



Ontology-based data integration in EPNet: Production and distribution of food during the Roman Empire

Diego Calvanese^a, Pietro Liuzzo^d, Alessandro Mosca^b, José Remesal^c, Martin Rezk^a, Guillem Rull^c

^a KRDB Research Centre, Free University of Bozen-Bolzano, Italy

^b SIRIS Lab, Research Division of SIRIS Academic, Spain

^c CEIPAC, University of Barcelona, Spain

^d ZAW, University of Heidelberg, Germany

ARTICLE INFO

Keywords:

E-Culture
Ontology-Based Data Access
Ontology-Based Data Integration
EPNet
Ontop
Knowledge Representation and Reasoning

ABSTRACT

Semantic technologies are rapidly changing the historical research. Over the last decades, an immense amount of new quantifiable data have been accumulated, and made available in interchangeable formats, in social sciences and humanities, opening up new possibilities for solving old questions and posing new ones. This paper introduces a framework that eases the access of scholars to historical and cultural data about food production and commercial trade system during the Roman Empire, distributed across different data sources. The proposed approach relies on the Ontology-Based Data Access (OBDA) paradigm, where the different datasets are virtually integrated by a conceptual layer (an ontology) that provides to the user a clear point of access and a unified and unambiguous conceptual view.

© 2016 Published by Elsevier Ltd.

1. Introduction

Semantic technologies (Hitzler et al., 2009; Shadbolt et al., 2006; Domingue et al., 2011) are rapidly changing historical research and, more in general, research in humanities. Over the last decades, an immense amount of new quantifiable data have been accumulated, and made available in interchangeable formats, from social sciences to economics, opening up new possibilities for solving old questions and posing new ones (Raghavan, 2014). Historians, especially in Digital Humanities, are starting to use new datasets to aggregate information about history: collections of data, information, and knowledge that are devoted to the preservation of the legacy of tangible and intangible culture inherited from previous generations. Moreover, the recent advances in computing and computational tools make it feasible to meaningfully manipulate, manage, and analyse these datasets.

Since a sustainable maturity in the development of Semantic Web and Linked Open Data technologies has been reached – think of data exchange protocols, standardised knowledge representation languages, common data formats, etc.¹ – a considerable number of public initiatives and projects have been funded to address the issue of building, and making public through the web, historical and cultural data. Among others, the following projects are worth to be mentioned here, since they represent pioneering efforts in the application of semantic technologies toward the

development of e-Culture portals providing multimedia access to distributed collections of cultural heritage objects, of tools for publishing Linked Data, of ontologies, and of lexical resources in the area of ‘Semantic Web and History’: EUROPEANA², CULTURESAMPO³, STICHATCH⁴, Multimediana N9C⁵, CHIP⁶, EAGLE⁷, CIDOC Conceptual Reference Model (CRM)⁸, GETTY Vocabularies⁹, INCONCLASS¹⁰, TEI/EpiDoc¹¹. A more comprehensive list of available joint initiatives of historians and computer scientists in the use of Semantic Web methods and technologies in historical research can be found in Merono-Penuela et al. (2015).

Some of these projects are explicitly meant to expose data structures, integrated datasets, vocabularies, and ontologies to support further initiatives in the design and development of computer applications in the Digital Heritage area, some others simply represent implementations of the envisioned applications. Notable examples of the first kind are EAGLE, the Europeana network of Ancient Greek and Latin Epigraphy, CIDOC CRM, and

² <http://www.europeana.eu>.

³ <http://www.kulttuurisampo.fi>.

⁴ <http://www.cs.vu.nl/STITCH>.

⁵ <http://e-culture.multimediana.nl>.

⁶ <http://chip.win.tue.nl>.

⁷ <http://http://www.eagle-network.eu>.

⁸ <http://www.cidoc-crm.org>.

⁹ <http://www.getty.edu/research/tools/vocabularies>.

¹⁰ <http://www.iconclass.org>.

¹¹ <http://epidoc.sourceforge.net>.

