Using Semantic Web to Create and Explore an Index of Toponyms Cited in Medieval Geographical Works

VALENTINA BARTALESI*, Institute of Information Science and Technologies "Alessandro Faedo" (ISTI), National Research Council (CNR), Italy
NICOLO' PRATELLI*, Institute of Information Science and Technologies "Alessandro Faedo" (ISTI), National Research Council (CNR), Italy and Department of Information Engineering, University of Pisa, Italy
EMANUELE LENZI*, Institute of Information Science and Technologies "Alessandro Faedo" (ISTI), National Research Council (CNR), Italy
PAOLO PONTARI, Department of Philology, Literature and Linguistics - University of Pisa, Italy

Western thought in European history was mainly affected by the image of the world created during the Middle Ages and Renaissance. The most popular reason to travel during the Middle Ages was taking a pilgrimage. Jerusalem, Rome, and Santiago de Compostela were the most popular destinations. It is not surprising that a lot of works written by travellers as guides for pilgrims exist. By the beginning of the Renaissance, a more precise image of the world was defined thanks to the discovery of ancient geographical models, especially the work of Ptolemy. The three years (2020-2023) Italian National research project IMAGO - Index Medii Aevi Geographiae Operum - aims to provide a systematic overview of the medieval and renaissance Latin geographical literature using the Semantic Web technologies and the LOD paradigm. Indeed, until now, this literature has not been studied using digital methods. In particular, this paper presents how we formally represented the knowledge about the toponyms, or place names, in the IMAGO ontology. To maximise the interoperability, we developed the IMAGO ontology as an extension of two reference vocabularies: the CIDOC CRM and its extension FRBRoo, including its in-progress reformulation, LRMoo. Furthermore, we used Wikidata as reference knowledge base. As case study, we chose to represent the knowledge related to the toponyms cited by the Italian poet Dante Alighieri in his Latin works. We carried out a first experiment for visualising the knowledge about these toponyms on a map and in the form of tables and CSV files.

Additional Key Words and Phrases: Semantic Web, Linked Open Data, ontology, toponyms, CIDOC CRM, Wikidata, Dante Alighieri

1 INTRODUCTION

The world's image created during the Middle Ages and Renaissance was crucial to the development of Western thought in European history. Especially in the Middle Ages, most people did not travel, and their knowledge and experience of the world were restricted to their immediate location. The most popular reason to travel was

*Both authors contributed equally to this research.

Authors' addresses: Valentina Bartalesi, Institute of Information Science and Technologies "Alessandro Faedo" (ISTI), National Research Council (CNR), via G. Moruzzi 1, Pisa, , Italy, 56124, valentina.bartalesi@isti.cnr.it; Nicolo' Pratelli, Institute of Information Science and Technologies "Alessandro Faedo" (ISTI), National Research Council (CNR), via G. Moruzzi 1, Pisa, , Italy, 56124 and Department of Information Engineering, University of Pisa, Via G. Caruso 16, Pisa, Italy, nicolo.pratelli@isti.cnr.it; Emanuele Lenzi, Institute of Information Science and Technologies "Alessandro Faedo" (ISTI), National Research Council (CNR), via G.Moruzzi 1, Pisa, , Italy, 56124, emanuele.lenzi@isti.cnr.it; Paolo Pontari, Department of Philology, Literature and Linguistics - University of Pisa, piazza Evangelista Torricelli, 2, Pisa, Italy, paolo.pontari@unipi.it.

1556-4673/2023/8-ART111 $15.00
https://doi.org/10.1145/3582263

Fig. 1. Matthew Paris's Map of the Route to Jerusalem. St. Albans, c.1250. London, British Library, Royal MS, 14 C vii, f. 5.

taking a pilgrimage. Jerusalem, Rome, and Santiago de Compostela were the most popular destinations. These trips would take months and require elaborate planning [17]. It is not surprising that there are works written by travellers as guides for pilgrims [25]. These guides report specific descriptions of lands, cities, places, monuments and buildings, as well as advice about what to pack, what to wear, and where and how to sleep [1]. An example of these guides is Matthew Paris's itinerary maps from London to Palestine. Around 1250, Mathew Paris, a monk, chronicled events from the creation of the world to the mid 13th century in his work *Chronica Major*. At the beginning of a manuscript of this work, we find a map showing the pilgrimage route from London to Jerusalem (Figure 1). This medieval road map is spread over seven pages and takes the viewer on a journey from London through significant stops on the pilgrimage [10].

By the end of the Middle Ages and the beginning of the Renaissance, a more precise image of the world was defined thanks to the discovery of ancient geographical models, especially the work of Ptolemy. Ptolemy was a Greco-Roman geographer, author of *Geographia*, a work written about AD 150, in which all of the knowledge about the world's geography in the Roman Empire of the 2nd century was reported and described. In the 15th century, his work was rediscovered and translated into Latin from the original Greek. The influence of Ptolemy on the medieval cartography is due to his use of applying geographic coordinates to 8,000 geographic locations, as well as the use of mathematical calculations to depict geographical places. The genre of geographical description had a further and fundamental turning point during the period between the mid-15th to the mid-16th century, when a combination of circumstances encouraged people to seek new routes and to start the *Age of Discovery*. Indeed, toward the end of the 14th century, the empire of the Mongols had been broken up, and Western merchants could no longer be assured of safe conduct along the land routes. Furthermore, the Ottoman Empire and the Republic of Venice controlled commercial access to the Mediterranean and the ancient sea routes from the East. Thus, nations, especially on the Atlantic shores of Europe, started to seek overseas trade and explorations. These travels and discoveries allowed the reassessment of the physical space and the description and representation of the new routes and lands, and the New World gave the basis of modern geography [11].

The three years (2020-2023) Italian National research project IMAGO - Index Medii Aevi Geographiae Operum - aims to provide a systematic overview of the medieval and renaissance Latin geographical literature using the Semantic Web technologies [18]. Indeed, until now, this literature has not been studied using digital methods. 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 and to publish this knowledge as Linked Open Data (LOD) [6] and Findable, Accessible, Interoperable, and Reusable (FAIR) data [39]. In particular, the project wants to provide: (i) a collection of the manuscript tradition and printed editions for each work; (ii) a classification of authors, genres and contents; (iii) a collection of critical editions of some more representative works; (iv) a medieval Latin toponym index. As the first step to developing tools to support scholars in creating, evolving and consulting a knowledge base (KB) of the geographical works, we created an OWL 2 DL ontology [23] that formally represents this knowledge. OWL 2 DL guarantees the decidability of the language, and it has a sufficient representational power for the project aims. Following the re-use logic and to maximise interoperability, we developed our ontology as an extension of two reference ontologies, that is, the CIDOC CRM vocabulary [13] and its extension FRBRoo [14], including its in-progress reformulation, LRMoo [32]. We have also designed and developed a Web tool that allows scholars to populate the ontology. Then, we carried out a first experiment for visualising the knowledge about Dante Alighieri's Latin toponyms collected in our KB on a map and in the form of tables and CSV files. To guarantee the quality of the collected dataset, all data were produced by the authoritative scholars involved in the project. After the creation of the KB, two or more scholars manually checked all data for consistency to guarantee the absence of systematic errors.

This paper presents the process we followed to formally represent the knowledge about the toponyms in the IMAGO ontology (Section 3). To contextualise our work, in Section 2, we report an overview of the current datasets of geographical places that use Semantic Web technologies and the LOD paradigm to facilitate the interlinking of different sources. Section 4 reports a description of the semi-automatic tool we designed and developed to allow scholars to populate the ontology. In Section 5, we describe the visualisation components to access and explore the knowledge about the toponyms. As a case study, we chose to represent and visualise the knowledge related to the toponyms cited by the Italian poet Dante Alighieri in his Latin works. This choice is motivated by the fact that the ontology population phase is far from the end, and in this first period, the scholars involved in our project focused their research on Dante Alighieri's works. Finally, in Section 6, some concluding remarks are given.

