Towards a knowledge base of medieval and renaissance geographical Latin works: The IMAGO ontology

Valentina Bartalesi
[1]ISTI-CNR, Pisa, Italy

[1]**Daniele Metilli**
[2]ISTI-CNR, Pisa, Italy and Dipartimento di Informatica, Università di Pisa, Pisa, Italy

Nicolò Pratelli
[3]ISTI-CNR, Pisa, Italy

Paolo Pontari
[4]Dipartimento di Filologia, Letteratura e Linguistica, Università di Pisa, Pisa, Italy

Abstract

In this article we present the first achievement of the Index Medii Aevi Geographiae Operum (IMAGO)—Italian National Research Project (2020–23), that is, the ontology we have created in order to formally represent the knowledge about the geographical works written in Middle Ages and Renaissance (6th–15th centuries). The IMAGO ontology is derived from a strict collaboration between the Institute of Information Science and Technologies (ISTI) of the Italian National Research Council (CNR) and the scholars who are involved in the project, who have supported ISTI-CNR in defining a conceptualization of the domain of knowledge. Following the re-use logic, we have selected as reference ontologies the International Committee on Documentation CRM vocabulary and its extension FRBRoo, including its in-progress reformulation, LRMoo. This research is included in a wider project context whose final aim is the creation of a knowledge base (KB) of Latin geographic literature of the Middle Ages and Renaissance Humanism in which the data are formally represented following the Linked Open Data paradigm and using the Semantic Web languages. At the end of the project, this KB will be accessed through a Web application that allows retrieving and consulting the collected data in a user-friendly way for scholars and general users, e.g. tables, maps, CSV files.

Correspondence:
Valentina Bartalesi, ISTI-CNR, via Moruzzi 1, Pisa 56124, Italy.
E-mail:
valentina.bartalesi@isti.cnr.it

doi:10.1093/llc/fqab060

1 Introduction

Index Medii Aevi Geographiae Operum (IMAGO) is a 3-year (2020–23) Italian National Research Project (PRIN) that aims at realizing new tools, based on Semantic Web technologies, to support scholars in the study of geographical works written in Middle Ages and Renaissance. The tools will allow scholars to create and access a collection of Latin works that define the knowledge, description, and representation of the world in the 6th–15th centuries. In particular, the project aims at creating a knowledge base (KB) in which the data are formally represented following the Linked Open Data paradigm and using the languages of the Semantic Web. This KB will be accessed through a Web application that allows retrieving and consulting the collected data in a user-friendly way for scholars and general users, e.g. tables, maps, CSV files.

The image of the world that the Medieval and Renaissance culture created throughout ten centuries is crucial to understand the level of geographical knowledge and the development of western thought in European history. During the Middle Ages, geographical descriptions were mostly functional to collect the human knowledge into encyclopaedic works or to provide universal chronicles with an essential overview. Specific descriptions of lands, cities, places, monuments, and buildings were also supplied as a guide to the pilgrims travelling to the Holy Land, Rome, and Santiago de Compostela. By the end of the Middle Ages and the beginning of Renaissance Humanism, a more and more clear image of the World was defined thanks to the discovery of ancient geographical models (especially Greek works by Ptolemy and Strabo). After this period, the genre of geographical description had a further and decisive turning point, due to the exploration travels and discoveries; the description and representation of the New World, together with the reassessment of the physical space, brought about an epochal revolution. To the best of our knowledge, until now in this field of studies, no scientific research that has applied digital methods in a systematic way was conducted and an overall study, which highlights the importance of this literature from a historical–literary point of view is needed.

As the first step in order to develop tools to support scholars in creating, evolving and consulting a KB of the geographical works written in Medieval and Renaissance Humanism, we created an ontology that formally represents this knowledge. The IMAGO ontology is derived from a strict collaboration between ISTI-CNR and the scholars who are involved in the project who defined together a conceptualization of the domain of knowledge. Following the re-use logic, we selected as reference ontologies the International Committee on Documentation (CIDOC) CRM (Doerr, 2003) vocabulary and its extension FRBRoo (Doerr *et al.*, 2008), including its in-progress reformulation, LRMoo (Riva and Žumer, 2018).

The article is organized as follows: in Section 2, we report the state of the art of the studies of Latin geographic literature of the Middle Ages and Renaissance Humanism and the digital projects, archives, and ontologies that are useful to represent knowledge in this field. Section 3 describes the methodological approach we have followed to develop an ontology for the IMAGO project. Section 4 introduces in an informal way the knowledge about the geographical works we are interested in representing. In Section 5, we formally express the conceptualization and we describe the IMAGO ontology. In Section 6, our conclusions are reported.

2 State of the Art

Latin geographic literature of the Middle Ages and Renaissance Humanism (6th–15th centuries) has never been the subject of an overall and systematic scientific examination using digital methods so far. To start a research in this field, the recovery of the Medieval and Renaissance Humanistic geographical texts is necessary in order to make a full and complete screening of this literature from both historical–critical and philological–ecdotic point of views.

A fundamental framework to categorize Medieval travel literature is in Menestò (1993), in which the author defines the specific features of this literature across the Medieval centuries, making a clear classification of the literary genres (i.e. *itineraria*; *descriptiones*; narrations of the crusades, ambassadors, and missionaries reports; imaginary journeys; and

mirabilia). About the rebirth of geographical science in Renaissance Humanism, the most extensive and comprehensive survey is reported in Defilippis (2001). However, historical–geographical overviews are also reported in Bouloux (1999) and in Defilippis (2009). A significant starting point for the study of Latin travel and geographical texts in the Middle Ages and Renaissance Humanism and of their manuscript tradition can be found in critical editions and critical studies of specific works, such as the ones reported inStocchi (1963); Chiesa (2002); de Rubrouck (2011); and Pontari (2016). Furthermore, studies and critical editions realized by SISMEL[1] (Società Internazionale per lo Studio del Medioevo Latino/Italian Society for the study of Latin Middle Ages) constitute an authoritative corpus to study Medieval geography. Another important source of medieval geographical texts is the 'Repertorium fontium Historiae Medii Aevi' by Potthast (1962), which contains many information about works belonging to the geographical genre, their tradition, and critical bibliography.

To perform a complete and systematic analysis of geographical works, authoritative digital archives that are particularly interesting in our research are ALIM[2]—Archivio della Latinità Italiana del Medioevo (Archive of the Italian Latinity of the Middle Ages), MIRABILE[3]—Archivio digitale della cultura medievale (Digital Archive for Medieval Culture), and ENSU—Edizione Nazionale dei testi della Storiografia Umanistica (National Edition of Texts of the Humanism Storeography).

