A Knowledge Base of Medieval and Renaissance Geographic Latin Works

Valentina Bartalesi [1], Nicolò Pratelli [1,2] and Emanuele Lenzi [1]

[1] *ISTI-CNR,Via G. Moruzzi 1, 56124 Pisa City, Italy*
[2] *Dipartimento di Ingegneria dell'Informazione, Via G.Caruso 16, 56122 Pisa, Italy*

Abstract

The geography of the world created during the Middle Ages and Renaissance (VI-XV centuries) was crucial to the development of Western thought in the European history. Until now, to the best of our knowledge, Medieval and Renaissance geographic Latin literature has not been studied using digital methods. The Italian National research project IMAGO - Index Medii Aevi Geographiae Operum - (2020-2023) aims at providing a systematic overview of this literature using Semantic Web technologies and the Linked Open Data paradigm. As the first step to develop tools to support scholars in creating, evolving and consulting a knowledge base (KB) of the geographic works, we created an OWL 2 DL ontology. To maximize its interoperability, we developed the ontology as an extension of two reference vocabularies: the CIDOC CRM and FRBRoo (including its in-progress reformulation LRMoo). In this paper, we briefly present the project, the ontology, and the automatic and semi-automatic tools we developed to populate it. The final aim of the project is the creation of a Web application allowing scholars to freely access and visualise the data collected in the IMAGO knowledge base.

Keywords

Semantic Web, Ontology, Digital Humanities, Medieval and Renaissance Geography, Knowledge Base, Medieval Manuscript.

1. Introduction

Medieval and Renaissance culture developed an image of the world that was at the basis of Western thought in European history. During the Middle Ages, geographical descriptions were mostly functional to collect human knowledge into encyclopedic works or to provide universal chronicles. Specific descriptions of lands, cities, places, monuments and buildings were also used as a guide by the pilgrims who travelled to the Holy Land, Rome or Santiago de Compostela. The genre of geographical description had a turning point in the Renaissance, due to the exploration travels and discoveries that produced several descriptions and representations of the New World.

Currently, the information about this kind of literature is dispersed on paper books, and this makes a systematic overview of the geographic literature impossible, preventing a well-ordered perception of how it was gradually set up in time. Indeed, this literature has not been studied yet using digital methods. The PRIN 2017 IMAGO - Index Medii Aevi Geographiae Operum – (2020-2023) aims at developing tools to support scholars in creating, evolving and consulting a knowledge base (KB) of the geographical Latin works written in the Middle Ages and Renaissance. In particular, IMAGO aims at (i) creating a knowledge base using the languages of the Semantic Web [6], (ii) producing data according to the Findable, Accessible, Interoperable and Reusable (FAIR) paradigm [11,12], (iii) publishing the data as Linked Open Data (LOD) [2]. The final aim of the project is the creation of a Web application allowing scholars to freely access and visualize the data collected in the IMAGO KB. The idea is to improve the studies of Medieval and Renaissance geography by providing scholars with a better insight into this field from different research perspectives.

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy

The main novelty introduced by our research is the use of the Semantic Web technologies to formally represent the scientific domain of the geographical Latin works written during the Middle Ages and the Renaissance. Despite in other research projects semantic technologies have been used to represent ancient manuscripts no scientific research that applies a Semantic Web approach has been conducted in this specific research field. In this article, we briefly report the work done until now to create a knowledge base of Medieval and Renaissance geographic works.

2. The IMAGO Ontology

As reported in the Introduction, the IMAGO project aims at providing a systematic overview of the Latin Medieval and Renaissance geographic literature using the Semantic Web technologies to make available this knowledge as Linked Open Data (LOD) and to develop automatic search and visualization services on the collected data. To do this, we created an OWL 2 DL[2] ontology to formally represent the collected data. This ontology is the result of a strict collaboration between ISTI-CNR and the experts in Latin and Italian Literature and Linguistics of the University of Pisa and the University of Salento who are involved in the project. As the first step, we defined a conceptualization of the domain of interest. Then, we provided a logic formalization of this conceptualization. In particular, following the re-use logic and in order to maximize the interoperability, we developed our ontology as an extension of two reference ontologies, i.e. the CIDOC CRM [4] vocabulary and its extension FRBRoo [5], including its in-progress reformulation, LRMoo [7].