2 RELATED WORKS

Medieval and renaissance Latin geographical works report many place names or toponyms and natural descriptions. Having toponyms in a standardised digital form allows scholars to compare their toponym datasets with other datasets, enrich their datasets with additional knowledge, and produce aggregated views on the data, e.g. aggregation of current and historical information about a place. Semantic Web technologies and the LOD paradigm facilitate the sharing and the comparison of the data. A dataset is linked and open when it is structured and published according to the LOD principles [7], so that it can be both interlinked and made openly accessible and shareable on the Semantic Web. Indeed, the goal of LOD is to allow data from different resources to be interconnected and queried. To make this data understood and automatically processed by computers, it must be expressed in a standard format. In the Semantic Web, each entity (for example, a museum object, a place, or a person) must be represented by a persistent identifier (known as an Internationalised Resource Identifier (IRI)[1]). Furthermore, RDF/S (Resource Description Framework/Schema) [27] and OWL (Web Ontology Language) [24] are formal languages for describing entities, as well as the relationships between them as simple properties and values (known as *triples*). The Semantic Web and LOD approaches are currently used to represent knowledge

[1]https://www.w3.org/International/articles/idn-and-iri/

about toponyms. In the following, we report some well-known datasets of geographical places that use Semantic Web technologies and the LOD paradigm to facilitate the interlinking of different sources.

The first example is the Getty Thesaurus of Geographic Names[2] (TGN). It is a structured vocabulary containing over four million toponyms and associated information about places. Places include administrative entities (e.g., cities, nations) and physical features (e.g., mountains, rivers). Other information related to history, population, culture, art and architecture is included. While many records in TGN include geographic coordinates, these coordinates are approximate and are intended for reference only. Indeed, they typically represent a single point, corresponding to a point in or near the centre of the inhabited place, political entity, or physical feature. TGN is constructed to allow publishing its data as LOD, i.e. JSON, RDF, N3/Turtle, N-Triples.

Another widely adopted open dataset for geographical entities is the Geonames Geographical Database[3]. It contains over 27 million geographical names. In addition to the names of places in various languages, Geonames includes latitude, longitude, elevation, population, administrative subdivision and postal codes. The geographical coordinates, stored with all the names, enable combining the municipality with the place names used for an inhabited place that is not an administrative unit. As a result of this spatial combination, the web pages provide lists of all known settlements in a municipality. Each GeoNames feature is represented as a web resource identified by an IRI. This IRI provides access either to the HTML wiki page or to an RDF description of the feature, using elements of the GeoNames ontology[4]. This ontology describes the GeoNames features properties using the Web Ontology Language, the feature classes and codes being described in the SKOS language.

LinkedGeoData [12] is another effort to add a spatial dimension to the Web of Data. LinkedGeoData uses the information collected by the OpenStreetMap (OSM) project[5] and makes it available as an RDF knowledge graph according to the Linked Data principles. LinkedGeoData is based on an ontology that is derived from the OSM tags [35]. The ontology includes around 1,200 classes, 250 data properties, and 80 object properties.

Another example of a toponyms dataset is Wikidata [38], an open collaborative general-purpose knowledge base created by the Wikimedia Foundation. The knowledge base accepts editing by any user, following the model of Wikipedia. Wikidata currently contains more than 17 million items and thousands of descriptions of geographical entities. For each geographical entity, Wikidata reports the latitude and longitude of single points. Wikidata has reached full compatibility with Semantic Web technologies [15], and it provides a SPARQL endpoint to query the knowledge base. Since October 2014, a GeoNames ID external identifier has been reported for many geographical Wikidata entities.

Pleiades[6] gazetteer provides historical geographical information about the ancient world in digital form. At present, Pleiades has extensive coverage for the Greek and Roman world and is expanding into Ancient Near Eastern, Byzantine, Celtic, and Early Medieval geography. Pleiades is compliant with the Semantic Web technologies and uses terms from different vocabularies to represent the collected knowledge, e.g. Dublin Core, SKOS, FOAF[7], Citation Ontology[8]. The Pleiades dataset may be downloaded in RDF format.

The project Mapping Manuscript Migations (MMM) makes available a dataset of places relating to the history and provenance of Western European medieval and early modern manuscripts [9]. The MMM links disparate datasets from Europe and North America to provide a view of the provenance of medieval and renaissance manuscripts. The MMM data are available as LOD for reuse under a CC-BY-NC 4.0 licence. Another project that makes available a geographic dataset as LOD is HiGeoMes. This project aims to integrate information on places mentioned in

[2]http://www.getty.edu/research/tools/vocabularies/tgn/index.html
[3]http://www.geonames.org/
[4]http://www.geonames.org/ontology/documentation.html
[5]https://www.openstreetmap.org
[6]https://pleiades.stoa.org/home
[7]http://xmlns.com/foaf/0.1/
[8]http://purl.org/spar/cito/

Babylonian and Assyrian texts with archaeological sites [21]. The linking of documented settlements with place names from written sources is intended to provide a better understanding of political, social and environmental developments in Upper Mesopotamia in the 2nd millennium BC. Finally, we want to mention an ongoing project LINKEDOPENNEPAL, which aims at creating an ontology for representing the Nepalese toponyms [37] attested in the documents collected in a text corpus that is part of the Documenta Nepalica[9]. This corpus comprises texts and documents on the history of religion and law of pre-modern Nepal. The ontology extends CIDOC CRM and increases interoperability by aligning the ontology individuals to the respective entries of GeoNames. Also, a mapping of the individuals to DBpedia entities was established.
As reported in detail in the following Section, to represent the knowledge collected in the IMAGO project, we defined an ontology as an extension of the standard vocabulary CIDOC-CRM, and its extension FRBRoo, including its in-progress reformulation, LRMoo. Indeed, the CRM provides terms to represent places, toponyms and geographic coordinates. In this representational context, we used Wikidata as reference knowledge base to represent the geographical entities we found in the works collected in the IMAGO project. We decided to use the IRIs of Wikidata since it is a general-purpose KB; thus, we could also find IRIs for other entities (e.g. authors, works, libraries) from the same source.

3 THE IMAGO ONTOLOGY

3.1 General Description of the Ontology

Despite in other research projects semantic technologies have been used to represent ancient manuscript corpora [5, 9, 16, 20], no scientific research that applies a Semantic Web approach has been conducted in the medieval and renaissance Latin geographical literature. Indeed, the main novelty introduced by our research is the use of Semantic Web technologies to formally represent this scientific domain. The knowledge related to these ancient works is currently dispersed on paper books. This makes a systematic overview of this geographical literature impossible, preventing a well-ordered perception of how it was gradually set up in time. The IMAGO project aims at making this information easily available in digital form for scholars, students and general users.
As the first step to develop tools to support scholars in creating, evolving and consulting a KB of the medieval and renaissance geographical works, we created an ontology that formally represents this knowledge. The ontology is extensively described in [2] and [3]. The IMAGO ontology is derived from a conceptualisation of the knowledge domain, which identifies the categories and the relations that have to be formally represented. This conceptualisation is the result of the collaboration between the CNR and the experts in Latin geographical literature who are involved in the project, i.e. scholars from the University of Pisa and the University of Salento[10]. The scholars identified the categories and relationships of the knowledge domain they were interested in formally representing. In particular, the conceptualisation is based on authoritative studies of the medieval and renaissance Latin geographical works, i.e. [28], [29], [25], [22], [26], [11], [8]. In our conceptualisation, we identified some main categories that represent the domain of the geographical works. The first categories are the author and title of a work. For each work, the literary genre is specified along with the toponyms that represent the places that are described or reported in the work. Furthermore, for each work, several metadata about the related manuscripts and printed editions are defined. In particular, for each manuscript, the following categories are identified:

- 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 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;

[9]https://nepalica.hadw-bw.de/nepal/
[10]https://imagoarchive.it/gruppo.html

- the *incipit* and *explicit* of the text;
- the date of the creation of the manuscript;
- the secondary sources;
- the iconographic apparatus;
- the link to the description or reproduction of the manuscript;
- the scholars' notes.