With regard to specific Web resources devoted to the field of geography, there are several online dictionaries, especially useful for detecting and normalizing toponyms. Among these, the most relevant in our project are the Getty Thesaurus of Geographical Names,[4] Histograph,[5] and Trismegistos.[6] However, these geographic dictionaries are limited resources that do not allow data interconnection. A step forward in sharing and reusing data is represented by two collaborative projects, such as Pelagios[7] and Pleiades.[8]

During the last years, some specific vocabularies have been developed to represent geographic knowledge. For example, the GeoNames ontology[9] allows representing the features of geographic places using the Web Ontology Language (OWL) (McGuinness and Van Harmelen, 2004). The GeoNames KB has collected over 11 million geographic places represented using the terms defined in the GeoNames ontology, and each place is denoted by an Internationalized Resource Identifier (IRI), following the Linked Data paradigm. The GO! ontology (Lana and Tambassi, 2017) is another vocabulary, developed within the Geolat project that allows access to the geographical knowledge contained in the classical Latin texts included in the digilibLT[10] digital library (DL). GO! describes the geographical entities with their boundaries, the mereological and topological relationships, the coordinates, their spatial representation and their literary, historical, and cultural features. A further geographical vocabulary is the Geographical Entity Ontology.[11] This ontology was developed to represent geopolitical entities (such as sovereign states and their administrative subdivisions) as well as various geographical regions (including but not limited to the specific ones over which the governments have jurisdiction). The Geographical Entity Ontology is implemented in OWL and based on the Basic Formal Ontology.[12] Finally, the Wikidata project[13] has defined a very large set of terms for representing geographic knowledge, including more than 29,000 classes and more than 700 properties expressing geographical relations.[14]

3 Methodology

As the first step of the project, the scholars have started working on a census of the Medieval and Renaissance Humanism geographical Latin texts. They are using as reference study the 'Repertorium fontium Historiae Medii Aevi' by Potthast (1962). However, the work of the census will not be limited to the collection of data from the repertory by Potthast, but it will be extended to other bibliographic tools and catalogues, such as the 'Iter italicum' by Kristeller (1963). A strong contribution comes from the MIRABILE and ENSU databases, especially with regards to the methods for classifying authors, texts and genres, manuscripts, editions, and historical–critical bibliography. At the same time, the scholars plan to create a Medieval Latin toponymy index. This index will be the first step towards the realization of an exhaustive catalogue that will collect specific lemmas related to Medieval Latin toponyms, providing a reference point, not available until

now, for detecting recurring place names into the texts of the Middle Ages and Renaissance Humanism.

The census of the Medieval and Renaissance geographical Latin texts is now concluded and at the same time, an ontology for representing the knowledge collected by the scholars was created. To develop the ontology, we followed these steps: (1) definition of a conceptualization of the domain of knowledge; (2) formalization of the conceptualization using standard ontologies as reference vocabularies; (3) development of the IMAGO ontology starting from the reference vocabularies; (4) population of the ontology; and (5) evaluation and refinement of the ontology. At the current stage of the project, we have gone through the first four steps. In particular, the conceptualization of the research domain has been defined in a formal way through the analysis of existing ontologies that are relevant to our work. These ontologies were selected with a preference for standards due to interoperability reasons. As reference vocabularies, we adopted the CIDOC CRM and FRBRoo (and its ongoing reformulation LRMoo). We created the IMAGO ontology as an extension of these two vocabularies. Finally, we developed a semi-automatic Web tool based on the ontology model to allow scholars to populate the ontology through a user-friendly interface, and we added functionalities to reduce the time needed to insert the knowledge. The steps of the development process are better described in the following sections of the article.

The methodology we followed to develop the ontology is well known and it is the one usually adopted to create formal vocabularies in the Semantic Web research field. The main novelty introduced by our work and our ontology 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. Indeed, no scientific research that has applied digital methods, and in particular Semantic Web approach, in a systematic way has been conducted in this specific field of research. Currently, such information 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. The IMAGO project aims at making this information available in digital form to both the scholars and the general users. The development of the ontology is integral to reach our project aim, that is, the creation of a KB in which the data are represented as a semantic graph that can be published as Linked Open Data. The choice of using the CRM and FRBRoo as reference ontologies and to develop the IMAGO ontology as an extension of both these vocabularies is also supported by the use of these standards within other recent research projects about manuscripts, e.g. the Mapping Manuscript Migrations[15] (MMM) (Burrows *et al.*, 2020) or the IRNERIO project (Barzaghi *et al.*, 2020). The MMM project transformed three separate datasets relating to the history and provenance of medieval and Renaissance manuscripts into a unified knowledge graph. The source databases are as follows: Schoenberg Database of Manuscripts, of the University of Pennsylvania; Bibale, from the Institut de recherche et d'histoire des textes (IRHT-CNRS, Paris); and medieval manuscripts in Oxford Libraries. The data consist of more than 20 million RDF triples that have been mapped to the MMM Data Model. The model combines classes and properties from CIDOC-CRM and FRBRoo, together with some specific MMM elements.

The IRNERIO project digitalized a collection of medieval texts of the Royal College of Spain in Bologna, Italy. One of the aims of the project is to provide a data model that allows to preserve the layering of information provided by different agents over time. Moreover, this data model allows preserving the description of the structure of the material, from both a physical and a conceptual point of view, in its particular components and features. To address these needs, the Medieval Manuscripts Ontology was developed and it was mainly designed around the FRBRoo vocabulary.

Although the above projects had different aims than IMAGO, having a KB compliant with them is a potentially huge advantage to link our data with other KBs. Indeed, this compliance allows extending the domain of representation and provides more knowledge on some aspects of the manuscripts present in different KBs. For example, the information on a manuscript present in the IMAGO KB can be extended with data about the history and the provenance of a manuscript from MMM KB or about glosses, annotations, and illustrations from IRNERIO KB. For this reason, we have conducted a preliminary study to map our

works and the manuscripts stored in the MMM KB. Indeed, especially the information about the manuscript migrations could significantly enrich the IMAGO KB. Since both MMM and IMAGO use the same reference vocabularies, the level of interoperability between the two ontologies is high. Querying the MMM KB through its SPARQL end-point, we measured that about 20% of the works collected in the IMAGO KB are also present in the MMM KB. We plan to integrate the knowledge related to these shared manuscripts to give more complete information to the users of the IMAGO Web application.

4 Conceptualization

This section introduces in an informal way the knowledge about the geographical works that we are interested in representing. On the basis of the studies and the methodological approach reported in Sections 2 and 3, respectively, the idea is that the domain of the geographical work can be represented using some main categories. The first ones are the author (this is, the author's name in Italian) and title (in Latin) of the works that were analysed. For each work, the literary genre has to be specified along with the toponyms that represent the locations that are described or reported into the work. Furthermore, for each work, several metadata about the related manuscripts and printed editions are reported.

Each manuscript and printed edition related to a work are described using several pieces of information.

In particular, for each manuscript the following knowledge is reported: the name of the author and the title of the work in the forms that appear in the manuscript; the library in which the manuscript is stored; 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, if they exist; the date of the creation of the manuscript; and 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. Figure 1 shows the categories described above in tabular format and for each category a value was reported from the work we chose as case study, that is, 'Descriptio insulae Cretae' by Cristoforo Buondelmonti (Florence, 1380/1390–1430).