In the conceptualization we identified some main categories to formally represent. The first ones are the author and title of a work. For each work, we identified its literary genre/s and the toponyms reported into the work. Furthermore, for each work, knowledge about the related manuscripts and printed editions is reported. In particular, for each manuscript we reported: the name of the author and the title of the work as it appears in the manuscript; the library in which the manuscript is collected; the location of the library; the signature and the folios of the manuscript; the incipit and explicit of the dedication/proem, if they exist; the incipit and explicit of the text; the date of the creation of the manuscript; the secondary sources.

On the other hand, for each printed edition the following knowledge is reported: the author, the title, and curator's name of the edition; the place and the date of publication; the publisher; the format of the edition; the number of pages; the information about the images reported in the edition; some general notes that the scholars intend to add as comment to the edition; the name of the author of the introduction, the text of the introduction, the text of the dedications; information about whether the edition is a first edition or a reprint; primary and secondary sources of the edition; the ecdotic typology.

As reported in detail in [1], in order to create our ontology, we mapped the pieces of knowledge of the conceptualisation with classes and properties of the CIDOC CRM and FRBRoo, including its in-progress reformulation, LRMoo. Figure 1 shows the classes, object and data properties we used in our ontology. As a notational convention, the CIDOC CRM uses the letters “E” and “P” to indicate classes and properties respectively, whereas FRBRoo (and its revisions LRMoo) uses the letters “F” and “R” to indicate classes and properties, respectively. We performed a first evaluation of the ontological model. We used the automatic OntoQA system [9]. OntoQA is a feature-based approach for evaluating ontologies and it does not require data training. OntoQA evaluates the ontologies using a predefined set of metrics. The results of this evaluation are reported in deep in [1]. In summary, analyzing the evaluation results, we can infer that the IMAGO ontology is significantly richer than a simple taxonomy (it has a value of relationship richness equal to 68.75%). Furthermore, the inheritance richness of the ontology is equal to 1.66 and this value indicates the vertical nature of the vocabulary, defining it as a domain-specific ontology.