On the other hand, for each printed edition, the following categories are defined:

- the author's name and the title as reported in the edition;
- the curator's name;
- 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 to the edition;
- 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;
- the primary and secondary sources of the edition;
- and the ecdotic typology.

Once the conceptualisation was defined, we formalised it using classes and properties from two existing ontologies we chose as reference vocabularies, i.e. the CIDOC CRM and its extension FRBRoo, including its in-progress reformulation LRMoo. We adopted a lot of terms from these ontologies to maximise the interoperability of our representation. Where necessary, we added our own classes and properties to represent the terms that we did not find in the reference vocabularies. The resulting ontology is an extension of the CRM, and it is expressed in OWL 2 DL [23]. In Table 1, we report the classes that represent our main concepts and in Table 2, the properties to express the semantic relations among concepts are listed. 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.

Figure 2 shows the main classes and properties of the IMAGO ontology.

Following the Semantic Web approach, all resources stored in the IMAGO KB have a unique identifier, i.e. the Internationalized Resource Identifier (IRI). These IRIs are automatically retrieved from different sources freely available on the Web. In particular, we used the IRIs stored in the following datasets: (i) Wikidata knowledge base [38]; (ii) MIRABILE digital archive[11]; (iii) Mapping Manuscript Migrations knowledge base[19]; and (vi) Plaiades gazeteer[12]. When the IRIs were 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 and Mirabile. 95% of place IRIs and 41% of library IRIs from Wikidata. 50% of ancient place IRIs from Pleiades. Finally, we were able to map 20% of the works collected in the IMAGO project with those collected in the MMM KB.

Linking our KB to these different datasets allows significantly enriching the knowledge collected in the IMAGO project. Indeed, each resource has an associated description reporting new knowledge in comparison with the one collected by the scholars participating in the IMAGO project.

[11]www.mirabileweb.it

[12]https://pleiades.stoa.org/

Table 1. Classes used to represent the IMAGO main concepts

Concept	Class
Author	subclass of E39 Actor
Work	equivalent to F2 Expression
Work creation	equivalent to F28 Expression Creation
Genre	subclass of E55 Type
Toponym	subclass of E41 Appellation
Manuscript	subclass of F5 Item
Printed Edition	subclass of F3 Manifestation
Library	subclass of F11 Corporate Body
Place	equivalent to E53 Place
Geographic Coordinate	equivalent to E94 Space Primitive
Signature	equivalent to E42 Identifier
Folios	subclass to E19 Physical Object
Date	equivalent to E52 Time-Span
Curator/Publisher	subclass of E39 Actor

Table 2. Properties used to represent the relations among the IMAGO main concepts

Relation (R) between concepts	Property
R(Work creation event,Author)	equivalent to P14 is carried out by
R(Work creation event,Work)	equivalent to R17 created
R(Manuscript,Title)	equivalent to P102 has title
R(Printed Edition,Title)	equivalent to P102 has title
R(Manuscript,Library)	equivalent to P50 has current keeper
R(Place,Geographic coordinates)	equivalent to P168 place is defined by
R(Manuscript,Signature)	equivalent to P1 is identified by
R(Manuscript,Folios)	equivalent to P46 is composed of
R(Manuscript,Date)	equivalent to P4 has time span
R(Printed edition,Date)	equivalent to P4 has time span
R(Printed edition,Curator)	subproperty of P14 carried out by
R(Printed edition,Publisher)	subproperty of P14 carried out by
R(Printed edition,Format)	equivalent to R69 specifies physical form
R(Printed edition,Page)	equivalent to P106 is composed of

3.2 Representing Toponyms in the IMAGO Ontology

One of the aims of our ontology is the representation of the knowledge related to the toponyms that are reported in the medieval works we collected. In particular, the ontology is able to represent the following pieces of knowledge:

- the place identified by the toponym;
- the geographic coordinates associated with the place identified by the toponym;
- the work/s in which the toponym appears;
- the textual fragment in which the toponym appears;
- the textual place of the toponym;
- the author of the work in which the toponym is reported.

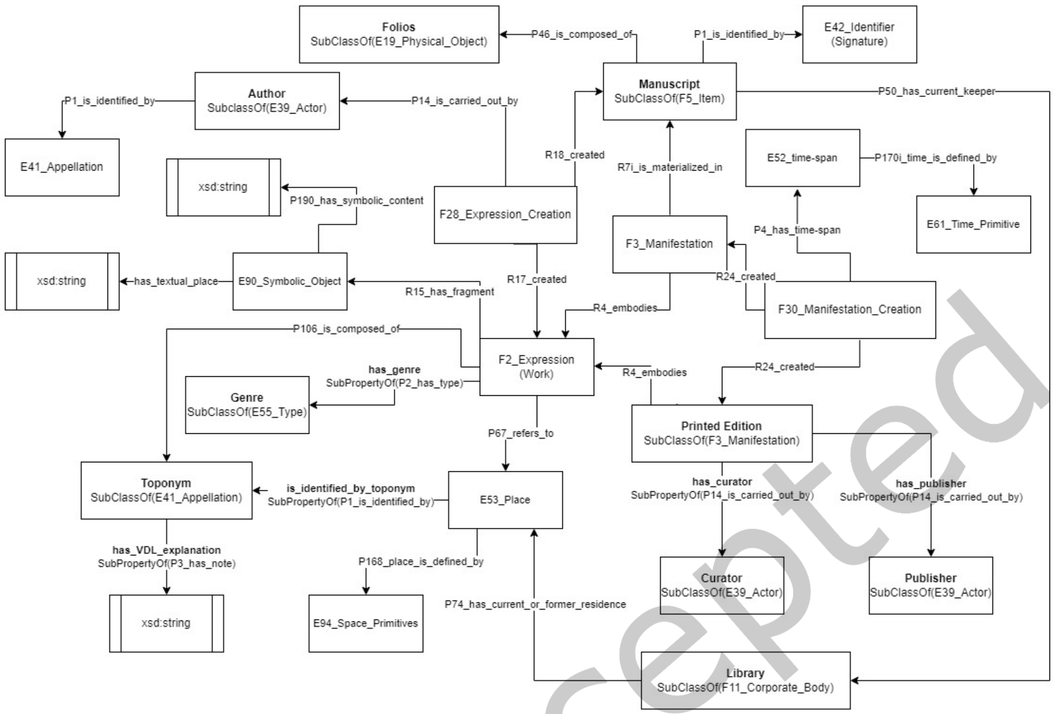


Fig. 2. The main classes and properties of the IMAGO ontology.

In the following, we report a running example to better clarify the meaning of the pieces of knowledge reported above. To test our model, we started to populate the ontology with the toponyms, and the related data, cited in the Latin works of the Italian poet Dante Alighieri[13]. The works we took into account are *De Vulgari Eloquentia*, *Monarchia*, *Epistole*, *Egloge*, and *Quaestio de Aqua et Terra*[14]. We chose the works of this author for the fundamental impact that they had on Italian and international literature. Furthermore, the scholars involved in the IMAGO project are authoritative experts in Dante Alighieri's literary production, thus they started to collect the toponyms cited in his works.

