CIDOC-CRM and the First Prototype of a Semantic Portal for the CHExRISH project

Luiz do Valle Miranda[1,*], 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*

Abstract

In this paper we present work in progress towards the integration of cultural heritage resources from different units of the Jagiellonian University as linked data and the prototyping of their presentation in a semantic portal. CIDOC-CRM has been chosen as the data model behind such an interoperability given its wide use and its flexibility. Challenges arose when converting bibliographical authorship relations into CRM's event-centered structure and when migrating instances' hierarchical classification into CRM's "E55 Type" and "P127 has broader term" conversion standard. Despite following CRM's data modeling best practices, these challenges reappeared while publishing the data in an Omeka S-powered website, thus showing some lack of compatibility between these two frameworks.

Keywords

CIDOC-CRM, Cultural Heritage, Semantic Portal, Linked Data

1. Introduction

A major motivational factor for the use of Linked Data (LD) in Cultural Heritage (CH) projects is the aspiration towards interoperability among differently collected data within an institution or between different institutions. Aligned with such interoperability, modeling, and publishing CH data as LD aims at the aggregation of resource with data published within the Semantic Web (SW) framework that are accessible via their unique URIs. Furthermore, using Knowledge Graphs (KG) and ontologies as the base technology for such a CH-LD endeavor provides a rich description of spatial, temporal, and personal relations of the included collections that allows for further processing in the form of algorithms that enrich the collection [1].

Taking into account such benefits of LD in CH, the ongoing CHExRISH project at the Jagiellonian University (JU) aims to create a prototype for the Jagiellonian University Heritage Metadata Portal (JUHMP). JUHMP is envisioned as a semantic portal that first and foremost provides browsing and searching capabilities over JU heritage data currently being stored and analyzed separately in different units, including the Archive (AUJ), the Museum (MUJ), and the Library (BJ). JUHMP is also conceived to incorporate tools for assisting researchers in solving Digital Humanities (DH) problems, such as Network Analysis [2] and VR visualization [3], as well as the use of Artificial Intelligence (AI) tools for recommendation and knowledge discovery [4], thus approximating it to what [5] calls a third generation semantic portal. For the first prototype of JUHMP (JUHMP_v1.1), Omeka S[1] was chosen given its purported ease of installation at its support for linked data (see [6, 7]).

A significant challenge in this project is the integration of diverse data from every unit, additionally since each unit has collections modeled (e.g., Dublin Core, MARC-21), stored (e.g., SQL, CSV, XML, RDF) and accessed (e.g., OAI-PMH, custom REST API) in different formats. This challenge is addressed with the selection of a flexible and extendable ontology as the data model for the unified base. After considering

[1]https://omeka.org/s/

different ontologies such as the Europeana Data Model[2] (EDM) [8], the Records in Contexts Ontology[3] (RiC-O) [9] and the Gemeinsame Normdatei (GND) Ontology[4], the decision was the CIDOC-CRM[5] [10]. CIDOC-CRM is a widely used ontology where cultural heritage data from different institutions, including libraries, museums, and archives, can be connected by means of the events relating records with, among others, people, places, types, concepts, and institutions. CIDOC-CRM is an international standard (ISO 21127:2014) designed to serve as a shared language among CH institutions. Given its status and its adoption, many tools for semantic portal's enrichment are being developed based on this data model (see [11, 12]).

The purpose of this paper is to present and discuss the current workflow and the challenges encountered in the process of data integration and the development of the semantic portal based on CIDOC-CRM v7.1.2[6]. While the CHExRISH project aims to integrate a variety of collections related to JU heritage, JUHMP_v1.1 currently only includes selected data from the AUJ's Corpus Academicum Cracoviense[7] (CAC) and BJ's Jagiellonian Digital Library[8] (JDL).

2. Integrating data of the CAC and JDL

CAC is an electronic database with around 67,000 records on students and graduates of the University of Kraków during the period 1364–1780[9]. The Web application for accessing CAC data displays records on academic, occupational, personal, and editorial events in a given person's life. Fig. 1 partially reproduces the CAC record for Nicolaus Copernicus. Among the records, we can see Copernicus's birth and death date, some institutions where he has worked and his position, two of his publications, and the place where he studied and where he finished his doctoral studies.