[2] https://www.w3.org/TR/owl2-overview/

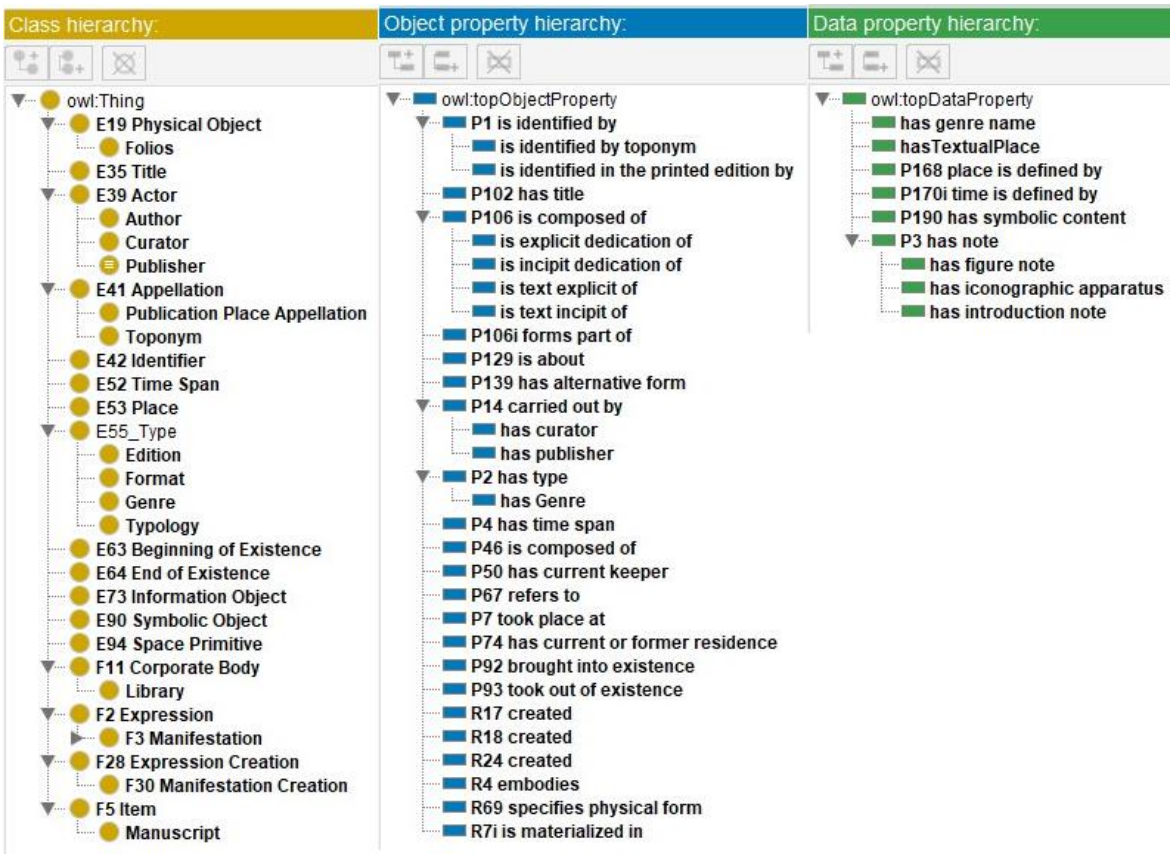

Figure 1: Classes, object and data properties in the IMAGO ontology, as visualized in Protégé.

3. Automatic Population of the Ontology using Different Data Sources

The resources stored in the IMAGO KB have a unique identifier that is an Internationalized Resource Identifier (IRI). These IRIs are automatically retrieved from different sources freely available on the Web. Linking different sources needs extensive use of ontologies in order to discover different knowledge across multiple repositories. In particular, we link the following datasets:

- Wikidata knowledge base[3]
- MIRABILE digital archive[4]
- Nuovo Soggettario thesaurus[5]
- Mapping Manuscript Migration knowledge base[6]
- Plaiades gazeteer[7]

When IRIs are not available in the previous datasets, we automatically created and assigned custom IRIs to the resources. We were able to retrieve: 96% of author IRIs and 98% of work IRIs from Wikidata [10] and Mirabile; 95% of place IRIs and 41% of library IRIs from Wikidata; 95% of genre IRIs from Soggettario Nazionale; 50% of ancient place IRIs from Pleiades [8]. Finally, we were able to map 20% of the works collected in the IMAGO project with those collected in the MMM KB [3]. The linking among different datasets allows to considerably enrich the knowledge collected in the IMAGO KB. At the current stage of the project, we were able to automatically populate our KB with 250 works, 206 authors, 614 libraries, and 310 places. The KB also includes 7 different literary genres, 4 types of editions, 6 ecdotic typologies.

[3] https://www.wikidata.org
[4] https://www.mirabileweb.it
[5] https://thes.bncf.firenze.sbn.it/
[6] https://mappingmanuscriptmigrations.org/en/
[7] https://pleiades.stoa.org/

4. Populating the Ontology with a Semi-Automatic Tool

To support the scholars in populating the ontology, we developed a semi-automatic tool that allows reducing the time for inserting the knowledge in the ontology, facilitates the retrieval of this knowledge, and prevents the possibility to make mistakes while inserting the data manually. The tool is written in Python[8] with the framework Django[9] and is connected to a PostgreSQL[10] with jsonb datatype. Using Django has some advantages, e.g. it is a web framework focused on reducing web application development time; it can easily be customizable, scalable, and extendable by making changes to its components.

Along with the IRIs automatically retrieved, we have embedded in the tool other pieces of knowledge (and the corresponding IRIs) about some resources that have to be added in our KB, i.e. authors, works, places and libraries.