¹ W3C Standards – <http://www.w3.org>.

TEI/EpiDoc. These initiatives became the reference points for what concerns the digital representation of cultural heritage information, and the XML and OWL-based data structures they produced are widely applied today. Most examples of the second kind, that is, implementations of e-Culture portals, resulted mostly in web-based virtual museum applications, collecting heterogeneous contents (e.g., digital representations of paintings, music, movies, books) from various distributed sources (Merono-Penuela et al., 2015). In these portals, the users experience a digital environment where, for instance, they can look for a painting of a historic event, find information on the event along with other artwork depicting it, geo-localise it, and see where nearby events occurred and how they are represented in artwork Schreiber et al. (2006).

Notice that a vocabulary is not an ontology: both vocabulary and ontology account for the way human beings refer to things in the real world, however they have differences. A controlled vocabulary is a list of terms that have been enumerated explicitly, and it may have no formally specified meaning. No defined interrelationships between terms is specified in a vocabulary, and in the absence of a formal semantics no automatic reasoning technique can be exploited. On the other hand, ontologies applied to information sources may be seen as explicit formal conceptualisations that describe the semantics of the data. Ideally, an ontology should contain the vocabulary as a building block. Ontologies are usually specified in Description Logics (DLs) (Baader et al., 2007), a family of knowledge representation languages that provide one of the main underpinnings for the OWL Web Ontology Language¹² as standardised by the World Wide Web Consortium (W3C). DLs are equipped with a formal semantics based on First-Order Logic. This formal semantics allows humans and computer systems to exchange DL ontologies without ambiguity as to their meaning, and also makes it possible to use logical deduction to infer additional information from the facts stated explicitly in an ontology (Krötzsch et al., 2012).

A shortcoming of the existing models is that they cannot be directly understood by non-experts since (i) the concept names are often *not* self-explanatory (for instance, the concept name for 'Information Carrier' is 'E84' in CIDOC CRM); and (ii) the concepts are intentionally defined at a very abstract level in order to be useful for any domain in the digital humanities field (for instance, E75: 'Conceptual Object Appellation'). Moreover, the existing portals rely on extract-transform-load (ETL) processes to integrate the data into a single database (e.g., a triple store). The resulting technological infrastructures is a materialised dataset, where the content is actually provided by different subjects but, at the end, managed as a single, unified, dataset. This implies that the data is often outdated, and periodic costly processes are required to bring it up to date. Observe that none of these approaches exploits the existing hierarchies and constraints in the ontology, neither to ease the access of the end users to the data, nor to detect inconsistencies that might be present in the data. For instance, in the EAGLE Portal (providing faceted browsing relying on a set of controlled vocabularies) the semantic information provided by the ontology (such as hierarchies, domain and range properties, etc.) is neither available nor exploitable by the user. Indeed, as pointed out in Merono-Penuela et al. (2015), reasoning is one of the key mechanisms of the Semantic Web still to be used in historical research.

A recently introduced paradigm that combines the possibility of using reasoning with respect to domain knowledge encoded in an ontology, with a mechanism to use the same ontology also for high level, integrated access to data sources is that of *Ontology-Based Data Access and Integration* (OBDA/OBDI) (Calvanese et al.,

2009a, 2009b). In OBDI¹³, a domain ontology is connected to the data sources through a declarative specification given in terms of *mappings* (Poggi et al., 2008; Das et al., 2012) that relate symbols in the ontology (classes and properties) to views over the data expressed by means of SQL queries. The ontology and mappings together expose the data in the sources in the form of an RDF graph, which however is not materialized. Queries, which can be formulated over the classes and properties of the ontology, are interpreted over this *virtual* RDF graph, and are translated, making use of the mappings, into SQL queries over the data sources. In this setting, users simply query the ontology, and no longer need an understanding of the data sources, the relation between them, or the encoding of the data. As query language, we rely on SPARQL¹⁴, the standard query language for the Semantic Web. Instead, for expressing ontologies in the OBDA and OBDI setting, the most commonly used language is OWL 2 QL¹⁵, which is the profile (i.e., sub-language, in W3C terminology) of OWL 2 that is specifically tailored for efficiently querying large amounts of data. In our work we rely on OBDA and OBDI systems that support such language, together with the standard mapping language R2RML¹⁶, and that provide the standard functionalities of a SPARQL endpoint to query the underlying data sources.