5 The IMAGO Ontology

In order to formally express the conceptualization, we took into account some existing ontologies as references vocabularies. Of course, existing ontologies have been extended with notions that are suited to describe the domain we are interested in. However, it was paramount to minimize the number of such extensions, in order to reduce the idiosyncrasies in our research.

Two top ontologies were analysed in order to understand whether these are rich enough to capture the concepts described in Section 4. The first ontology is the CIDOC CRM (CRM for short), a high-level ontology that allows integrating the information contained in data of the cultural heritage domain along with their correlation with knowledge stored in libraries and archives (Doerr, 2003). The CRM achieves this by providing definitions and a formal structure for describing the implicit and explicit concepts and relationships used in cultural heritage documentation and of general interest for the querying and exploration of such data. Since December 2006, the CRM has been recognized as an official ISO standard. This status was renewed in 2014 and can be found at ISO 21127:2014.[16] As such, it offers a stronger guarantee under many aspects: it is widely known, it is regularly revised, and it is universally accessible. The Special Interest Group[17] of the CRM continuously works for expanding the domain of applicability of the ontology, and a number of extensions have been already devised.[18] The CRM has been successfully applied to the representation of knowledge in several fields, including narrative representation (Bartalesi *et al.*, 2017; Meghini et al., 2021), biography modelling

WORK	
AUTHOR	Buondelmonti Cristoforo (Firenze, 1380/1390 – 1430)
WORK	*Descriptio insulae Cretae*

MANUSCRIPT	
AUTHOR	Christofori Bondelmontis
WORK	Descriptio insulae Cretae
PLACE	Vatican City
LIBRARY	Vatican Apostolic Library
SEGNATURE	Rossiano 703
FOLIOS	ff. 1r-50v
INCIPIT OF THE DEDICATION/PROEM	-
EXPLICIT OF THE DEDICATION/PROEM	-
INCIPIT OF THE TEXT	LOREM IPSUM...
EXPLICIT OF THE TEXT	... LOREM IPSUM
DATE	1417/1422
SECONDARY SOURCES	Potthast, p. 1967, p. 606; DBI, XV, p. 199

PRINT EDITION	
AUTHOR	Cristoforo Buondelmonti
WORK	Descriptio insulae Cretae
CURATOR	E. Legrand
PLACE	Paris
DATE	1897
PUBLISHER	E. Letoux
FORMAT	-
PAGES	I-XL, 1-258
IMAGES	44 images of geographic maps out of the text
NOTES	-
AUTHOR OF INTRODUCTION – INTRODUCTION- DEDICATIONS	-
FIRST EDITION/REPRINT	First edition
PRIMARY SOURCES	Manuscrit du Serail
ECDOTIC TYPOLOGY	Critical edition with commentary with French translation
SECONDARY SOURCES	Potthast, p. 1967, p. 606; DBI, XV, p. 199

Fig. 1 The categories defined in the conceptualization along with the values related to 'Descriptio insulae Cretae' by Cristoforo Buondelmonti (1380/1390–1430)

(Tuominen *et al.*, 2018), craft heritage (Zabulis *et al.*, 2020), and archaeology (Niccolucci, 2017).

The second ontology we took into account is FRBRoo (Doerr *et al.*, 2008), including its in-progress reformulation, LRMoo.[19] FRBRoo is a formal ontology intended to capture and represent the underlying semantics of bibliographic information and to facilitate the integration, mediation, and interchange of bibliographic and museum information. The FRBR model was originally designed as an entity-relationship model by a study group appointed by the International Federation of Library Associations and Institutions during the period 1991–97 and was published in 1998. At the same time, the CRM was being developed independently from 1996 by the International Council for Museums–CIDOC (Documentation Standards Working Group. FRBRoo is based on the idea that both the library and museum communities might benefit from harmonizing FRBR with the CRM. A first version of FRBRoo was expressed in 2000 and was expanded in the following years. The latest major version of FRBRoo was published in October 2017, and a new version called LRMoo is currently in draft status. FRBRoo provides fundamental notions for text modelling that are important for our aims.

Analysing these two ontologies, we verified that they contain terms for representing all the categories and their characteristics described in the conceptualization. In the following sections, a detailed mapping is reported.

5.1 Representing authors, works, literary genres, manuscripts, and printed editions

As a notational convention, the CIDOC CRM uses the letters 'E' and 'P' to indicate classes and properties, respectively. On the other hand, FRBRoo (and its revisions LRMoo) uses the letters 'F' and 'R' to indicate classes and properties, respectively.

The two main categories of the conceptualization are Author and Work. In the IMAGO ontology, this knowledge is represented using the classes E39 Actor and F2 Expression. As reported in Fig. 2, the class F2 Expression is related to the class E39 Actor through the class F28 Expression Creation. F28 Expression Creation is linked to F2 Expression by the property

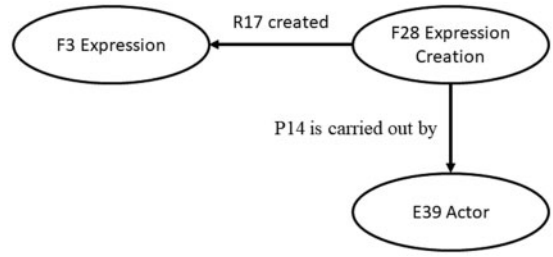

Fig. 2 A graphical view of the classes and properties used to represent the authors and works

R17 created and to the class E39 Actor by the property P14 is carried out by.

The class E39 Actor has linked with the class E41 Appellation (the author's name in Italian) through the property P1 is identified by. In general, to link the Appellation IRI with the corresponding literal (a string), we use the CRM property P190 has symbolic content.

The literary genre of the F2 Expression is represented using the Genre class that we defined as a subclass of E55 Type. F2 Expression is linked to the class Genre by the property 'has genre' we defined as a subproperty of P2 has type. The individuals of the class Genre are geographic work and travel literature.

As shown in Fig. 3, the toponyms are represented using the class Toponym we defined as a subclass of E41 Appellation. The class F2 Expression representing a work is linked to the class E53 Place by the property P67 refers to. The Place is linked to the class Toponym by the property 'is identified by toponym' that we defined as a subproperty of P1 is identified by. For each place, the corresponding geographical coordinates are reported in order to show this knowledge on a map in a later stage of the project. To represent this knowledge, the class Place is linked to the class E94 Space Primitive, representing the geographical coordinates, by the property P168 place is defined by. The Expression is linked to the Toponym contained in it by the property P106 is composed of.

The IMAGO ontology has to represent two types of resources: Manuscript and Printed edition.

The manuscript is represented using the class Manuscript that we defined as a subclass of F5 Item, and the printed edition using the class Printed edition that we defined as a subclass of F3 Manifestation.

