulfilling R7. By mapping to the EDM, TAO also covers the basic description in MARC 21 and Dublin Core formats (R4), since concepts from Dublin Core are explicitly used in the EDM. The mapping to the Kalliope model (R3) was more difficult because this system is based on Encoded Archival Description, which is not an ontology, but a rich set of keywords that can be used to describe documents. Integration with EAD was achieved in two ways: on the one hand through mappings to RiC-O and EDM (which are based on EAD, among other things), and on the other hand through direct mappings that will allow Arcarium to export data in EAD format [13].

4. Description of the TAO ontology

TAO is composed of 14 main classes, including “Agent”, “Place”, “TimeSpan” and “CollectionEntity”. Agents are divided into 2 subclasses “CollectiveBody” and “Person”, while collection entities include shelfmarks, documents, and their digital representation. The class “Document” accounts for all the cultural heritage objects contained in the Berlin collection. Documents can be divided into 7 subclasses: “Correspondence”, “Effigy”, “HistoricDocument”, “LibraryMaterial”, “Oeuvre”, “PersonalDocument”, “Print”. Fig. 5 presents such a class hierarchy and the relation of particular classes to the equivalent in EDM. Figs. 6 and 7 provide a graphical representation of the two most important classes in TAO and a subset of the classes and subclasses they are linked to.

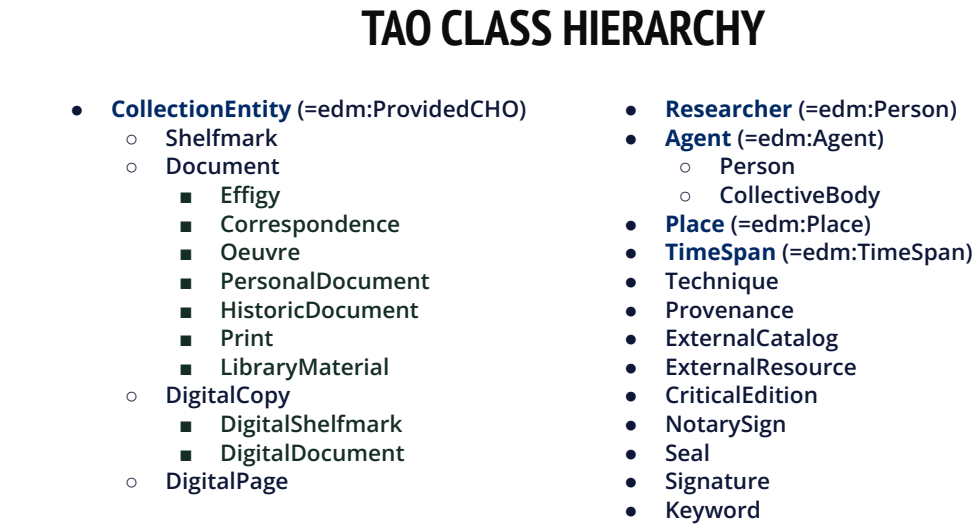

Figure 5: TAO class hierarchy and its relation to EDM.

The set of object properties linking documents and agents is intrinsically related to the subclasses of documents to which an object belongs. A pair of properties such as “isSenderOf”/“hasSender” and “isRecipientOf”/“hasRecipient” are necessary properties to represent a letter, while the pair “isDepictedOn”/“depicts” is rather used for effigies. Other properties apply to all kinds of documents, those properties usually are connected to modeling the history of the document, including “isAuthorOf”/“hasAuthor” and “isProvenanceOf”/“hasProvenance”. Another set of properties is used to represent the very organization of the collection, including those that represent the relation between documents and shelfmarks. In this set belong properties such as “isOrWasIncluded”/“includes”, “next”/“previous”, “hasDigitalCopy”/“isDigitalCopyOf”.

As mentioned above, TAO can be both represented from a document-centric approach and an agent-centric approach. Even though TAO is not developed from an event-centric approach, it is still possible to augment it to make it more easily interoperable with CIDOC CRM. One step towards such a development is the inclusion of an event class.