Specifically, we provide the following contributions, thus overcoming the shortcomings of existing approaches for the integration of historical data:

- We define a conceptual reference model that integrates and extends already existing ontologies and standards for the representation of historical data (e.g., CIDOC CRM, the EAGLE metadata model¹⁷, and FaBio¹⁸).
- Using a historical domain vocabulary (linked to the existing abstract ontologies), we specifically build an OWL 2 QL ontology that can be easily understood by historians, to ease their access to heterogeneous datasets. The ontology has been semi-automatically derived from the EPNet conceptual model introduced in Section 3.
- We define a set of mappings linking the different concepts in the ontology to the data stored in three well-known datasets: the EPNet relational repository (see Section 4.1), the Epigraphic Database Heidelberg (EDH) (see Section 4.2), and the Pleiades dataset (see Section 4.3).
- We develop a web-based framework that uses the ontology to virtually integrate these three datasets (and that is easily extendable to more) without performing expensive ETL processes, and avoiding the need of periodically updating the data. Notably, in the proposed framework, reasoning can be exploited to ease the access of scholars to the data.
- We provide a SPARQL endpoint¹⁹ to allow other systems to query the integrated data.

The example below intuitively shows the rationale behind choosing this approach for integration:

Example 1. Suppose the user looks for all the objects (information carriers) produced in 'La Corregidora' and is interested in the *complete* information that is available in *all* three datasets. The EPNet dataset and EDH contain information about different types

¹³ Ontology-Based Data Access refers to the case of a single data source accessed through an ontology, whereas in the presence of multiple data sources, as in this work, we talk about *Ontology-Based Data Integration*.

¹⁴ <http://www.w3.org/TR/rdf-sparql-query/>.

¹⁵ <http://www.w3.org/TR/owl-profiles/>.

¹⁶ <http://www.w3.org/TR/r2rml/>.

¹⁷ <http://www.eagle-network.eu/about/documents-deliverables/>.

¹⁸ <http://vocab.ox.ac.uk/fabio>.

¹⁹ <http://www.w3.org/TR/rdf-sparql-protocol/>.

¹² <http://www.w3.org/TR/owl-overview/>

of information carriers (amphoras, monuments, stones) and some (potentially incomplete) information about geo-coordinates. On the other hand, the Pleiades dataset contains complete geo-coordinates but has no information about amphoras. Traditionally, to query this information requires (i) a deep understanding of the datasets to create the necessary queries that extract all information about each of these types of objects from each dataset; and (ii) to merge the answers and filter out the objects that were not

produced in 'La Corregidora'. Thus obtaining an answer for this simple information need is not only extremely complex, but requires the user to know all database schemas, the encoding used in each data source, and to manually merge the information obtained from each of them. Ideally the user should instead be able to execute a single simple query that does not require any specific knowledge about the underlying data sources, and get all the available information coming from all three datasets.

This work is contextualised within the EPNet Project (ERC Advanced Grant EPNet "Production and distribution of food during the Roman Empire: Economics and Political Dynamics", ERC-2013-ADG 340828). The emphasis of the EPNet Project is on providing historians with computational tools to compare, aggregate, measure, geo-localise, and search data about Latin and Greek inscriptions on amphoras for food transportation, and this is done by relying on the OBDI paradigm.