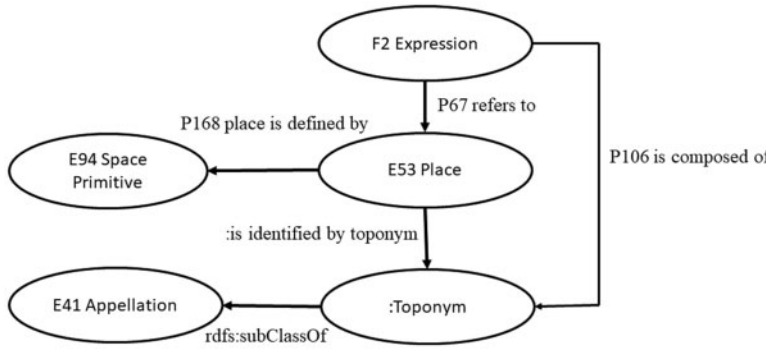

Fig. 3 A graphical view of the classes and properties used to represent the toponyms

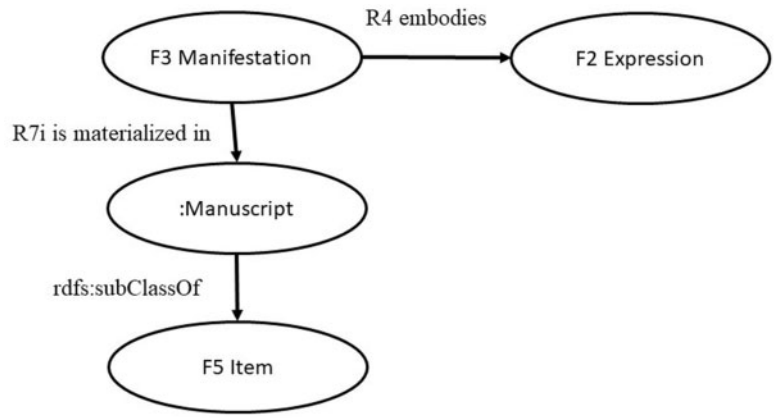

Fig. 4 A graphical view of the classes and properties used to represent the manuscript

In the following Sections 5.2 and 5.3, we report the classes and properties we used to represent the knowledge about manuscripts and printed editions.

5.2 Representing knowledge about manuscripts

The manuscript of a work is represented through the class Manuscript that we defined as a subclass of F5 Item. The class Manuscript is linked to the corresponding F3 Manifestation through the property R7i is materialized in. The Manifestation R4 embodies F2 Expression. Figure 4 shows the representation of the manuscript.

As reported in the conceptualization, we are interested in representing the following knowledge about a manuscript:

- *The name of the author as it is reported in the manuscript.* To represent this knowledge, we linked the class F28 Expression Creation with the class Manuscript through the direct property R18 created. F28 is related to the name of the author as it is reported in the manuscript using the class E41 Appellation. To link each appellation to the manuscript in which it appears, we use the property P106i forms part of.
- *The title of the work as it is reported in the manuscript* (this is the title of the whole manuscript). The Manuscript class is linked to the class E35 Title using the property P102 has title.
- *The library in which the manuscript is stored.* It is represented with the class F11 Corporate Body and it is related to the class Manuscript using the property P50 has current keeper.

- *The location of the library*. It is represented with the class E53 Place and it is linked to the class F11 Corporate Body through the property P74 has current or former residence. The class E53 Place is linked to the class E94 Space Primitive, representing the geographical coordinates, by the property P168 place is defined by.
- *Signature*. It is represented using the class E42 Identifier and it is related to the class Manuscript through the property P1 is identified by. To link the signature IRI with the corresponding string, we use the CRM property P190 has symbolic content.
- *Folios*. The class Manuscript is P46 is composed of E19 Physical Object that is P1 is identified by E41 Appellation.
- *The Incipit dedication/proem*. It is represented with the class E90 Symbolic Object. Each instance of the Symbolic Object class is linked to the corresponding string using the CRM property P190 has symbolic content. The Symbolic Object class is related with the corresponding manuscript using the property 'is incipit dedication of' that we defined as a subproperty of P106 is composed of.
- *The Explicit dedication/proem*. It is represented with the class E90 Symbolic Object. Each instance of the Symbolic Object class is linked to the corresponding string using the CRM property P190 has symbolic content. The Symbolic Object class is related with the corresponding manuscript using the property 'is explicit dedication of' that we defined as a subproperty of P106 is composed of.
- *The Incipit of the text*. It is represented with the class E90 Symbolic Object. Each instance of the Symbolic Object class is linked to the corresponding string using the CRM property P190 has symbolic content. The Symbolic Object class is related with the corresponding manuscript using the property 'is text incipit of' that we defined as a subproperty of P106 is composed of.
- *The Explicit of the text*. It is represented with the class E90 Symbolic Object. Each instance of the Symbolic Object class is linked to the corresponding string using the CRM property P190 has symbolic content. The Symbolic Object class is related with the corresponding manuscript using the property 'is text explicit of'' that we defined as a subproperty of P106 is composed of.
- *Date*. The date is represented with the class E52 Time Span and it is related to the class F30 Manifestation Creation using the property P4 has time span. F30 Manifestation Creation is linked to the Manuscript using the property R24 created.
- *Secondary sources*. To represent the secondary sources, we used the class Secondary Sources that we defined as a subclass of F3 Manifestation. We linked the class Secondary Sources with the class Manuscript using the class P129 is about.

5.3 Representing knowledge about printed editions

As reported in the conceptualization (Section 4), we are interested in representing the following knowledge about a printed edition:

- *Author*. To represent this knowledge, we linked the class F28 Expression Creation with the class F2 Expression using the property R17 created. Then, we linked the class Printed edition with the class F2 Expression using the property R4 embodies. Finally, we linked the F28 Expression Creation with the E39 Actor and then Actor with the E41 Appellation. To link each Appellation to the Printed edition in which it appears, we use the property P106i forms part of. Figure 5 shows this formal representation.
- *Title*. The class Printed edition is linked to the class E35 Title using the property P102 has title. To link the title IRI with the corresponding string we use the CRM property P190 has symbolic content.
- *Curator*. To represent this knowledge, we have introduced the class F30 Manifestation Creation that is linked to the class Printed Edition using the property R24 created. F30 is linked to the class Curator we defined as a subclass of the class E39 Actor using the property 'has curator' that is a subproperty of P14 carried out by. We linked the Curator class to the E41 Appellation.
- *Place*. We use the class E53 Place to represent this knowledge. The class F30 Manifestation Creation (linked to the class Printed edition) is linked to the class E53 Place using the property P7 took place at. The class E53 Place is linked to the class E94 Space

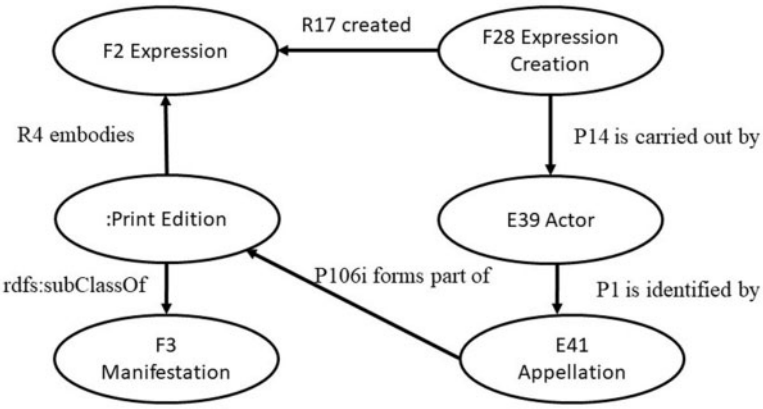

Fig. 5 A graphical view of the classes and properties used to represent the printed edition

Primitive, representing the geographical coordinates, by the property P168 place is defined by.