Regarding the information about the authors, though the tool it is possible to insert: the author's name; the author's name aliases; some dates relating to the life and the work of the author. In particular, the scholar can define: (i) the *floruit* date that identifies the highest activity period of an author; (ii) the *episcopus* date that refers to the range of time in which an author has covered the role of bishop; (iii) the dates of birth and death of an author. The *floruit* and *episcopus* dates are reported when the dates of birth and death of an author are not available. All the dates have been modelled with an interval, with a start date and an end date. Furthermore, we have inserted four checkboxes to specify other information about dates, that is: (i) uncertainty, to indicate that the reported date is not uncertain, (ii) before, for events whose dates are uncertain but the scholar knows that they happened before a certain date, (iii) after, for events whose dates are uncertain but the scholar knows that they happened after a certain date, (iv) Roman numerals, to allow the scholars to visualize the dates in roman numerals. Figure 2 shows the interface of the tool that allows inserting this knowledge.

Regarding the works, the tool supports the scholars suggesting a precompiled list of titles and title aliases. For what concerns the libraries, the scholars can choose a library from a predefined list or can insert a new library if it is not already present in the list. Then, the tool automatically adds information about the place in which the library is located.

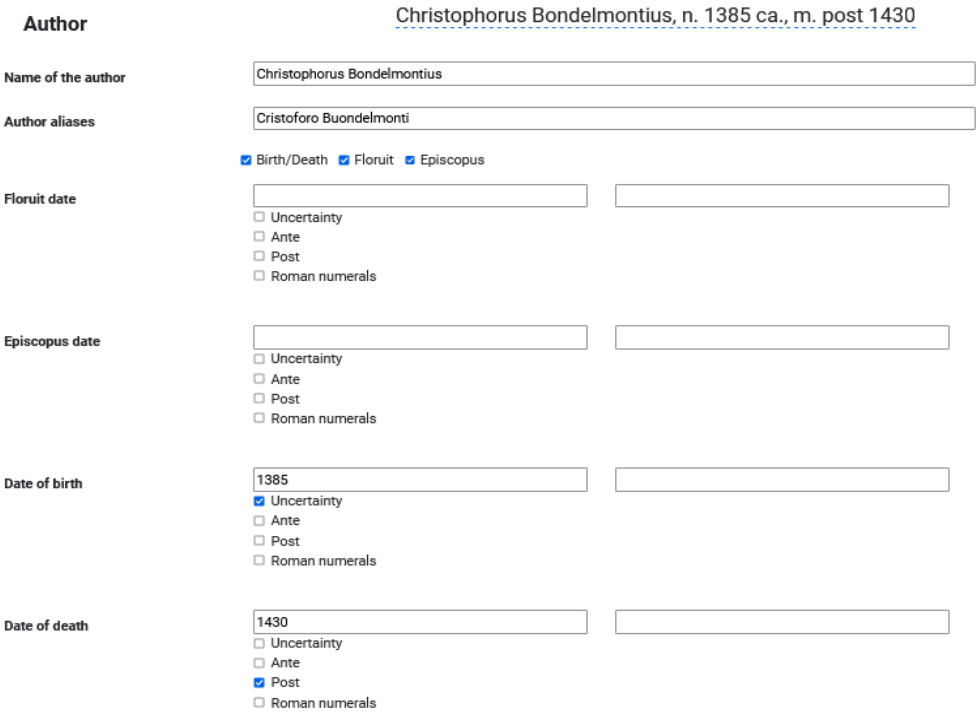

Figure 2: The tool interface to insert the dates about the authors.

[8] https://www.python.org/
[9] https://www.djangoproject.com/
[10] https://www.postgresql.org/

5. Conclusions and Future Work

In this paper, we have presented the Italian National Research Project IMAGO - Index Medii Aevi Geographiae Operum (2020-2023). IMAGO aims at creating a knowledge base of the Medieval and Renaissance geographic works which report the description and representation of the world in the VI-XV centuries.