The paper is organised as follows: Section 2 provides an overview of the historical context of the EPNet project, together with its main objectives related to the development of innovative computational means for historical research. Section 3 briefly describes the different artefact that we developed to build our OBDA-based data management system. Section 4 introduces the main characteristics of the datasets whose semantic-based integration is presented here. Section 5 is devoted to the introduction, by means of examples, of the OBDA framework we implemented, explaining how this solution deals with data access, integration, and consistency issues. A preliminary testing-oriented interface, together with a SPARQL endpoint, are hyperlinked in the same section. Section 6 concludes the paper.

Preliminary results concerning the knowledge representation and data modelling effort in the EPNet project have been submitted to the 'First International Workshop on Semantic Web for Cultural Heritage (SW4CH 2015)' and to the 'Digital Heritage International Congress 2015', and they are still under revision. These submissions briefly report about a much smaller ontology, and the integration of only two datasets (as opposed to the three presented here). Moreover, in the previous two submissions, no



Fig. 1. Stamp engraved on a Dressel 20 amphora that belonged to Septimius Severus and his sons.



Fig. 2. Titulus pictus in 'delta' position over a Dressel 20 amphora, transcribed as: "[R A]stigis arca p(endo) ccxl / [act]us agatephori p(ensit) atimetion/[d(omino)] n(ostro) antonino iii et comazonte co(n)s(ulibus)" [222 A.D.].



Fig. 3. The result of a query over the stamp ACIRGI in the EPNet dataset.

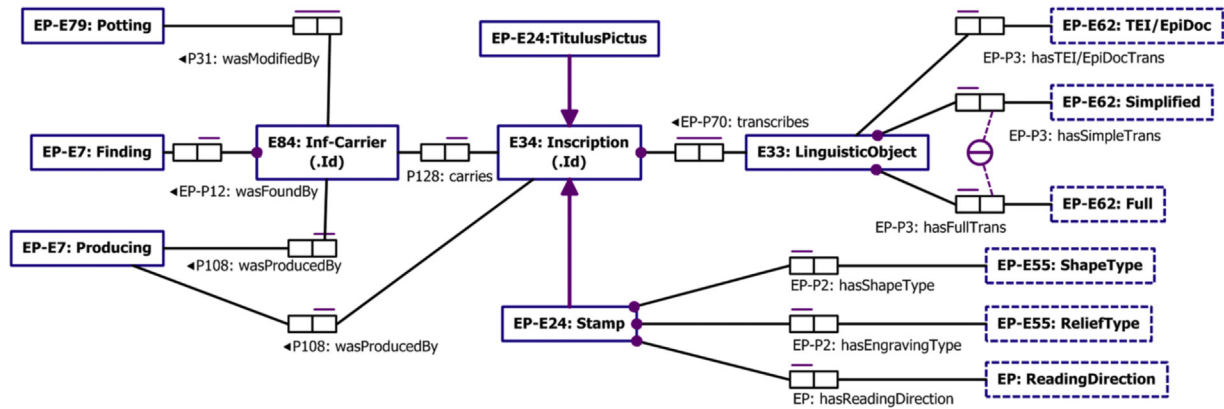


Fig. 4. A fragment of the EPNet CRM, where *InformationCarrier* is related with activities like *Producing*, *Potting*, and *Finding*. *Stamp* and *TitulusPictus* are inscriptions. In particular, instances of *Stamp* are characterised, among others, by their *Relief type*, *Shape type*, and *ReadingDirection*. The model also shows that inscriptions are connected with their 'simplified', 'full', and *EpiDoc* transcriptions, bringing information about their translation into contemporary alphabets, and conservation status. The model is written in ORM2: continuous line rectangles represent concepts, while dashed rectangles stand for datatypes (*xsd:String*, *xsd:Date*, *xsd:Integer*, etc.); each relation has an explicit number of arguments typed by the linked concepts, and is annotated with the preferred reading (e.g., 'P31:wasModifiedBy'); pink coloured symbols indicate cardinality constraints that have been superimposed to the schema (dots stand for 'at least one' mandatory constraints, and horizontal lines for 'at most one' cardinality constraints), while an arrow stands for the usual *is-a* relation.