- *Date*. We use the class E52 Time Span to represent a date. The class F30 Manifestation Creation is linked to the E52 Time Span using the class P4 has time span.
- *Editor/Publisher*. F30 Manifestation Creation is linked to the class Publisher we defined as a subclass of the class E39 using the property 'has publisher' that we defined as a subproperty of P14 carried out by. We linked the Publisher class to the E41 Appellation.
- *Format*. To represent this knowledge, we used a subclass of the E55 Type, we called Format. The class Printed edition is linked to the class Format using the property R69 specifies physical form. We linked the class Format to E41 Appellation.
- *Pages*. To represent the pages of a printed edition, we adopted the class E90 Symbolic Object. The printed edition is linked to the class E90 using the property P106 is composed of. The class E90 is P1 identified by E41 Appellation.
- *Information about figures*. The class Printed edition is linked to a literal (string) that reports information about the figures present in the edition using the property 'has figure note' that we defined as subproperty of P3 has note.
- *Notes*. The class Printed edition is linked to a literal (string) using the property P3 has note.
- *Author of the introduction, dedications, introductions*. The class Printed edition is linked to a literal (string) using the property 'has introduction note' that we defined as a subproperty of P3 has note.
- *First edition/reprint*. This knowledge is represented using the class Edition that we defined as a subclass of the class E55 Type. The individuals of this class are first edition, reprint (associated with the year of publication) facsimile and anastatica. A printed edition is linked to the class Edition through the property P2 has type.
- *Primary Sources*. To represent the primary sources, we used the class Primary Source that we defined as a subclass of F3 Manifestation. We linked the class Printed edition with the class Primary Source using the property P67 refers to.
- *Ecdotic typology*. This knowledge is represented using the class Typology that we defined as a subclass of the class E55 Type. The printed edition is linked to the class Typology through the property P2 has type.
- *Secondary sources*. To represent the secondary sources, we used the class Secondary Source that we defined as a subclass of F3 Manifestation. We linked the class Secondary Source to the printed edition using the property P129 is about.

5.4 Representing knowledge as linked open data

Following the Linked Open Data paradigm, each resource that we will create in the KB will be identified by an IRI that allows accessing a description of the

resource. For identifying authors and works, we decided to use, where possible, IRIs from two existing KBs: the Wikidata KB and the MIRABILE database.[20] We selected Wikidata because it is one of the largest general-purpose KBs and contains thousands of descriptions of geographic entities, and MIRABILE because it is a specialized KB that describes many of the works, authors, and manuscripts that we aim to represent. The scholars provided us with a list of works and authors they intend to investigate during the project, and we mapped the entries of this list to the corresponding IRIs that we found in Wikipedia and MIRABILE. The mapping has been accomplished using a semi-automatic tool that we developed. The tool queries the two KBs and retrieves a set of matching IRIs. These IRIs are then checked by a human, who approves the result or, in case of multiple results, selects the correct one. If the tool finds an existing connection between the KBs (e.g. Wikidata links to MIRABILE), this connection is automatically imported in our KB.

We were able to automatically find 98% of the IRIs of the works that the scholars provided us in MIRABILE and Wikidata. For what concerns the authors, we found 96% of the IRIs from the same sources.

To identify the IRIs for libraries in which the manuscripts are present, we first automatically created a list of libraries (∼700 libraries) based on the manuscript names as reported in MIRABILE. Indeed, the names of the manuscripts contain, at the beginning, the information about the libraries in which they are collected, e.g. 'Città del Vaticano, Biblioteca Apostolica Vaticana, Vat. lat. 10497'. Secondly, performing SPARQL queries, we searched for the corresponding IRIs in Wikidata and we found 80% of the IRIs. For each library, we also extracted the IRIs of the countries and of the administrative–territorial entities in which each library is located, along with the corresponding geographical coordinates.

Finally, we automatically assign IRIs—in the form https://imagoarchive.it/resource/ID—to the other resources represented in our KB (e.g. printed edition, format, editor, signature, folios).

Using a formal ontology and the LOD paradigm to represent knowledge allows publishing KB content as Findability, Accessibility, Interoperability, and Reuse (FAIR) data. The 'FAIR Guiding Principles for scientific data management and stewardship' (Wilkinson *et al.*, 2016) provides guidelines to improve the FAIRof digital assets. The principles 'emphasise machine-actionability because humans increasingly rely on computational support to deal with data as the result of the increase in volume, complexity, and creation speed of data'.

Regarding the Findability principle, the data collected in the IMAGO KB are assigned a globally unique and persistent identifier (an IRI) and are indexed in a searchable resource.

The IMAGO data are accessible because they are retrievable through their identifier using a standardized communication protocol, i.e. the SPARQL query language.

The interoperability of the IMAGO data is guaranteed by the use of a formal, accessible, shared, and broadly applicable language for knowledge representation, i.e. OWL 2 DL.

Finally, the IMAGO data are easily reusable since the data are enriched with metadata that describe them, thus they can be reused in different contexts.

Notably, some of these fundamental principles, i.e. the importance to guarantee the discoverability and accessibility of the data and the possibility to easily share data within a research community, were already introduced informally by experts of Middle Ages studies in 2015 (Turnator *et al.*, 2015).

6 Ontology Population

To allow a simple and rapid ontology population, we decided to develop a semi-automatic Web tool. As specified in the ISO 9241-210 standard, understanding the needs and requirements of the users is the first step to develop successful systems and products (Doll and Torkzadeh, 1988). The first step in users' requirements analysis is to collect background information about the users and the processes that currently take place through structured interviews. In our project, we used this approach to identify a set of requirements on the modality of insertion, modification, and cancellation of the knowledge that has to be collected in the IMAGO KB. The users of this research are the scholars involved in our project. These scholars commonly use Web tools for their work; they use the tools made available by the DLs to carry out their work, they

use the search service of DLs to discover and access the relevant resources; finally, they rely on the DLs to disseminate and preserve the result of their work.

We interviewed two scholars, both experts in Italian literature and linguistics, who derived the user requirements reported below. For their experience and authoritativeness, these scholars can be considered as representatives of a large community of users. The tool should support the population of the ontology in an assisted way and should minimize the cognitive and technical burden on the user in the selection and identification of the involved resources. This goal requires supporting the following features:

- Automatically associating the IRIs to the resources collected through the tool.
- Identifying works, authors and libraries from a list of pre-defined options.
- Supporting the search on the Web of the geographical places that are interesting for the scholars.

We developed a prototype of the tool to illustrate the implemented requirements. We used feedback from the scholars to validate our technical solutions and to refine the requirements.

The population of the ontology has started and is being carried out by five scholars, who are experts in geographical medieval and Renaissance studies and who performed the census of the works to collect into the IMAGO KB.