CAC interface is integrated with a PostgreSQL database. For the initial prototyping of JUHMP_v1.1, a dump of the database with the database diagram was used. A first challenge arose when academic and occupational events were modeled as CRM's "E7 Activity" class. As shown in Fig. 1, academic events have several properties, including event type, degree type, and the scientific discipline. The CRM standard solution for aggregating details in an activity is by creating an instance of class "E55 Type" connected to the former with property "P2 has type". Even though this solution was able to capture the academic events' properties, it was not able to represent each of the properties as a related group. This grouping was achieved by creating more instances of the class "E55 Type", respectively "Educational Activity", "Academic Degree", and "Field of Study" linked to the former types with CRM's "P2 has type" property. A similar structure was taken for occupational activities. The problem of attaching properties to instances of a class for further specification was found in the case of names and surnames. They were modeled as instances of "E41 Appellation", and also linked to instances of "E55 Type", respectively "First Name" and "Surname" to differentiate between them. Fig. 2 presents a visualization of the current solution of using CIDOC-CRM to model Copernicus' personal record and bibliography. At the current moment publication activities as described in CAC were not included in JUHMP_v1.1 given the difficulty of resolving naming differences to the works described in JDL.

JDL, a BJ's digital library, is a platform for the preservation and dissemination of (not only) antique collections of the Jagiellonian Library as digitized resources. Among the digitized resources are the collection of manuscripts, the collection of old prints, the collection of prints from the 19th and 20th centuries among others. Data from JDL follows the Dublin Core[10] (DC) data model and can be harvested

---

Mikołaj Kopernik (Copernicus) z Torunia, syn Mikołaja

	+
First name:	Mikołaj
Surname:	Kopernik
Father's name:	Mikołaj
Place of origin:	Toruń
Place and date of birth:	Feb 19, 1473
Place and date of death:	May 24, 1543
Known events from period:	Feb 19, 1473 — the end of 1543 year

Education and academic degrees

	⇕	Event type ⇕	Education stage/academic degree ⇕	Scientific discipline ▲	Institution ⇕	Date ⇕
⊞	1	koniec studiów/nauki	student	medycyna	Uniwersytet Padewski	the beginning of summer 1503 — the end of summer 1503
⊞	1	początek studiów/nauki	student	medycyna	Uniwersytet Padewski	the beginning of August 1501 — the end of December 1501
⊞	1	immatrykulacja	student	prawo	Uniwersytet Boloński	after Jan 6, 1497
⊞	1	immatrykulacja	student	sztuki wyzwolone/filozofia	Uniwersytet Krakowski (Akademia Krakowska)	the beginning of winter semester 1491-1492 — the end of winter semester 1491-1492
⊞	1	koniec studiów/nauki	student	sztuki wyzwolone/filozofia	Uniwersytet Krakowski (Akademia Krakowska)	the beginning of summer 1495 — the end of autumn 1495

Functions/offices/roles

	⇕	Event type ⇕	Function/office/role ▲	Place/Institution ⇕	Date ⇕
⊞	1	wybór/mianowanie	kanonik	katedra we Fromborku	Aug 26, 1495 — the end of 1495 year
⊞	1	objęcie funkcji	kanonik	katedra we Fromborku	Oct 20, 1497
⊞	1	rezygnacja	scholastyk (koadiutor)	katedra we Wrocławiu	the beginning of 1538 year — the end of 1538 year
⊞	1	początek funkcji	scholastyk (koadiutor)	katedra we Wrocławiu	Jan 10, 1503

Works and books

	Event type	Tilte	Shelf mark	Date	Source of information ▲	
🔍	autor	Nicolai Copernici Torinensis De Revolutionibus orbium coelestium libri VI (1543)		the beginning of 1543 year — the end of 1543 year	PSB, t. 14	
🔍	autor	O szacunku monety (De aestimatione monetae)		the beginning of March 1522 — the end of March 1522	AM, t. 8	💬

Figure 1: Excerpt of Nicolaus Copernicus' record from CAC.

using the OAI-PMH protocol. Listing 1 contains the metadata for Copernicus' "De Revolvtionibvs Orbium Coelestium" as shared by JDL.