Figure 3 shows the ontology section representing the toponyms. A toponym is identified as an instance of the class Toponym that we created as a subclass of the class E41 Appellation. Each instance has an rdfs:label that links the toponym IRI with a string that represents its human-readable name. In our example, Dante Alighieri's toponym we want to represent is the Adriatic Sea, *Adria* in Latin. The class 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. Each instance of the class Place is identified by a Wikidata IRI, and each IRI is linked through the corresponding Pleiades instance using the property OWL:sameAs, which indicates that two IRIs actually refer to the same place. In the example, the place linked to *Adria* is identified with the corresponding Wikidata IRI (i.e. https://www.wikidata.org/wiki/Q13924), and also to the corresponding Pleiades IRI (i.e. https://pleiades.stoa.org/places/1004). For each place, the geographic coordinates are reported. To represent the geographic coordinates, we used the class E94 Space Primitive. The

[13]https://en.wikipedia.org/wiki/Dante_Alighieri
[14]https://dante.princeton.edu/pdp/

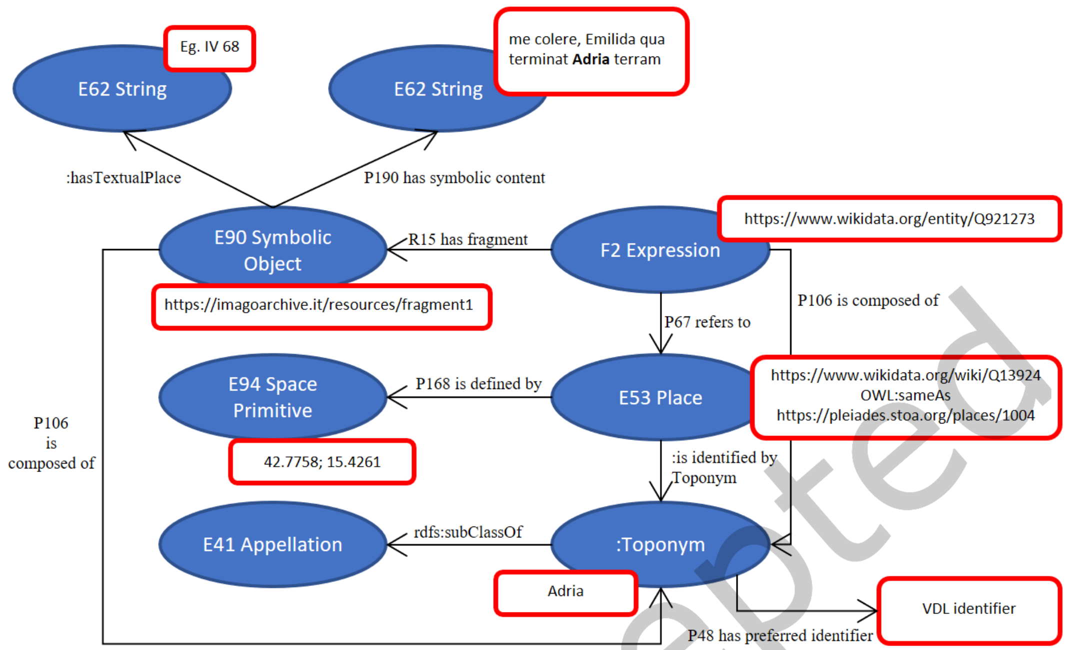


Fig. 3. The ontology section representing toponyms, along with a running example.

class Place is linked to the class E94 Space Primitive by the property P168 place is defined by. The class F2 Expression, which represents the work in which the toponym is cited, is linked to the class Place by the property P67 refers to. In our example, the work in which the Adriatic Sea is cited is the *Egloge*, identified with an Wikidata IRI (i.e. https://www.wikidata.org/entity/Q921273). The class F2 Expression is linked to the class Toponym by the property P106 is composed of. To provide the textual context of a toponym along with the textual place in which the toponym is cited, we used the property R15 has fragment, which links the classes F2 Expression and E90 Symbolic Object. Thus the E90 Symbolic Object is related to the textual context (a string) through the property P190 has symbolic content and to the textual place (a string) through the property :has textual place we defined. In our example, the textual place in which *Adria* is cited is "Eg. IV 68" (that means that the toponym is reported in the *Egloge*, 4th piece, versus 68) and the textual content is the sentence: "*Emilida qua terminat Adria terram*". Finally, the class Toponym is linked by the property "has preferred identifier" to the Vocabolario Dantesco Latino (VDL)[15], a web resource that reports linguistic explanations of the toponyms cited by Dante Alighieri.

4 THE ONTOLOGY POPULATION

To populate the ontology, we developed a semi-automatic Web tool to allow scholars to insert knowledge through a user-friendly interface. The tool was created to reduce the time to insert knowledge and to avoid the insertion of mistakes thanks to the use of predefined lists of works, authors, libraries, places, and literary genres. The geographic coordinates of the places are also automatically assigned. The labels and the IRIs associated with the resources included in these predefined lists are extracted from the datasets reported in Section 3. Figure 4 shows the main interface of the tool. Through this interface, the scholars can insert the title of a work, the author, the

[15] https://www.vocabolariodantescolatino.it/

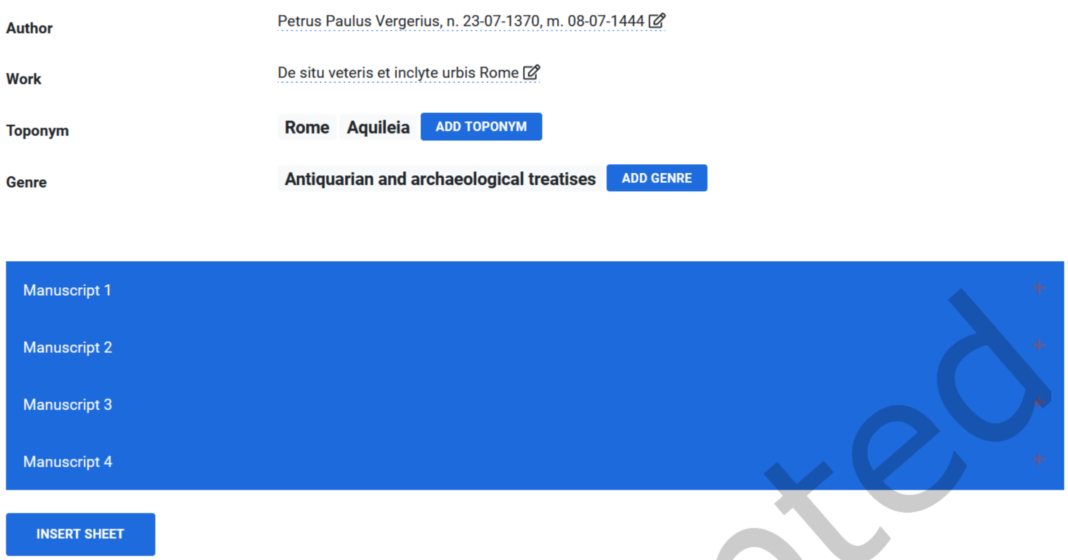
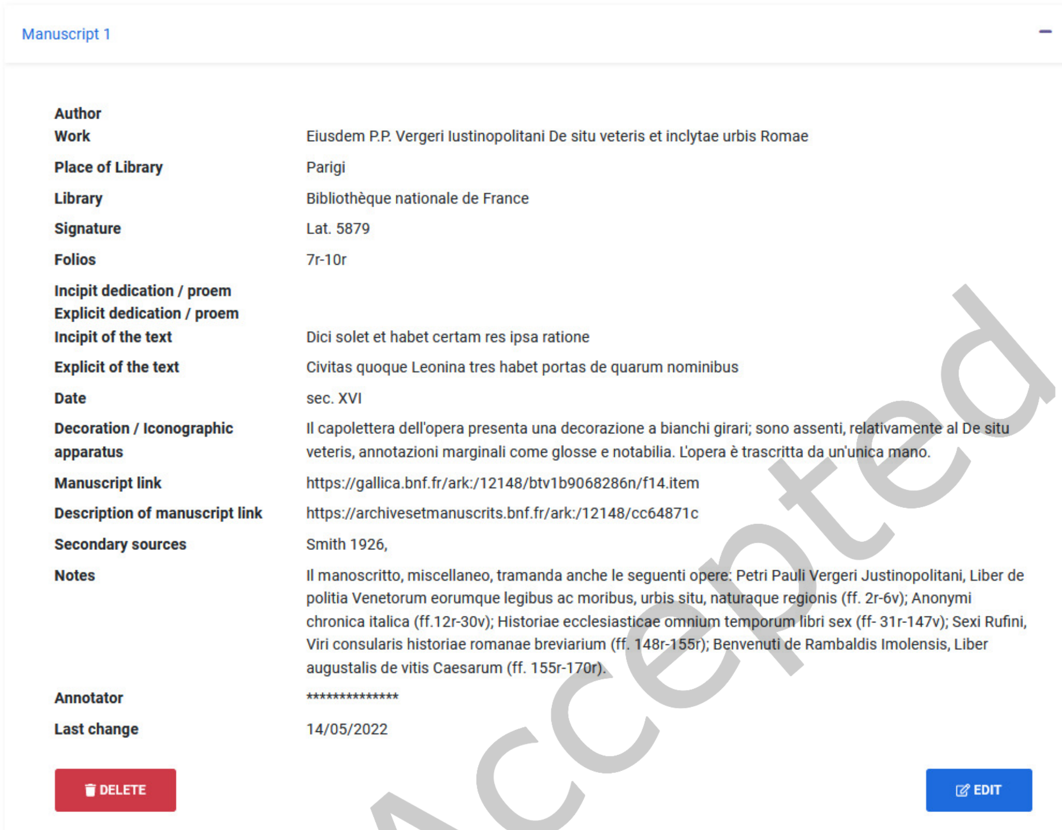