Finally, it is worth mentioning the “TimeSpan” class, since dates related to both “Agents” and “Documents” in the “Berlin Collection” are not simple. While for some entities dates are precise

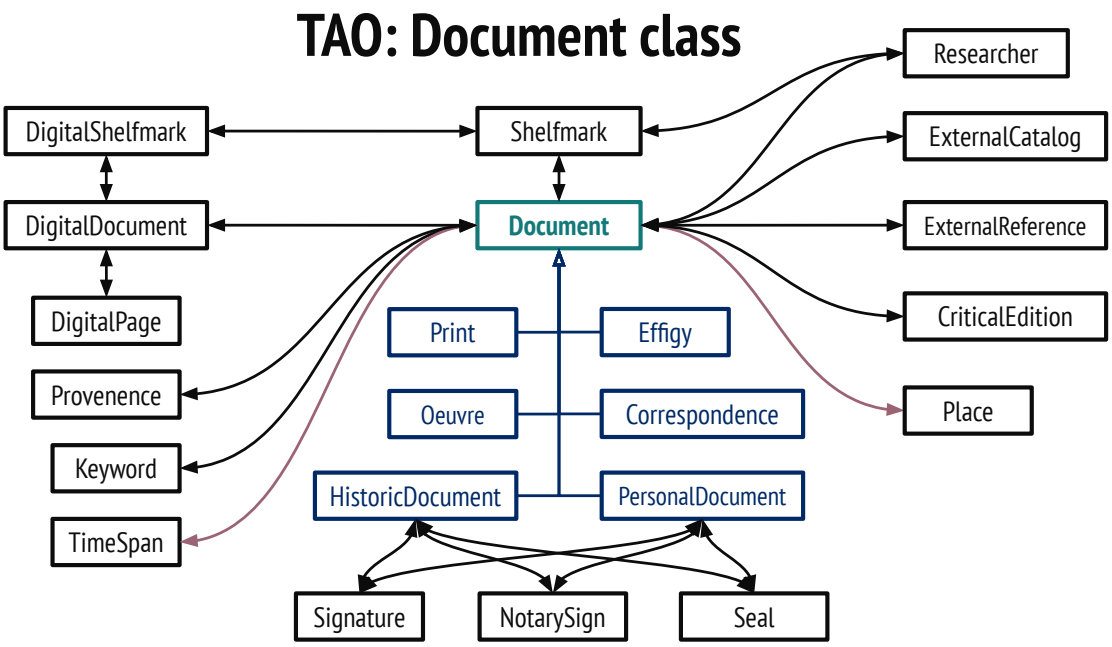

Figure 6: Document-centric view on TAO.

to the day, month, and year, for others they are vaguely known. For example, the "Letter from Cardinal Francesco Barberini to Count Ferdinando di Lodrone" is dated to be sent in October of the year 1722. In this case, the date entity with which the creation of this letter is associated is a timespan with a start date at the beginning of this month, and an end date the October, 31st, 1722. Another interesting peculiarity of dates in this collection is that there might be more than one date accepted by the research community to be associated with a given entity. In this case, the mostly accepted dates are included as begin and end dates of a timespan, and the other possible dates are represented as alternative dates. An example of this case is the birthdate of Ugo Foscolo, which has as its main date the year 1778—i.e., a timespan from 01.01.1778 until 31.12.1778—and as an alternative date the year 1789—i.e., a timespan from 01.01.1789 until 31.12.1789. Both dates are included as one individual, the timespan entity called TS1778(1789).

In addition to the main classes and properties mentioned above, there are other classes such as "Technique", "Seal", "NotarySign", "ExternalCatalog", "CriticalEdition", etc. These classes and their relation to the given document are used to enrich the latter with non-necessary information. Furthermore, information about agents can be enriched with relations such as "hasChild"/"hasParent" and "hasoOrHadSpouse"/"isSpouseOf".