Since the digital objects that are associated with CAC entities are not distinguished in any way in the JDL, it was necessary to harvest all data accessible via OAI-PMH to be processed into the JUHMP Ontology. Around 700,000 bibliographic records were retrieved. Only bibliographic records pertaining to CAC entities were taken into consideration. However, the mapping of authors from CAC to JDL is not obvious, since in Poland there is no authority file for the organization of personal names shared by all GLAM institutions (there are standards for Polish libraries [14], followed by BJ, but they are not followed by archives and museums, incl. AUJ and MUJ). For JUHMP_v1.1 a mapping file for 10 records was manually created. Additionally, a simple conversion script was written in Python that was able to map with certainty 99 entities from CAC to JDL. These records were also included in JUHMP_v1.1. In possession of such a conversion file for personal names from CAC to JDL, it was possible to filter the bibliographic records to those referring to CAC entities in the fields "dc:creator", "dc:contributor" and "dc:subject".

Modeling the relationship between a bibliographic record and the people related to it is another challenge. In CRM, a bibliographic record can be modeled both with the "E33 Linguistic Object" and the "E22 Human-Made Object" to represent respectively the linguistic expression and the physical inscription of the expression. There is, however, no property in CRM that directly connects an author or a contributor to instances of these classes. This link is achieved by mediation of events that represent
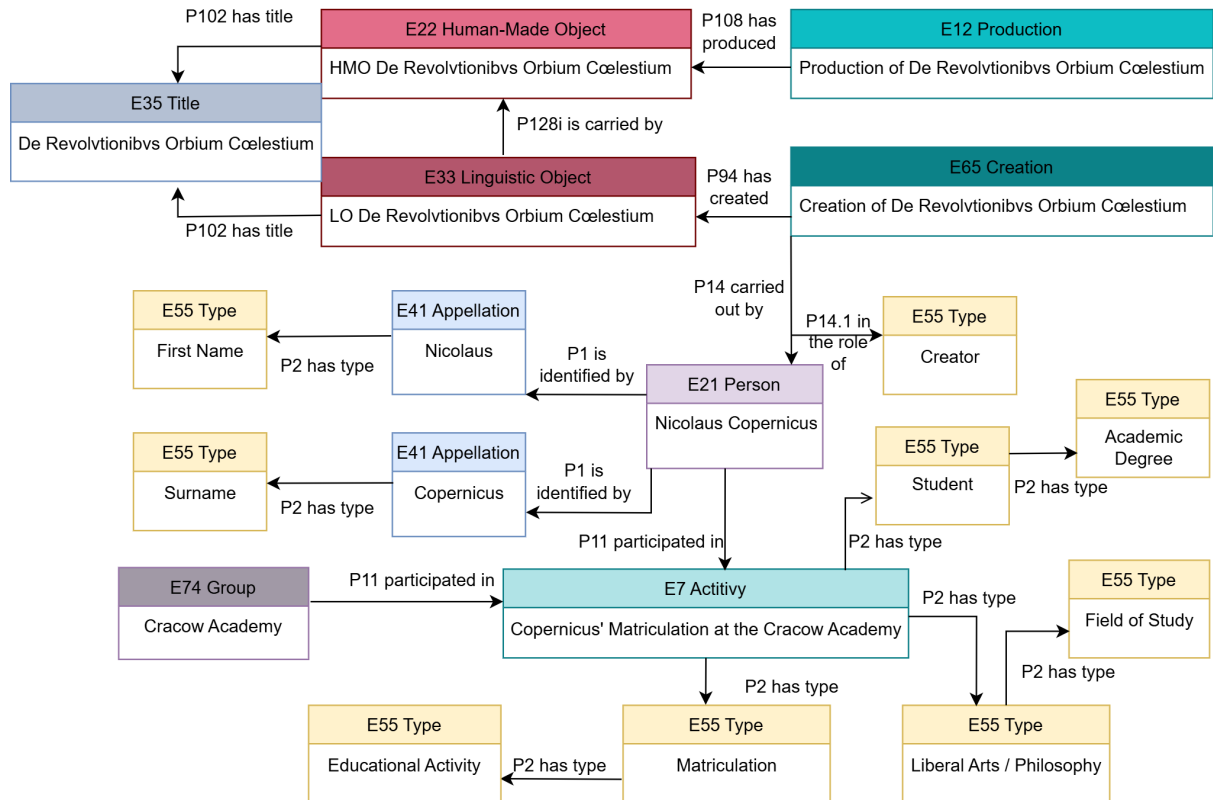
Figure 2: Diagram of the implementation of CIDOC-CRM for an excerpt of Copernicus' personal record and bibliography

```
1   @prefix dc <http://purl.org/dc/elements/1.1/> .
2   @prefix jbc <http://www.jbc.bj.uj.edu.pl/Content/> .
3
4   jbc:219947   dc:coverage "16 w."@pl
5                dc:creator "Kopernik, Mikolaj (1473-1543)"@pl ;
6                dc:date "1543"@pl ;
7                dc:language "lat"@pl ;
8                dc:publisher "Andreas Osiander"@pl ;
9                dc:rights "Domena publiczna (public domain)"@pl ;
10               dc:source "Biblioteka Jagiellonska, BJ St. Dr. Cim. 8288"@pl ;
11               dc:title "De Revolvtionibvs Orbium Coelestium, Libri VI [...]"@lat ;
12               dc:contributor "Schoenberg, Nikolaus von (1472-1537)"@pl ;
13               dc:type "starodruk"@pl .
14
```