Fig. 4. The start page of the IMAGO Web tool.

literary genre and the toponyms which are cited in the work. Figure 5 shows the interface to add data about a manuscript. A similar interface was developed for the print editions. Each field of these interfaces maps a class of the IMAGO ontology.
Once the data about a work is inserted through the tool, this is encoded as an OWL graph and stored in a triple store. The overall process is described in Figure 6. The data collected by the tool is exported as a JSON object. Indeed our software uses a JSON schema to represent the data [4], structured according to the IMAGO ontology classes. The JSON object is processed by a Java software, which transforms it into an OWL graph encoded in RDF/XML and Turtle formats. This software carries out its task by relying on the Apache Jena library[16]. The graph is finally stored into a Fuseki triple store[17], and it can be queried through a SPARQL endpoint[18].

At the current stage of the project, our KB includes 250 works, 206 authors and 614 libraries, and the scholars have started to insert detailed knowledge about manuscripts and printed editions of these works. The KB also includes ten different literary genres, four types of editions, and six ecdotic typologies. For what concerns the toponyms, at the moment, we only collected those cited by Dante Alighieri in his Latin works, which are 88 corresponding to 80 places. In our case study, we used DanteSearch[19] to automatically retrieve the linguistic data related to the toponyms. DanteSearch is a Web resource that allows querying a complete corpus of Dante Alighieri's works. The corpus is lemmatized and grammatically and syntactically annotated. Using an API[20] we developed to query DanteSearch, for each toponym, we were able to automatically retrieve the textual context

[16]https://jena.apache.org/
[17]https://jena.apache.org/documentation/fuseki2/
[18]https://imagoarchive.it/sparql/
[19]https://dantesearch.dantenetwork.it/
[20]http://vdl.isti.cnr.it/dantesearch.html


Manuscript 1

Author	
Work	Eiusdem P.P. Vergeri Iustinopolitani De situ veteris et inclytae urbis Romae
Place of Library	Parigi
Library	Bibliothèque nationale de France
Signature	Lat. 5879
Folios	7r-10r
Incipit dedication / proem	
Explicit dedication / proem	
Incipit of the text	Dici solet et habet certam res ipsa ratione
Explicit of the text	Civitas quoque Leonina tres habet portas de quarum nominibus
Date	sec. XVI
Decoration / Iconographic apparatus	Il capolettera dell'opera presenta una decorazione a bianchi girari; sono assenti, relativamente al De situ veteris, annotazioni marginali come glosse e notabilia. L'opera è trascritta da un'unica mano.
Manuscript link	https://gallica.bnf.fr/ark:/12148/btv1b9068286n/f14.item
Description of manuscript link	https://archivesetmanuscrits.bnf.fr/ark:/12148/cc64871c
Secondary sources	Smith 1926,
Notes	Il manoscritto, miscellaneo, tramanda anche le seguenti opere: Petri Pauli Vergeri Justinopolitani, Liber de politia Venetorum eorumque legibus ac moribus, urbis situ, naturaque regionis (ff. 2r-6v); Anonymi chronica italica (ff.12r-30v); Historiae ecclesiasticae omnium temporum libri sex (ff- 31r-147v); Sexi Rufini, Viri consularis historiae romanae breviarium (ff. 148r-155r); Benvenuti de Rambaldis Imolensis, Liber augustalis de vitis Caesarum (ff. 155r-170r).
Annotator	*************
Last change	14/05/2022

DELETE EDIT

Fig. 5. Example of manuscript annotation form of the IMAGO Web tool.

in which the toponym occurs and the number of occurrences. Once this data was retrieved, we automatically inserted it in our KB. Table 3 summarises the data about the toponyms of our case study.

Table 3. Data about Dante Alighieri's toponyms collected in the IMAGO KB.

Data	Total Number
Toponyms	88
Places	80
Works	5
Textual contexts	136

The population tool is open source [31], as well as the Java software that creates the OWL graph [30]. The tool and the Java software are based on the ontological model we made freely available online[21]. Hence, researchers can adapt them to other domains of interest or extend them by adding additional pieces of knowledge. Once the knowledge is collected, and the graph is created, the data is published online as LOD. Having the IMAGO

[21] https://imagoarchive.it/documentation/doc/index-en.html

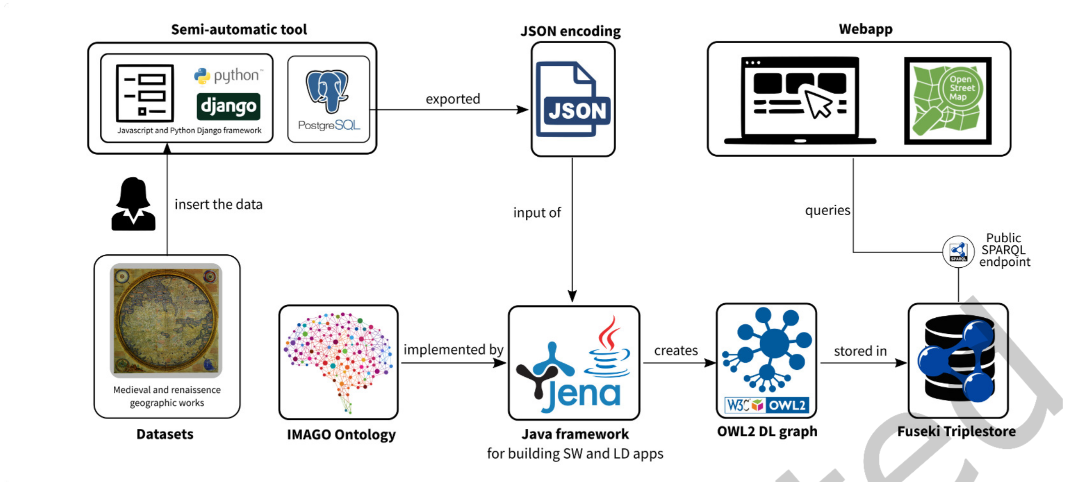


Fig. 6. The architectural schema of the IMAGO project.