From a technical point of view, the tool was developed using a Python backend with the Django[21] framework, and a frontend built with HTML5, JavaScript, and the Bootstrap[22] library. It takes as input a JSON file where the knowledge extracted from the census is stored and automatically shows each piece of knowledge in the corresponding field of the tool interface. To allow the scholars to easily retrieve the list of authors as resulted from the census, we implemented a search by author using a free text search or the list of authors' names paginated by initials, as shown in Fig. 6. After a scholar has retrieved the name of the author, she/he is interested in, the tool shows the list of the works of that author. We called lemma the pair 'author name/work title'.

At any time, the scholar can add an author name or a work title that is not present in the list at any time. Indeed, the tool provides a search functionality that allows retrieving entities from Wikidata through a pre-defined SPARQL query. If the searched entity is not present in Wikidata, the tool allows scholars to insert it manually, providing a name and a description of the entity.

Once the scholar has set the lemma through a user-friendly interface, she/he can insert the other pieces of knowledge, that is, (1) the place/s that the work refers to, (2) the genre/s of the work, and (3) information about the manuscripts and printed edition/s related to the work as it is described in detail in Section 4. The interface to insert place/s and genre/s is shown in Fig. 7.

The tool automatically assigns the IRIs to works, authors, places, genres, and libraries. For the works and authors, the IRIs are imported from MIRABILE or Wikidata. For the places and libraries, the IRIs are imported from Wikidata. The IRIs of a set of pre-defined genres are imported from the Nuovo Soggettario,[23] a standard thesaurus created and maintained by the National Central Library of Florence. When IRIs are not available in Wikidata, MIRABILE or in the Nuovo Soggettario, custom IRIs are automatically assigned by the tool.

This automatic assignment of the IRIs to the entities and the possibility of choosing an entity in a pre-defined list allow the scholars to reduce (1) the time for populating the ontology and (2) the possibility of making mistakes while inserting the data manually. The knowledge that is inserted by scholars through the tool is later converted into an OWL graph according to our ontology model, using a triplifier written in Java.

At the current stage of the project, our corpus includes 250 works, 206 authors, and 614 libraries and the scholars have started to insert detailed knowledge about manuscripts and printed edition of these works.

7 First Evaluation and Refinement of the Ontology

We performed a first evaluation of the ontological model. In particular, we conducted two different types of evaluation: an automatic evaluation and an

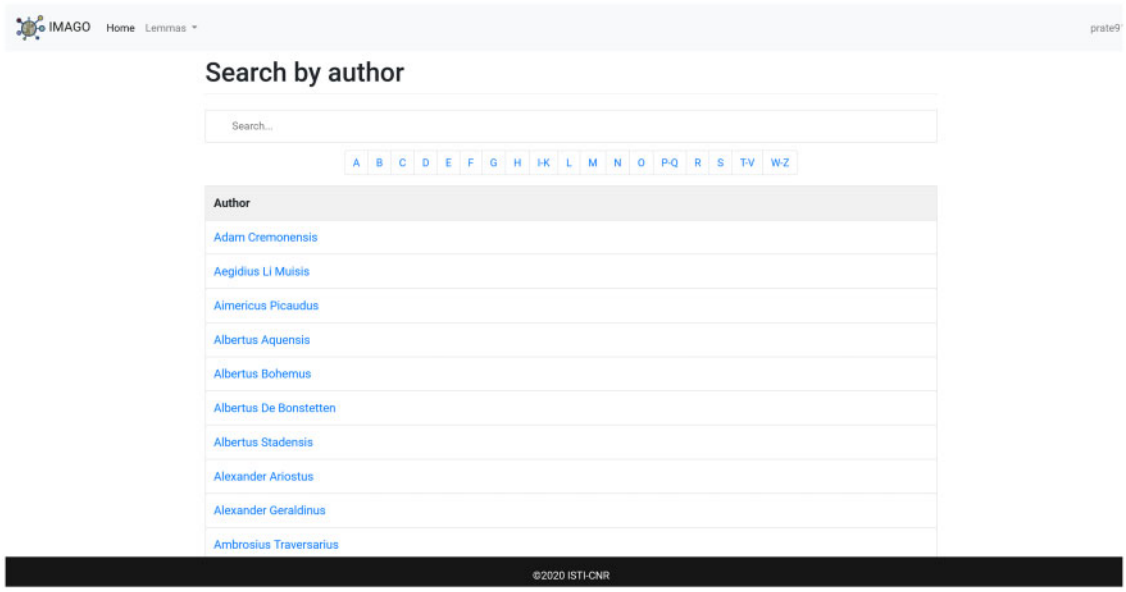

Fig. 6 The interface of the search by author

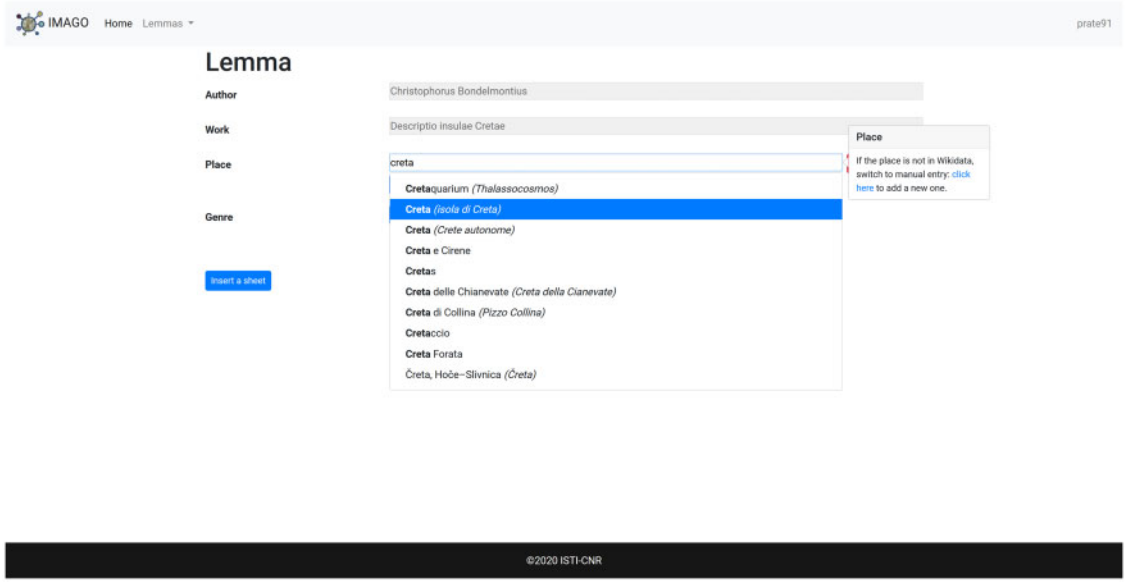

Fig. 7 The interface to insert place/s and genre/s

evaluation involving users. Methodologies and results of these two evaluation approaches are reported in the following sections. Since at this stage of the project the ontology population has just started, we only evaluated the ontological model, but we plan to also evaluate the IMAGO KB as future work.