Listing 1: DC description of Copernicus' "De Revolvtionibvs Orbium Coelestium, Libri VI" from JDL.

the generation of these objects linked to an author via the property "P14 carried out by". For linguistic objects, "E65 Creation" is used, while for human-made objects, "E12 Production". As discussed in [15], CIDOC-CRM supports the reification of the property "P14 carried out by" and the use of the property of properties "P14.1 in the role" to specify the role by connecting the latter to an instance of CRM's "E55 Type". Fig. 2 presents an example of such modelling practices. Finally, it is worth mentioning that the conversion between both JDL and CAC data from their respective original data formats to the population of a CIDOC-CRM-based ontology was achieved via custom Python scripts using the rdflib[11] library.

---

[11] https://rdflib.readthedocs.io/en/stable/

Mikołaj Kopernik (Copernicus) z Torunia, syn Mikołaja

Title	Mikołaj Kopernik (Copernicus) z Torunia, syn Mikołaja
is identified by	Mikołaj
	Kopernik
participated in	Copernicus' Matriculation at the Cracow Academy
was born	Birth of Mikołaj Kopernik
died in	Death of Mikołaj Kopernik

Linked resources

COLLAPSE ALL

Filter by resource type and property:

Items: All ⌄

Items with "carried out by: Mikołaj Kopernik (Copernicus) z Torunia, syn Mikołaja" ⌃

Creation of De Revolvtionibvs Orbium Cœlestium

Figure 3: Excerpt of Copernicus' record from Omeka S.

3. Prototyping with Omeka S

While the target JUHMP is planned as a complex portal including several modules to assist digital humanities research and intelligent search and exploration, the first prototype, JUHMP_v1.1, has significantly simpler requirements. JUHMP_v1.1 aims to be a web application to showcase the integration of two different data models originating from different JU units over a manually selected set of records. Furthermore, JUHMP_v1.1 is supposed to be based on existing semantic cultural heritage platforms, given the limited resources for developing such a prototype. In JUHMP_v1.1 the user's interaction with these records is supposed to happen in two ways. First, users should be able to browse over a list of included people's records, being able to easily access linked resources, such as bibliographic records, or connected themes. Second, users should be able to query the KG via a SPARQL endpoint.

Based on the comparison table presented by [6] and the analysis in [16], Omeka S appeared to be the most adequate platform for JUHMP_v1.1[12]. Omeka S is presented as an easy-to-use platform compatible with the principles of linked open data, with extensive documentation, a growing community of users, and allowing the addition of modules for the extension of its functionalities. The presentation of records in Omeka S is based on the creation of digital collections with items added either manually, imported via CSV or via an API. Furthermore, a SPARQL module for Omeka S[13] is already widely available, allowing easy integration in the web application of an interface for writing SPARQL queries and executing them against an Apache Jena Fuseki[14] server.

As CIDOC-CRM is not one of Omeka S pre-built ontologies, its import is one of the preliminary steps before the addition of any records. CRM can be easily imported, as Omeka S provides a page with forms to add new ontologies, which only requires an RDF file containing the ontology. This process ensures the availability of the CRM vocabulary in the Omeka S platform enabling its use for both importing data and presenting resources in a semantically enriched manner. However, the addition of records to Omeka S as linked resources is not so straightforward.