dataset available as LOD allows enriching the knowledge collected in the IMAGO KB by linking it to other datasets. At the same time, the LOD dataset can be used and exploited within other research projects. Indeed, the corpus of geographical knowledge collected in the IMAGO project is entirely new in the Digital Humanities domain. Furthermore, until now, the knowledge about medieval geographical works was dispersed on paper books and unavailable in digital format. Finally, we used Pellet, an open source OWL DL reasoner [34], to check the consistency of the IMAGO ontology to make sure to retrieve consistent knowledge as the result of the SPARQL queries. Indeed, by performing semantic queries on the Fuseky SPARQL endpoint, we are able to discover new knowledge starting from the knowledge explicitly represented in the data. For example, we can retrieve all works reported in the same manuscript, the manuscripts collected in the same library, the centuries in which the production of the manuscripts was more relevant, the manuscripts that describe a specific place or how many times Dante Alighieri cited a place.

5 KNOWLEDGE VISUALISATION OF THE DATA RELATED TO THE TOPONYMS

The final aim of the IMAGO project is to create a web application that allows visualising and exploring the data collected in the KB. This knowledge should be accessed by scholars but also by students and general users. For this reason, we implemented some user-friendly visualisations to access the data. The design of the visualisation components was defined after the analysis of the requirements we collected through several interviews with the scholars involved in the project. The first type of visualisation we implemented is the presentation of the toponyms on a map[22]. To implement this map, we used the Leaflet library, an open-source JavaScript library for mobile-friendly interactive maps[23]. This library was already successfully used within the Mapping the Enlightenment project for visualising, exploring, and analysing the intellectual and geographical networks developed by Greek-speaking scholars of the Ottoman Empire during the 17th and 18th centuries [33]. Leaflet allows the creation of geographic maps consisting of several layers. In our case, the first layer is the map provided by OpenStreetMap. The second layer shows the pins that identify the toponyms. In particular, Leaflet takes as input the JSON file produced by our tool. The JSON contains all the information about the toponyms. Each

[22]https://imagoarchive.it/en/toponyms/
[23]https://leafletjs.com/

toponym is reported as a pin on a map associated with a place that is uniquely identified by its geographic coordinates. Figure 7 shows the map of the toponyms reported by Dante Alighieri in his Latin works.

Fig. 7. The toponyms reported by Dante Alighieri in his Latin works.

Clicking on each toponym, a pop-up appears reporting the following data:

- the corresponding linguistic form;
- the work containing the toponym;
- the total number of toponym occurrences in the work;
- the textual place;
- the textual context;
- the link to the corresponding Wikipedia page;
- the link to the corresponding Pleiades page.

Figure 8 shows the pop-up corresponding to *Adria* containing all the data reported above.

A toponym can be cited in more than one work or more than once in the same work. This is the case of the toponym *Florence*, which is cited in different works but also more than once in the same work (e.g. *De Vulgari Eloquentia*, *Monarchia*). The map reported several pins on the Florence place, one for each citation of a toponym in Dante's works (Figure 9-a). Clicking on each pin, a pop-up reporting the data associated with the selected toponym citation is shown (Figure 9-b).

In addition to the map, we designed and implemented an interface that allows users to access the collected data through predefined SPARQL queries. The SPARQL queries are able to retrieve the following data:

(1) Selecting a toponym in a list[24], the user can retrieve:

[24] https://imagoarchive.it/en/toponyms/toponyms.html

Fig. 8. The data associated to Adria reported in a pop-up.

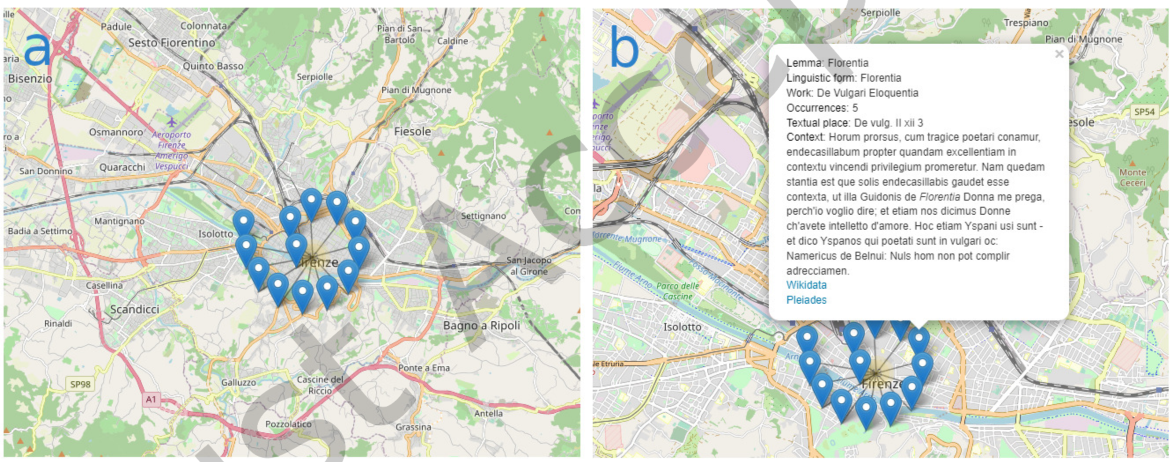


Fig. 9. a. Multiple pins associated to Florence; b. The pop-up reporting the data associated with a citation of the toponym Florence in the De Vulgari Eloquentia.

- the works in which it appears, the textual contexts, the textual places;
- the works in which it appears and the number of occurrences in those works;
- the related place along with the geographic coordinates and the corresponding links to Wikidata and Pleiades pages;
- the authors who cite the toponym and the works in which it appears.

(2) Selecting a work in a list[25], the user can retrieve:

[25]https://imagoarchive.it/en/toponyms/works.html

- the cited toponyms and the number of citations;
- the cited places and the corresponding toponyms, which can be more than one, e.g. the toponyms *Ausonia* and *Hesperia* for Italy.

(3) In the case that the KB includes more than one author, selecting an author in a list, the user can retrieve:
- all the toponyms cited by the author;
- all the works and for each of them the list of the cited toponyms;
- for each work, the total number of the cited toponyms.

The results of all these queries are shown to the user in the form of maps, tables and CSV files.