7.1 Automatic evaluation of the IMAGO ontological model

For the automatic evaluation of the ontological model, we used the automatic OntoQA system (Tartir *et al*, 2005). OntoQA is a feature-based approach for evaluating ontologies that do not require data training. OntoQA evaluates the ontologies using a pre-defined set of metrics. In particular, the model considers how classes are organized in the model and how instances are distributed across the model. We chose OntoQA instead of other similar approaches, such as those reported in Tartir *et al.* (2010), because OntoQA provides metrics for evaluating both the model and the KB. Since the ontology population has just started, at this stage we only evaluated the model, but we plan to evaluate the IMAGO KB using this same software once the ontology population process is concluded. Furthermore, OntoQA provides a user-friendly application system that automatically computes metrics out of an input ontology.

The metrics and the results of the IMAGO ontological model evaluation are reported in Table 1.

The relationship richness (RR) metric evaluates the diversity of relations in the ontology. An ontology that contains many relations different from hierarchical relations (i.e. class–subclass relations or is–a) is richer than a taxonomy. The result of this metric is a percentage. If an ontology has an RR close to zero, then most of the relationships are is–a relationships. In contrast, an ontology with an RR close to 100 indicates that most of the relationships are different from an is–a relationship. The IMAGO ontology has a value of RR equal to 68.75%. This value denotes that an ontology is significantly richer than a simple taxonomy.

The inheritance richness (IR) indicates how the knowledge is grouped into different categories and subcategories in the ontology. Using this measure, we can distinguish between horizontal and vertical ontologies. A horizontal ontology has a small number of inheritance levels. In contrast, a vertical ontology contains a larger number of inheritance levels where classes have a small number of subclasses. The result of the IR metric is a real number representing the average number of subclasses per class. An ontology with a low IR is vertical and denotes that the ontology represents a very detailed knowledge. An ontology with a high IR is horizontal, i.e. the ontology represents general knowledge. The IR value of 1.66 of the IMAGO ontology indicates its vertical nature and defines it as a domain-specific ontology as indicated by its small number of subclasses per class. Notably, the IR values of general-purpose ontologies like SWETO (Aleman-Meza *et al.*, 2004) or TAP (Guha and McCool, 2003) have 4 and 5.36 IR values, respectively.

The attribute richness (AR) metric calculates the average number of attributes per class. The result is a real number. An ontology with a high AR value indicates that each class has averagely a high number of attributes, i.e. it is specified in detail, whereas a low value indicates that little information is provided about each class. Up to now, we have only introduced, for each class and property, the attribute rdfs: comment; thus, the AR value of the IMAGO ontology is still low (0.43).

Table 1. The OntoQA metrics and the values of the IMAGO ontology

Metrics	IMAGO ontology value
RR	68.75%
IR	1.66
AR	0.42

7.2 User evaluation of the IMAGO ontological model

The IMAGO ontology is currently being populated by four scholars, two University professors and two PhD students, all experts in Italian and Latin literature and linguistics, with a special focus on the Medieval and Renaissance geographical literature. The tool was used for 3 weeks by the scholars, and after that we organized a meeting to gather and discuss their comments about the ontological model. Since this was the first evaluation of the ontology, we chose the think-aloud method (Van Someren *et al.*, 1994) to allow scholars to freely express problems, suggestions, and requests. Following this method, the scholars have simply verbalized their thoughts as they move through the user interface of the tool and consequently through the ontological model on the top of which the tool was developed. In this section, we only reported the comments of the scholars about the ontological model. Indeed, a usability evaluation of the population tool is outside the scope of this article.

The scholars have appreciated the ability of the model to represent in a satisfactory way all concepts and relations reported in the conceptualization. However, managing the model through the tool, they have realized that some pieces of knowledge are still missing. In particular, all the scholars agreed that the following additional pieces of knowledge have to be represented by the model to provide a complete description of the manuscript and the printed edition:

- the manuscript has to be linked to a Web page containing a description or a digitalized reproduction of the manuscript;
- the manuscript has to be linked to a note field that reports the comments of the experts about the manuscript;
- the manuscript has to be linked to a string that reports the information about its iconographic apparatus; and
- the printed edition is currently linked to the publication place (current name and IRI). The experts suggested to also link the print edition to the name of the publication place as it is reported in the printed edition, e.g. Argentoratum, the ancient name of Strasbourg.

The requests of the scholars were easily integrated into our ontological model. The ability of the ontology to be expanded in a simple way denotes the strength of the model, thanks to its semantic interoperability. We reported below the technical representation of the pieces of knowledge listed above.

- *Link to the description or reproduction of the manuscript*. The class Manuscript is linked through the property P129 is about to the class E73 Information Object that represents the Web page in which the manuscript is described or reproduced.
- *Notes*. The class Manuscript is linked to a literal (string) using the property P3 has note.
- *Iconographic apparatus*. The class Manuscript is linked to a literal (string) that reports information about the iconographic apparatus using the property 'has iconographic apparatus' that we defined as subproperty of P3 has note.
- *Publication place*. We use the class E53 Place to represent this knowledge. The class F30 Manifestation Creation (linked to the class Printed edition) is linked to the class E53 Place using the property P7 took place at. The class E53 Place is linked to the class E94 Space Primitive, representing the geographical coordinates, by the property P168 place is defined by. Furthermore, the class E35 Place is P1 identified by the class E41 Appellation that reports the current name of the place. Finally, E35 Place is linked to the name as reported in the printed edition through the property 'is identified in the printed edition by' that we defined as a subproperty of P1. To manage the case of more than one printed editions for the same work, we added also a direct link between the publication place and a subclass of E41 Appellation we called Publication Place Appellation.

8 Conclusion

In this article, we have presented an ontology developed within the PRIN IMAGO. The ontology aims to formally represent the knowledge about geographical Latin works, including manuscripts and printed editions, which report the description and representation of the world in the 6th–15th centuries. Generally speaking, IMAGO aims at creating a KB in which the data about these works are formally represented following the Linked Open Data paradigm and using the languages of the Semantic Web. Indeed, to the best of our knowledge, until now no scientific research has applied digital methods in a systematic way in this field of studies. In this article, we have reported the methodological approach that we have followed to develop the ontology. First, we have defined a conceptualization of our domain of interest, and then we have formally expressed it using two standard ontologies as reference vocabularies: the CIDOC CRM and FRBRoo (and its ongoing revision LRMoo). A detailed mapping between the concepts of the conceptualization and the classes of these ontologies are also reported.

We have developed a semi-automatic Web tool that allows scholars to populate the ontology with the data they are retrieving and collecting about geographical Latin works. We have performed a preliminary evaluation of the ontological model, showing that the model is adequate for the representational

aims of the project. We plan to conduct a full evaluation of the ontology (model and KB) and of the tool to populate the ontology as an output of IMAGO project.

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 a better insight into this field from many perspectives, such as the Medieval Latin toponyms and the identification of historical places. The Web application will host a special section of Medieval and Renaissance cartography as well, in order to provide a digital collection of the most interesting maps and drawings.