While the structure of the payload of a successful request to the "\api\items" endpoint is based on the JSON-LD format[15], the mere conversion of an RDF file to its equivalent JSON-LD is not enough to add a

---

[12]One alternative to Omeka S that was considered was ResearchSpace (https://researchspace.org/). Although ResearchSpace offers native compatibility with CIDOC-CRM, configuring a user-friendly interface requires significantly more development effort compared to Omeka S. Consequently, Omeka S was selected for this initial prototype.

[13]https://omeka.org/s/modules/Sparql/

[14]https://jena.apache.org/documentation/fuseki2/

[15]https://omeka.org/s/docs/developer/api/rest_api_reference/

linked resource to Omeka S. The "\api\items" endpoint's payload requires fields such as "property_id" and "type" for each item, that need to be added with additional scripting. Furthermore, to draw on the LD browsing capabilities of Omeka S, the resources cannot be linked via their URIs, but rather using Omeka S' internal "value_resource_id". Thus, another additional step needed to programmatically add resources to Omeka S is the creation of empty items for each resource, getting their "value_resource_id", substituting the URI references to the resource id references, and finally patching the empty items with their respective linked data. Although Omeka S does not contain an integrated solution for a seamless integration of linked items from an RDF file, with the help of additional scripting it is possible to import a set of linked data points to the Omeka S repository preserving a structure that closely mirrors the RDF description. Fig. 3 presents the excerpt from Copernicus' personal record from CAC presented in Omeka S.

The situation of presenting CIDOC-CRM data in Omeka S poses another challenge. After adding an item to be displayed in a collection, all the triples that have such a given item as a subject will be displayed in the page with the name of the property according to the respective ontology. One way to customize the presentation of an item is by associating a resource template with it. Among other things, resource templates can create an alternate label to some properties or make some properties invisible. However, two shortcomings were identified with item presentation. First, Omeka S does not allow for displaying fields with values nested in another item. A widespread practice in CIDOC-CRM (and CH in a whole) is modeling dates as time-spans. In CRM, for example, the class "E52 Time-Span" is used, with its instances being linked to date data via the properties "P82a begin of the begin" and "P82b end of the end". Taking the Copernicus' personal record as an example, it is not possible to display a Copernicus' birthdate in his page, since it is nested in an instance of class "E67 Birth" and further in an instance of "E52 Time-Span", the same goes for death date, or publication date of a bibliographic record. The issue is present with not only dates, but also with listing books an author wrote or institutions where a person studied or worked.

The second shortcoming relates to the challenge mentioned in Sect. 2 on further classification of instances of CIDOC-CRM classes in hierarchies that are not in the ontology using the "E55 type" class. As visible on Fig. 3, both name and surname are indistinguishably presented under the header "is identified by". The same is the case for academic events and occupational events under "participated in". This limitation is due to the fact the Omeka S does not allow customization or filtering of fields according to some nested value.

Since CIDOC-CRM is a high-level ontology and intrinsically flexible, both nested data types and vocabulary-based classification are essential aspects of modeling according to this standard. Given the lack of these customization possibilities, the LD-based browsing and presenting capabilities of Omeka S are partially incompatible with data represented in CIDOC-CRM[16]. One way to create a customized presentation in Omeka S is expanding the CIDOC-CRM ontology with properties explicitly linking the data points to be presented, for example, creating a property called "death date" that is a short-cut between an instance of "E21 Person" and a date value. The drawback of this approach is the necessity for further ontology development work and maintenance, and the divergence between the data model present in Omeka S repository and in the SPARQL endpoint.

Despite the challenges encountered during the implementation of Omeka S, it remains, in our case, the most viable option for an early-stage prototype of a semantic portal, especially given the constraints in available time and resources for portal design and development.

---

[16]The NestedDataType module (https://github.com/sinanatra/NestedDataType) addresses a similar challenge and presents a solution to some of the issues here presented. However, it is not yet suitable for production use without additional configuration and bug fixes. Further experimentation is required to fully assess the value of adopting such a solution for this particular context.

4. Conclusion and further works

The current experience of implementing a prototype for the aggregation of data based on different controlled vocabularies and data standards has shown many challenges in this process. While some current issues faced while prototyping JUHMP in a first version were solved, such as modeling CAC and JDL data according to the CIDOC-CRM ontology or importing CIDOC-CRM based data into Omeka S, the issue of presenting CIDOC-CRM based data in Omeka S continues being open. Furthermore, the most pressing issues for data integration in the CHExRISH project were avoided by manual mapping or by exclusion from the first prototype. Among these issues, there is the mapping of personal data and bibliographical records between units, and the mapping of these data with external platforms, such as GeoNames[17] and Wikidata[18].