5. TAO in use

Part of the benchmarks in the development of TAO was a set of use cases that could potentially be transformed into competence questions. Revisiting such use cases makes possible a theoretical evaluation of the organization of classes and properties in TAO. One of the use cases for TAO is

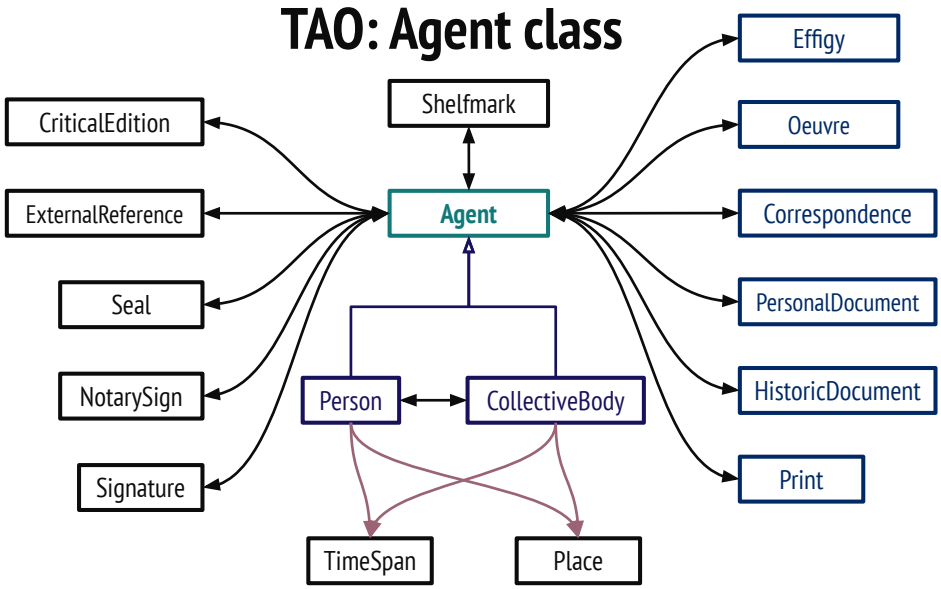

Figure 7: Agent-centric view on TAO.

serving as a system for querying information on individuals-centered networks. TAO should be able to answer questions such as a list of individuals who corresponded with a given person during a specific period, e.g., with Johann Gottfried von Herder between 1760 and 1780. The response list would contain information on correspondences to and from Herder with basic information about these letters, such as their creation date, the place from which they were sent, and their language.

The brief description of TAO in Sect. 4 shows what would be involved in such a query. The person Johann Gottfried von Herder would be represented by a "Person" class. Having such a representation, it is possible to query for the "Letter" entities that have "Herder" as the object of the properties "isSenderOf" or "isRecipientOf". Having these letters, it is possible to get the different authors, except Herder, who are senders or addressees of these letters. Finally, it is possible to filter for letters that "hasDate" a timespan with "begin" equal to or larger than "01.01.1760" and "end" equal to or less than "31.12.1780". Some refining options to this query would be to (1) include or not letters whose timespan is not precise enough, for example, a letter supposedly written in the second half of the 18th century; (2) include or not alternative dates; (3) if the addressee of a letter is a "CorporateBody", then if some kind of member of it should be included on the list.

Another application of TAO involves tracking connections, specifically verifying if there is a chain of correspondence connections between two selected individuals. This chain can be uncovered using a search algorithm that begins with a designated person and their associated set of letters—those either written by them or addressed to them. The algorithm then proceeds to explore the addressees and senders of these letters, retrieving further correspondence linked to these individuals and identifying additional addresses and senders. This iterative process continues until the shortest connection between the initial person and the target person is

revealed—or until it is determined that such a connection does not exist.

To enhance the algorithm's ability to discover connections, it can incorporate pairs of individuals who were present at the same location during the same period. The dates indicating when a "Person" stayed at a particular "Place" and authored a "Document" within a specific timeframe, along with the location where this document was created, represented by the property "happenedAt", are crucial in this context. Once again, the specificity of these dates serves as a crucial refining factor.

In addition to these more complex scenarios, straightforward competency questions also guided the development of TAO. Attempting to represent such queries in SPARQL can give a practical evaluation of the usefulness of the implemented classes and properties. Fig. 8 presents the questions in plain text and their corresponding SPARQL query.