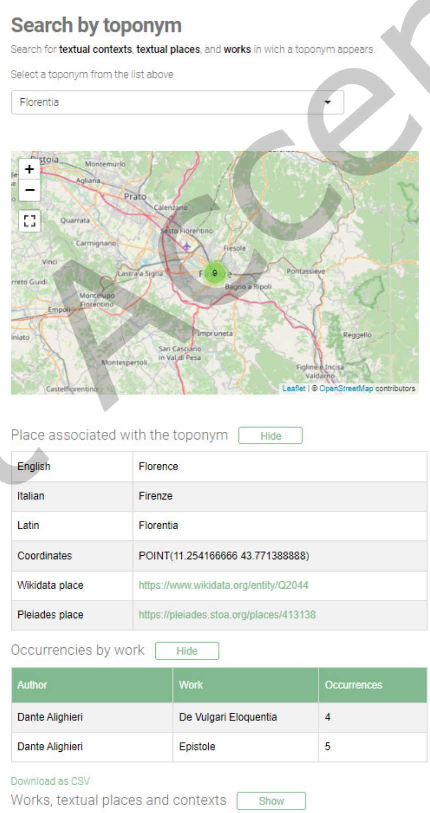


Fig. 10. The "Search by Toponym" interface showing some results associated to the toponym *Florentia*

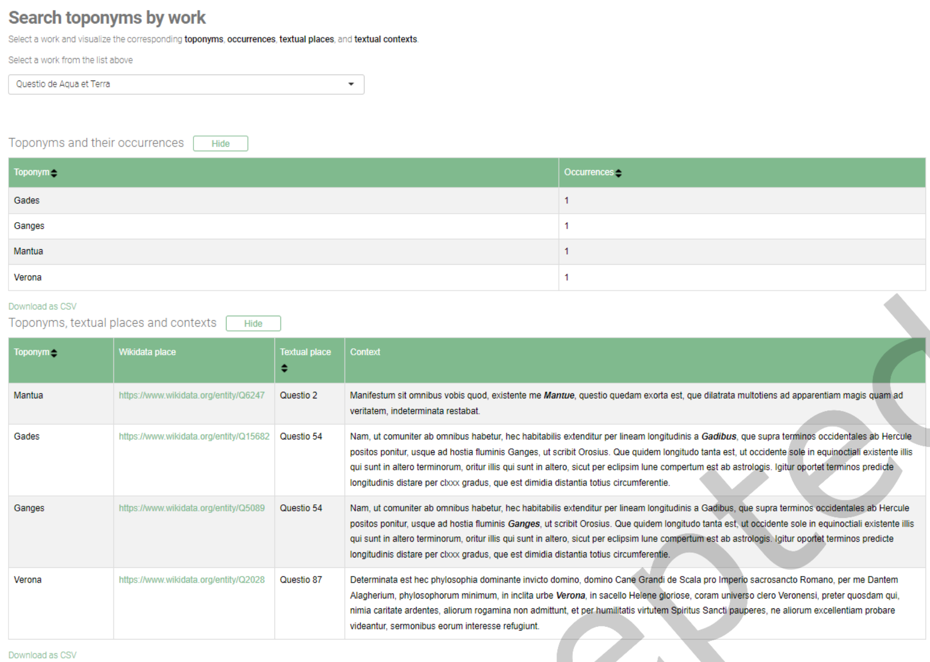


Fig. 11. The "Search toponym by work" interface showing the results associated to the work *Questio de Aqua et Terra*

Figure 10 shows the interface from which a user can select a toponym and visualise the information related to the place associated with the toponym and the occurrences of the toponym in Dante's works. The Figure reports some pieces of knowledge related to the toponym *Florentia*. Figure 11 shows the interface from which, by selecting one of Dante's works, a user can retrieve information about the toponyms that occur in that work, the number of occurrences, the textual places and contexts. In particular, Figure 11 reports the information related to the toponyms that occurr in the *Questio de Aqua et Terra*.

6 CONCLUSION AND FUTURE WORK

In this paper, we have presented how we have used the Semantic Web technologies for representing the knowledge about the toponyms cited in the medieval Latin geographical works collected within the IMAGO project. To reach this aim, we have developed an CIDOC-CRM-based ontology that allows representing the knowledge about this literature. In particular, the ontology allows representing different pieces of knowledge about toponyms, e.g. the place which the toponym refers to, the work in which it is cited, the textual context in which it appears, and the total number of citations in a work. Each resource collected in the IMAGO is represented by an IRI. In order to maximise the interoperability of our KB, we retrieve and reuse IRIs from existing and *standard* KBs, i.e. Wikidata, Pleiades, Mapping Manuscript Migrations. The ontology is expressed in OWL 2 DL and we have designed and developed a Web tool to populate it. On top of the ontology, different visualisations were created, which allow scholars, students and general users to retrieve and access the knowledge about the toponyms. In particular, we have implemented a visualisation of the toponyms on a map, where clicking on the pin that identified the toponym, a pop-up appears that shows all the information about the toponym we collected. Another visualisation is in the form of tables and CSV files that aggregate data about a toponym (e.g. works in which it appears, number of occurrences, the corresponding place). As case study, we considered the toponyms cited by the Italian poet

Dante Alighieri in his Latin works. However, one of the final aims of IMAGO project is to create a complete index of the toponyms cited in the medieval and renaissance Latin geographical works. The main advantages of using a Semantic Web approach instead of a traditional database to create such an index are the following: (i) the ontology allows to easily add classes and properties, thereby refining the ontology, (ii) the ontology allows representing a huge variety of semantic relationships easily, (iii) the ontology can be linked to other ontologies to extend the represented domain, (vi) any user can download and use the knowledge base freely, using the paradigm of Linked Open Data. Eventually, the digitisation of the knowledge about medieval toponyms collected in the IMAGO KB and their visualisation in the form of maps, tables and CSV files allow the scholars to have a complete overview of this data. For example, it is possible to easily carry out a first analysis of the spatial distribution of the toponyms of each author, visualising them on a map. Furthermore, scholars can study the evolution of geographical knowledge over time, automatically analysing which and how many toponyms were cited by different authors who lived in different times or by the same author in his/her works written in different years. For example, the visualisations of the results of the case study reported in this paper, are able to provide an overview of Dante Alighieri's geographical knowledge and allow understanding of how this knowledge evolved in his works. We think that it is a precious information for scholars, but also for students and generally for users interested in Dante's literature. To facilitate the perception of the evolution of geographical knowledge over time, as future work, we have also planned to show the temporal knowledge related to the ancient manuscripts and digital editions on a timeline. In [2] we have reported the results of a first evaluation of the ontological model we performed. As future work, we will carry out a complete assessment of the ontology, which include not only the model but also the KB of the project. In particular, we plan to conduct two different types of assessment: an automatic assessment and an assessment involving users. For the first type of assessment, we plan to use the automatic OntoQA system [36] that evaluates both the model and the KB. For the second type of assessment, we plan to propose a specific questionnaire to the scholars who are currently populating the ontology. After the analysis of the assessment results, if necessary, we will review and extend our ontology.

For what concerns the toponyms, we have planned to expand our KB with those cited in all works we are collecting. Furthermore, we are working to implement a data visualisation in the form of network graphs that will allow users to identify the toponyms shared by different works. The long term aim of the project is to develop a Web application that allows retrieving and consulting all the data collected in the IMAGO KB in a user-friendly way.

REFERENCES

[1] MG Aune. 2011. The Itineraries of William Wey. *Renaissance Quarterly* 64, 1 (2011), 310.