The next step in the CHExRISH project is integrating metadata of MUJ's objects and the complete bibliographic dataset from BJ, exported in MARC21[19]. Despite the complexity of this task, the initial prototype demonstrated that CIDOC-CRM is flexible enough to accommodate CH data from diverse formats and models. The challenge now is to select and develop systems that not only preserve this flexibility but actively leverage it to enhance interoperability and usability.

Acknowledgments

References

[1] E. Hyvönen, H. Rantala, Knowledge-based relation discovery in cultural heritage knowledge graphs, in: Proceedings of the Digital Humanities in the Nordic Countries 4th Conference (DHN 2019), 2019.

[2] C. N. Warren, D. Shore, J. Otis, L. Wang, M. Finegold, C. R. Shalizi, Six degrees of francis bacon: A statistical method for reconstructing large historical social networks, Digit. Humanit. Q. 10 (2016). URL: https://api.semanticscholar.org/CorpusID:40536368.

[3] L. Lu, M. Li, Development of a virtual interactive system for Dahua Lou loom based on knowledge ontology-driven technology, Heritage Science 11 (2023). doi:10.1186/s40494-023-01027-x, publisher: Springer Science and Business Media Deutschland GmbH Type: Article.

[4] D. Melo, I. P. Rodrigues, D. Varagnolo, A strategy for archives metadata representation on CIDOC-CRM and knowledge discovery, Semantic Web 14 (2023) 553–584. doi:10.3233/SW-222798, publisher: IOS Press.

[5] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187 – 193. doi:10.3233/SW-190386.

[6] R. A. Martinez, Omeka s como alternativa para el desarrollo de colecciones digitales y proyectos de humanidades digitales, BiD (2022). doi:10.1344/BID2022.48.05.

[7] A. Schimmenti, P. Bonora, Lod that picture: Leveraging linked open data to enhance navigation of the photographic archive of the teatro comunale of bologna, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 13541 LNCS (2022) 518 – 522. doi:10.1007/978-3-031-16802-4_56.

---

[17] https://www.geonames.org/
[18] https://www.wikidata.org/
[19] https://www.loc.gov/marc/bibliographic/

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

[9] 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.

[10] D. Oldman, The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER, Technical Report, CRM Labs, 2014. http://old.cidoc-crm.org/docs/CRMPrimer_v1.1.pdf.

[11] M. Mountantonakis, Y. Tzitzikas, Generating sparql queries over cidoc-crm using a two-stage ontology path patterns method in llm prompts, J. Comput. Cult. Herit. 18 (2025). URL: https://doi.org/10.1145/3708326. doi:10.1145/3708326.

[12] Y. Tzitzikas, M. Mountantonakis, P. Fafalios, Y. Marketakis, Cidoc-crm and machine learning: A survey and future research, Heritage (2022). URL: https://api.semanticscholar.org/CorpusID:250376251.

[13] M. Baster, Corpus academicum cracoviense: database of students and professors of the university of krakow (1364-1780), in: Universitätsstudium und Gesellschaft in Mitteleuropa vom 15. bis zum 18. Jahrhundert, volume 5 of *Historia et Monumenta Universitatis Jagellonicae*, Towarzystwo Naukowe Societas Vistulana, Kraków, 2017, pp. 265–276.

[14] J. Cieloch-Niewiadomska, Introducing the national library of poland descriptors to the polish national bibliography, Cataloging & Classification Quarterly 57 (2019) 37–58. doi:10.1080/01639374.2019.1573774.

[15] V. Alexiev, Types and annotations for cidoc crm properties, Digital Presentation and Preservation of Cultural and Scientific Heritage 2 (2012) 38–51. URL: https://dipp.math.bas.bg/dipp/article/view/dipp.2012.2.23. doi:10.55630/dipp.2012.2.23.

[16] I. Nishanbaev, E. Champion, D. A. McMeekin, A survey of geospatial semantic web for cultural heritage, Heritage 2 (2019) 1471–1498. URL: https://www.mdpi.com/2571-9408/2/2/93. doi:10.3390/heritage2020093.