Notes

1. https://www.sismelfirenze.it/.
2. http://en.alim.unisi.it.
3. http://www.mirabileweb.it.
4. http://www.getty.edu/research/tools/vocabularies/tgn/index.html.
5. http://histograph.io.
6. https://www.trismegistos.org.
7. http://pelagios.org.
8. https://pleiades.stoa.org.
9. http://geonames.org.
10. http://digiliblt.lett.unipmn.it.
11. https://www.ebi.ac.uk/ols/ontologies/geo.
12. https://github.com/bfo-ontology/BFO/wiki.
13. https://www.wikidata.org.
14. The classes are subclasses of "Q27096213 geographic entity" and can be retrieved through the following query: https://w.wiki/hsF. A list of geography-related properties is available at: https://www.wikidata.org/wiki/Wikidata: List_of_properties/geography.
15. https://mappingmanuscriptmigrations.org/.
16. https://www.iso.org/standard/57832.html.
17. http://www.cidoc-crm.org/sig-members.
18. http://www.cidoc-crm.org/collaborations.
19. http://www.cidoc-crm.org/frbroo/ModelVersion/lrmoo-f.k.a.-frbroo-v.0.6.
20. http://www.mirabileweb.it.
21. https://www.djangoproject.com.
22. https://getbootstrap.com.
23. https://thes.bncf.firenze.sbn.it/.

References

Aleman-Meza, B., Halaschek, C., Sheth, A., et al. (2004). SWETO: large-scale semantic Web test-bed. In *Proceedings of the 16th International Conference on Software Engineering and Knowledge Engineering (SEKE 2004): Workshop on Ontology in Action*, Banff, Canada, 21–24 June, pp. 490–93.

Bartalesi, V., Meghini, C., and Metilli, D. (2017). A conceptualisation of narratives and its expression in the CRM. *International Journal of Metadata, Semantics and Ontologies*, **12**(1): 35–46.

Barzaghi, S., Palmirani, M., and Peroni, S. (2020). Development of an ontology for modelling medieval manuscripts: the case of Progetto IRNERIO. *Umanistica Digitale*, **9**: 117–40.

Burrows, T., Emery, D., Fraas, M., et al. (2020). Mapping manuscript migrations knowledge graph: data for tracing the history and provenance of medieval and renaissance manuscripts. *Journal of Open Humanities Data*, **6**(1): 3.

Bouloux, N. (1999). *Culture et Savoirs Géographiques dans l'Italie du XIVe Siècle*. Doctoral dissertation, Paris 1.

Chiesa, P. (2002). Introduction. In *The Travels of Friar Odoric*, transl. by Sir Henry Yule, Grand Rapids (Michigan). Cambridge: William B. Eerdmans Publishing Company, pp. 1–62.

de Rubrouck, G. (2011). Viaggio in Mongolia: (Itinerarium). A cura di Paolo Chiesa. Fondazione Lorenzo Valla, Arnoldo Mondadori Editore.

Defilippis, D. (2001). La rinascita della corografia tra scienza ed erudizione. Bari, Adriatica editrice.

Defilippis, D. (2009). Da Flavio Biondo a Leandro Alberti: corografia e antiquaria tra quattro e cinquecento. Atti del Convegno di Studi, Foggia, 2 febbraio 2006, Volume 41 di Biblioteca di critica e letteratura.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine*, **24**(3): 75.

Doerr, M., Bekiari, C., and LeBoeuf, P. (2008). FRBRoo, a conceptual model for performing arts. In *2008 Annual Conference of CIDOC*, Athens, pp. 15–18.

Doll, W. J. and Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly*, **12**: 259–74.

Guha, R. and McCool, R. (2003). TAP: a semantic web test-bed. *Journal of Web Semantics*, **1**(1): 81–87.

Kristeller, P. O. (ed.) (1963). *Iter Italicum. Vol. 6: Italy III and Alia itinera IV*. Leiden, Netherlands: Brill.

Lana, M. and Tambassi, T. (2017). Eliciting the Ancient Geography from a Digital Library of Latin Texts. In Digital Libraries and Multimedia Archives. Communications in Computer and Information Science. Cham: Springer International Publishing, pp. 191–200.

McGuinness, D. L. and Van Harmelen, F. (2004). OWL web ontology language overview. *W3C Recommendation*, **10**(10): 2004.

Meghini, C., Bartalesi, V., and Metilli, D (2021). Representing narratives in digital libraries: the narrative ontology. *Semantic Web*, **12**(2): 241–64. doi:10.3233/SW-200421

Menestò, E. (1993). Relazioni di viaggi e di ambasciatori. In diretto da Cavallo, G., leonardi, C. and Menestò, E. (eds), *Lo spazio culturale del Medioevo latino, 1, Il Medioevo latino, II, La produzione del testo*, Roma: Salerno Editrice, pp. 535–600.

Niccolucci, F. (2017). Documenting archaeological science with CIDOC CRM. *International Journal on Digital Libraries*, **18**(3): 223–31.

Pontari, P. (2016). «Nedum mille qui effluxerunt annorum gesta sciamus». L'Italia di Biondo e l'invenzione del Medioevo.

Potthast, A. (1962). Repertorium fontium historiae medii aevi. Romae: Istituto storico italiano per il medio evo.

Riva, P. and Žumer, M. (2018). FRBRoo, the IFLA Library Reference Model, and Now LRMoo: A Circle of Development. In *Proceedings of IFLA WLIC 2018 – Kuala Lumpur, Malaysia – Transform Libraries, Transform Societies in Session 74.*

Stocchi, M. P. (1963). Tradizione medievale e gusto umanistico nel "De montibus" del Boccaccio.*Pubblicazioni della Facoltà di lettere e filosofia, Università di Padova Vol. 39. Firenze: Olschki.*

Tartir, S., Arpinar, I. B., Moore, M., Sheth, A. P., and Aleman-Meza, B. (2005). OntoQA: metric-based ontology quality analysis. *IEEE ICDM 2005 Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*, Houston, TX.

Tartir, S., Arpinar, I. B., and Sheth, A. P. (2010). Ontological evaluation and validation. In *Theory and Applications of Ontology: Computer Applications.* Dordrecht: Springer, pp. 115–30.

Tuominen, J. A., Hyvönen, E. A., and Leskinen, P. (2018). Bio CRM: a data model for representing biographical data for prosopographical research. In *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017), CEUR Workshop Proceedings.*

Turnator, E., Bolintineanu, A., Rose-Steel, T., Whearty, B., and Widner, M. (2015). *Summary of Proceedings of the "Linking the Middle Ages" Workshop*, 11–12 May 2015, University of Texas, Austin.

Van Someren, M. W., Barnard, Y. F., and Sandberg, J. A. C. (1994). *The Think Aloud Method: A Practical Approach to Modelling Cognitive.* London: Academic Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, **3**(1): 1–9.

Zabulis, X., Meghini, C., Partarakis, N., et al. (2020). Representation and preservation of heritage crafts. *Sustainability*, **12**(4): 1461.