[2] Valentina Bartalesi, Daniele Metilli, Nicolò Pratelli, and Paolo Pontari. 2022. Towards a knowledge base of medieval and renaissance geographical Latin works: The IMAGO ontology. *Digital Scholarship in the Humanities* 37, 1 (2022), 34–50. https://doi.org/10.1093/llc/fqab060

[3] Valentina Bartalesi and Nicolò Pratelli. 2021. The IMAGO Project: Towards a Knowledge Base of Medieval and Renaissance Geographical Works (extended abstract). In *Proceedings of the International Joint Workshop on Semantic Web and Ontology Design for Cultural Heritage co-located with the Bolzano Summer of Knowledge 2021 (BOSK 2021), Virtual Event / ozen-Bolzano, Italy, September 20-21, 2021 (CEUR Workshop Proceedings, Vol. 2949)*, Antonis Bikakis, Roberta Ferrario, Stéphane Jean, Béatrice Markhoff, Alessandro Mosca, and Marianna Nicolosi Asmundo (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-2949/short1.pdf

[4] Valentina Bartalesi, Nicolò Pratelli, and Emanuele Lenzi. 2022. *JSON schema of the internal IMAGO tool data representation.* Retrieved December 19, 2022 from https://imagoarchive.it/imago_json.html

[5] Sebastian Barzaghi, Monica Palmirani, and Silvio Peroni. 2020. Development of an ontology for modelling medieval manuscripts: The case of Progetto IRNERIO. *Umanistica Digitale* 9 (2020), 117–140. https://doi.org/10.6092/issn.2532-8816/11187

[6] Florian Bauer and Martin Kaltenböck. 2011. Linked open data: The essentials. *Edition mono/monochrom, Vienna* 710 (2011).

[7] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2011. Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*. IGI global, 205–227.

[8] Nathalie Bouloux. 1999. *Culture et savoirs géographiques dans l'Italie du XIVe siècle.* Ph. D. Dissertation. Paris 1.

[9] Toby Burrows, Douglas Emery, Mitch Fraas, Eero Hyvönen, Esko Ikkala, Mikko Koho, David Lewis, Andrew Morrison, Kevin Page, Lynn Ransom, Emma Thomson, Jouni Tuominen, Athanasios Velios, and Hanno Wijsman. 2020. Mapping Manuscript Migrations: Digging into Data for Researching the History and Provenance of Medieval and Renaissance Manuscripts: White Paper.

[10] Daniel K Connolly. 1999. Imagined pilgrimage in the itinerary maps of Matthew Paris. *The Art Bulletin* 81, 4 (1999), 598–622. https://doi.org/10.2307/3051336

[11] Domenico Defilippis. 2009. Da Flavio Biondo a Leandro Alberti. Corografia e antiquaria tra Quattro e Cinquecento, Atti del Convegno di Studi (Foggia, 2 febbraio 2006), a cura di D. Defilippis. (2009).

[12] Linfang Ding, Guohui Xiao, Albulen Pano, Claus Stadler, and Diego Calvanese. 2021. Towards the next generation of the LinkedGeoData project using virtual knowledge graphs. *Journal of Web Semantics* 71 (2021), 100662. https://doi.org/10.1016/j.websem.2021.100662

[13] Martin Doerr. 2003. The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine* 24, 3 (2003), 75–75. https://doi.org/10.1609/aimag.v24i3.1720

[14] Martin Doerr, Chryssoula Bekiari, Patrick LeBoeuf, and Bibliothèque nationale de France. 2008. FRBRoo, a conceptual model for performing arts. In *Proceedings of the 2008 Annual Conference of CIDOC, Athens*. 15–18.

[15] Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing Wikidata to the linked data web. In *Proceedings of the International semantic web conference*. Springer, 50–65. https://doi.org/10.1007/978-3-319-11964-9_4

[16] Stefanie Gehrke, Eduard Frunzeanu, Pauline Charbonnier, and Marie Muffat. 2015. Biblissima Prototype on Medieval Manuscript Illuminations and their Context.. In *Proceedings of the International Workshop Semantic Web for Scientific Heritage at ESWC Conference*. 43–48.

[17] Sergio Gensini. 2000. *Viaggiare nel Medioevo*. Pacini Centro di Studi sulla civiltà del tardo Medioevo, Pisa San Miniato.

[18] Pascal Hitzler, Markus Krotzsch, and Sebastian Rudolph. 2009. *Foundations of semantic web technologies*. Chapman and Hall/CRC.

[19] Eero Hyvönen, Esko Ikkala, Mikko Koho, Jouni Tuominen, Toby Burrows, Lynn Ransom, and Hanno Wijsman. 2021. Mapping Manuscript Migrations on the Semantic Web: A Semantic Portal and Linked Open Data Service for Premodern Manuscript Research. In *Proceedings of the International Semantic Web Conference*. Springer, 615–630. https://doi.org/10.1007/978-3-030-88361-4_36

[20] Anna Jordanous, K Faith Lawrence, Mark Hedges, and Charlotte Tupman. 2012. Exploring manuscripts: Sharing ancient wisdoms across the semantic web. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*. 1–12. https://doi.org/10.1145/2254129.2254184

[21] Tobias Kohr, Kai-Christian Bruhn, Ashish Karmacharya, Christophe Cruz, and Frank Boochs. 2013. An SDI for archaeological data with a RESTful interface to semantically modeled information. In *Geoinformatiks*.

[22] Paul Oskar Kristeller. 1963. *Iter Italicum: Alia itinera 1. Australia to Germany*. Vol. 3. Warburg Institute.

[23] Markus Krötzsch. 2012. OWL 2 profiles: An introduction to lightweight ontology languages. In *Reasoning Web International Summer School*. Springer, 112–183. https://doi.org/10.1007/978-3-642-33158-9_4

[24] Deborah L McGuinness, Frank Van Harmelen, et al. 2004. OWL web ontology language overview. *W3C recommendation* 10, 10 (2004), 2004.

[25] Enrico Menestò. 1993. Relazioni di viaggi e di ambasciatori. *Cavallo, Guglielmo; Leonardi, Claudio; Menestò, Enrico (dir. da), Lo spazio letterario del Medioevo* 1 (1993), 535–600.

[26] Friar Odoric. 2002. The Travels of Friar Odoric. *Edited by Henry Yule and Paolo Chiesa. Translated* (2002).

[27] Jeff Z Pan. 2009. Resource description framework. In *Handbook on ontologies*. Springer, 71–90. https://doi.org/10.1007/978-3-540-92673-3_3

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

[29] August Potthast. 1970. Repertorium fontium historiae medii aevi. (1970).

[30] Nicolò Pratelli and Valentina Bartalesi. 2022. *The Java sofware to create the IMAGO OWL graph*. Retrieved December 19, 2022 from https://github.com/prate91/imagoTriplifier

[31] Nicolò Pratelli, Valentina Bartalesi, and Emanuele Lenzi. 2022. *IMAGO Annotation Tool*. Retrieved December 19, 2022 from https://github.com/prate91/IMAGO_annotation_tool

[32] Pat Riva and Maja Žumer. 2017. FRBRoo, the IFLA Library Reference Model, and now LRMoo: a circle of development. (2017).

[33] Vasileios Routsis, Irene Goudarouli, and Manolis Patiniotis. 2017. Mapping the Enlightenment: Intellectual Networks and the Making of Knowledge in the European Periphery. In *DH*.

[34] Evren Sirin, Bijan Parsia, Bernardo Cuenca Grau, Aditya Kalyanpur, and Yarden Katz. 2007. Pellet: A practical owl-dl reasoner. *Journal of Web Semantics* 5, 2 (2007), 51–53. https://doi.org/10.1016/j.websem.2007.03.004

[35] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. 2012. Linkedgeodata: A core for a web of spatial open data. *Semantic Web* 3, 4 (2012), 333–354. https://doi.org/10.3233/SW-2011-0052

[36] Samir Tartir, I Budak Arpinar, Michael Moore, Amit P Sheth, and Boanerges Aleman-Meza. 2005. OntoQA: Metric-based ontology quality analysis. (2005).

[37] Sabine Tittel. 2022. Towards an Ontology for Toponyms in Nepalese Historical Documents. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation*

Conference. 7–16.

[38] Denny Vrandecic. 2013. The rise of Wikidata. *IEEE Intelligent Systems* 28, 4 (2013), 90–95. https://doi.org/10.1109/MIS.2013.119

[39] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 1 (2016), 1–9. https://doi.org/10.1038/sdata.2016.18