To the best of our knowledge, until now no scientific research has applied digital methods in a systematic way in this field of study. The knowledge included in the IMAGO KB is formally represented using the languages of the Semantic Web (OWL 2 DL) and published as Linked Open Data. Furthermore, the produced data are compliant with the Findable, Accessible, Interoperable and Reusable (FAIR) paradigm. In particular, in this paper we have presented the ontology we have developed to formally represent the knowledge. The IMAGO ontology has been implemented as an extension of two standard vocabularies: CIDOC CRM and FRBRoo (and its ongoing extension LRMoo). The ontology has been populated using automatic and semi-automatic tools. We have been able to automatically import knowledge from existing KBs and databases, e.g. Wikidata, Pleiades, Mapping Manuscript Migration. We have also developed a semi-automatic web tool that is currently used by the scholars who are inserting data in our KB. We have performed a first evaluation of the ontological model using an automatic system.

As future work we have planned to perform a complete evaluation of the IMAGO ontology, regarding not only the ontological model but also the instances, using the same automatic system. Furthermore, we have also planned to conduct an evaluation involving users. In particular, we have planned to propose a specific questionnaire to the scholars who are currently populating the ontology. After the analysis of the evaluation results, if necessary, we will review and extent the ontology. The final aim of the project is to develop a web application that allows users (i.e. scholars, students but also general users) to retrieve and consult the data collected in the IMAGO KB in a user-friendly way (e.g. tables, maps, CSV files).

6. References

[1] Bartalesi V., Metilli D., Pratelli N., Pontari P.: Towards a knowledge base of medieval and renaissance geographical Latin works: The IMAGO ontology, Digital Scholarship in the Humanities, 2021, fqab060, https://doi.org/10.1093/llc/fqab060.

[2] Bauer, F., Kaltenböck, M.: Linked open data: The essentials. Edition mono/monochrom, Vienna 710 (2011).

[3] Burrows, T., Emery, D., Fraas, M., Hyvönen, E., Ikkala, E., Koho, M., Lewis, D., Morrison, A., Page, K., Ransom, L., Thomson, E., Tuominen, J., Velios, A., Wijsman, H.: Mapping manuscript migrations: Digging into data for researching the history and provenance of medieval and renaissance manuscripts: White paper (August 2020), https://diggingintodata.org/file/1281/download?token=x59u8fFQ.

[4] Doerr, M., Bekiari, C., LeBoeuf, P.: Frbroo, a conceptual model for performing arts. In: 2008 Annual Conference of CIDOC, 2008, pp. 06–18. CIDOC – ICOM International Committee for Documentation.

[5] Doerr, M.: The cidoc conceptual reference module: An ontological approach to semantic interoperability of metadata. AI Mag. 24(3), 2003, pp. 75–92.

[6] Pauwels, P., Zhang, S., and Lee, Y. C.. Semantic web technologies in AEC industry: A literature overview. Automation in Construction, 73, 145-165. (2017).

[7] Riva, P., Žumer, M.: Frbroo, the ifla library reference model, and now lrmoo: a circle of development. In: IFLA WLIC 2018 Conference, Transform Libraries, Transform Societies (2017).

[8] Simon, R., Isaksen, L., Barker, E. T., and de Soto Canamares, P. The Pleiades gazetteer and the Pelagios project.Bloomington: Indiana University Press (2016).

[9] Tartir, S., Arpinar, I.B., Moore, M., Sheth, A.P., Aleman-Meza, B.: Ontoqa: Metric-based ontology quality analysis (2005).

[10] Vrandečić, D.: Wikidata: A new platform for collaborative data collection. In: Proceedings of the 21st international conference on world wide web, 2012, pp. 1063–1064.
[11] Wilkinson M., “The addendum: Fair guiding principles for scientific data management and stewardship,” Scientific Data, vol. 6, 2019. doi:http://doi.org/10.1038/s41597-019-0009-6.
[12] Wilkinson M., “The fair guiding principles for scientific data management and stewardship,” Scientific Data, vol. 3, 2016. doi:http://doi.org/160018.10.1038/sdata.2016.18.