Competency question	SPARQL query
To what shelfmark does a specific document belong?	SELECT **?x** WHERE{ **?x** rdf:type tao:Shelfmark . **?x** tao:includes tao:SA_Alter_Franz_Carl_2.08 }
How many letters (outgoing, incoming) are in a given shelfmark?	SELECT (COUNT(?x) as **?count**) WHERE{ ?x rdf:type tao:Correspondence . ?x tao:belongs tao:SA_Alter_Franz_Carl }
Who wrote to a given person?	SELECT DISTINCT **?senderName** WHERE{ ?senderPerson tao:hasName **?senderName** . ?senderPerson tao:isSenderOf ?letter . ?letter tao:hasRecipient tao:KelAda1 }
In what languages are the documents in a given shelfmark?	SELECT DISTINCT **?language** WHERE{ tao:SA_Alter_Franz_Carl tao:includes ?document . ?document tao:language **?language** }
Does a given person appear in the collection under other names, and if so, what ones?	SELECT DISTINCT **?alias** WHERE{ ?x tao:hasName "Alter, Franz Karl" . ?x tao:alias **?alias**}

Figure 8: Plain text competency questions and the corresponding SPARQL queries prepared for TAO. All of the queries included two prefixes: `rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>` and `tao: <http://www.arcarium.uj.edu.pl/tao/0.3/>`.

Evaluating the current progress in the development of TAO using both the use cases and the competency questions shows that the cultural heritage object metadata representation model within TAO is proving instrumental in addressing various requirements of the Arcarium system's end users. Through an examination of use cases and competency questions, it becomes evident that TAO's development aligns closely with the envisioned functionalities and demands. These evaluations not only validate the effectiveness of TAO in handling complex scenarios, such as querying information networks and tracking correspondence connections, but also highlight its adaptability to straightforward competency questions. By integrating feedback from these assessments, TAO has the potential to continue evolving as a robust framework that meets the diverse needs of cultural heritage management and research endeavors.

6. Summary and future work

This paper discusses the ongoing development of TAO, an ontology designed as the core structure for the Arcarium system. A key consideration for TAO is its ability to interoperate with existing platforms and ontologies. In our review of related work, we referenced four distinct ontologies that illustrate approaches to representing metadata for cultural heritage objects. These include EDM, GND, RiC-O, and CIDOC CRM. Each ontology presents both advantages and limitations in the context of the DiHeLib project. Specifically, EDM is crucial for ensuring interoperability with Europeana, adherence to GND is essential due to the prevalence of German language materials in the Berlin collections, RiC-O offers a detailed object-centric ontology for manuscripts, and CIDOC CRM is notable for its detailed coverage and broad adoption.

Arcarium is intended to be a knowledge base offering storage and searching of rich metadata—which current cataloging and digital library systems did not allow—about manuscript collections and enabling advanced research scenarios in the digital humanities. Thus, the development of TAO must reconcile a variety of requirements including the specifics of the digitized Autograph Collection, the need for universality of the entire solution, and the need for integration with the already existing infrastructure maintained at the Jagiellonian University and the Jagiellonian Library.

After highlighting the main principles behind the development of TAO, we showed the main classes and properties of the ontology. TAO encompasses not only information about documents, but also about individuals. Its structured class hierarchy, exemplified by main classes such as “Agent”, “Place”, and “Documents”, provides a comprehensive framework. TAO’s flexibility allows for both document-centric and agent-centric representations, with provisions for future interoperability enhancements, such as event modeling. Notably, the inclusion of the “TimeSpan” class addresses the nuanced nature of dates associated with entities in the Berlin collection. Additionally, supplementary classes like “Technique” and “Seal” enrich document metadata, while relational properties can augment agent information. Furthermore, the evaluation of TAO’s alignment with practical use cases and competency questions underscores its effective role in the emerging Arcarium system.

It is by adhering to a partially shared vocabulary, especially with EDM and RiC-O, that TAO aims at semantic interoperability with other systems. Examples of shared vocabularies are the use of a derivative class from EDM’s timespan or the linkage via owl:sameAs between the classes tao:Place, edm:Place and rico:Place. Another way to achieve such interoperability is by referencing GeoNames and WikiData. A future addition to the system is the possibility of multiple triple pattern functions allowing the user to query different conceptual models at the same time. In sum, we can claim that TAO currently achieves a limited degree of interoperability.

The development of TAO is currently linked to the DiHeLib project and the Berlin collection. As metadata collection is an ongoing process, TAO improvements depend on the nature of future documents under analysis. Furthermore, a publication of a prototype of Arcarium and TAO is planned, which will allow end-users to submit proposals on how to improve the ontology. Another possibility for the future of TAO is growing beyond the Berlin collection, to encompass other documents included in the Jagiellonian Library or other sub-units of the Jagiellonian University. These possibilities for expansion lay the groundwork for an exciting future for TAO, fostering interoperability not only among institutions but also within the Jagiellonian

University itself.

Acknowledgments

This publication was funded by a flagship project "CHExRISH: Cultural Heritage Exploration and Retrieval with Intelligent Systems at Jagiellonian University" under the Strategic Programme Excellence Initiative at Jagiellonian University. The research for this publication has been supported by a grant from the Priority Research Area DigiWorld under the Strategic Programme Excellence Initiative at Jagiellonian University.

References

[1] E. Bakowska, The Jagiellonian Library, Cracow: its history and recent developments, Library Review 54 (2005) 155–165. doi:10.1108/00242530510588917.

[2] B. Jurkowicz, The collection of the prussian state library. polish, german, or european cultural heritage?, in: K. Ziemer (Ed.), Memory and Politics of Cultural Heritage in Poland and Germany, Cardinal Stefan Wyszyński University in Warsaw, Warsaw, 2015, pp. 117–130.

[3] R. Sosnowski, P. Tylus, European treasure in the jagiellonian library. a flagship project, Polish Libraries 11 (2023) 235–244. doi:10.36155/PLib.11.00008.

[4] K. Kutt, J. Gomułka, L. do Valle Miranda, G. J. Nalepa, Microsoft cloud-based digitization workflow with rich metadata acquisition for cultural heritage objects, Multimedia Tools and Applications (2024). Submitted, in review.

[5] A. Isaac, Europeana Data Model Primer, Technical Report, Europeana, 2013. https://pro.europeana.eu/page/edm-documentation.

[6] J. Lindenthal, H.-L. Meiners, D. Balzer, LIDO Primer, Technical Report, CIDOC LIDO Working Group, 2023. https://lido-schema.org/documents/primer/2023-09-20/lido-primer.html.

[7] D. V. Pitti, Encoded archival description: An introduction and overview, The New Review of Information Networking 5 (1999) 61–69. URL: https://doi.org/10.1080/13614579909516936. doi:10.1080/13614579909516936.

[8] J. McDonough, METS: standardized encoding for digital library objects, International Journal on Digital Libraries 6 (2006) 148–158. URL: https://doi.org/10.1007/s00799-005-0132-1. doi:10.1007/S00799-005-0132-1.

[9] S. du Château, D. Boulanger, E. Mercier-Laurent, Voice knowledge acquisition system for the management of cultural heritage, in: Z. Shi, E. Mercier-Laurent, D. Leake (Eds.), Intelligent Information Processing IV, Springer US, Boston, MA, 2008, pp. 38–49.

[10] D. Pitti, G. McCarthy, B.-F. Popovici, Records in Contexts (RiC). An Archival Description Draft Standard, Technical Report, ICA Experts Group on Archival Description, 2017. https://www.alaarchivos.org/wp-content/uploads/2018/01/1.-Daniel-V.-Pitti.pdf.

[11] H. Döhn, Die Sammlung Autographa der ehemaligen Preussischen Staatsbibliothek zu Berlin: Autographenkatalog auf CD-ROM, Harrassowitz Verlag, Wiesbaden, 2005.

[12] J. Kita-Huber, M. Jaglarz, Re-cataloguing the varnhagen collection. a proposal of a new

description scheme and its application to the selected material, Polish Libraries 10 (2022) 135–161. doi:10.36155/PLib.10.00006.

[13] L. do Valle Miranda, K. Kutt, G. J. Nalepa, Advancing manuscript metadata: Work in progress at the Jagiellonian Library, in: The 28th International Conference on Theory and Practice of Digital Libraries (TPDL 2024), 2024. Submitted, in review.

[14] J.-B. Lamy, Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies, Artificial Intelligence in Medicine 80 (2017) 11–28. doi:10.1016/j.artmed.2017.07.002.