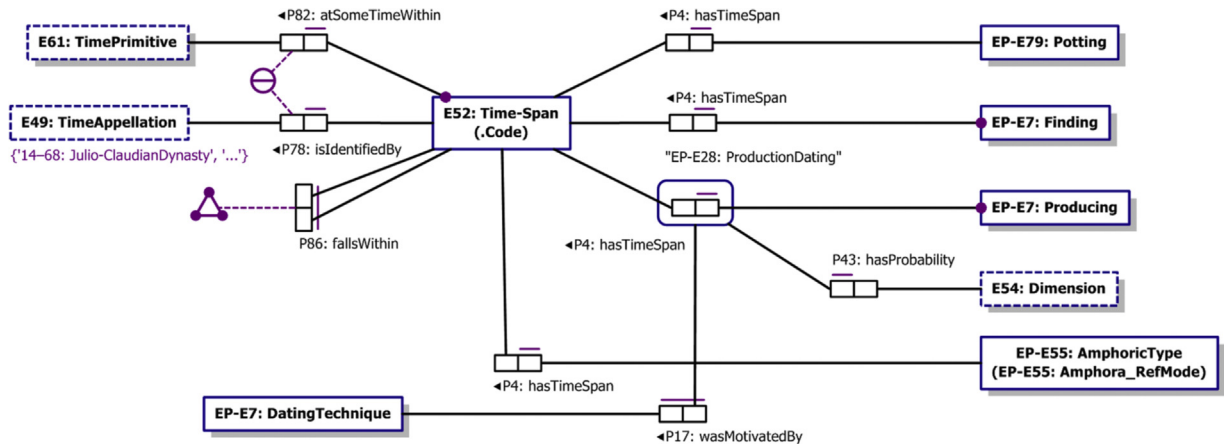


Fig. 5. The TIME module of the EPNet CRM, where the activities related with the discovery, the production, and the potting of the information carriers, and inscriptions, are connected with their associated *Time-Span*. Instances of *Time-Span* are then related to each other by means of the relation *fallsWithin* directly provided by the CIDOC CRM data model.

SPARQL endpoint is provided, and no evaluation of query times are shown.

2. The historical context

The Roman Empire trade system is generally considered to be the first complex European trade network. It formed an integrated system of interactions and inter-dependencies between the Mediterranean basin and northern Europe. Over the last couple of centuries, scholars have developed a variety of theories to explain the organisation of the Roman Empire trade system, but most of them continue to be speculative and difficult to falsify (Garnsey and Whittaker, 1983; Lo Cascio and Rathbone, 2000).

EPNet aims at setting up an innovative framework to investigate the mechanisms and characteristics of the commercial trade system during the Roman Empire. The main objective of EPNet is to create an interdisciplinary experimental laboratory (the project team includes specialists from Social Sciences and Humanities and from Physical, and Computer Sciences) for the exploration, validation and falsification of existing theories, and for the formulation of new ones. This approach is made possible on the one hand by a large dataset of existing empirical data about Roman

amphorae and their associated epigraphy (see, e.g., Figs. 1–3) that has been created during the last two decades, and on the other hand by the front line theoretical research done by historians on the political and economic aspects of the Roman trade system.

2.1. The economy of the Roman Empire: an ongoing debate

A crucial aspect of any society is the production, supply, and redistribution of food. This topic has long been, and still remains, one of the open problems for sustainable decision policies in a world scale perspective. The food distribution during the Roman Empire is commonly associated with the control of the army. It is argued that the emperor and his circle managed the relationship between food and army in order to supervise and control the whole Roman territory and to strengthen and maintain their own political power. Two approaches are particularly evident in the current debate over scales and modalities of the Roman economics system: (i) the Roman Empire trade system as a specific model not connected with modern global economies; (ii) the Roman Empire trade system as a sort of predecessor of modern global economies perfectly explainable through modern economic theories. Assuming or not an analogy between past and present, the scientific debate has focused mostly on the influence of the capital of

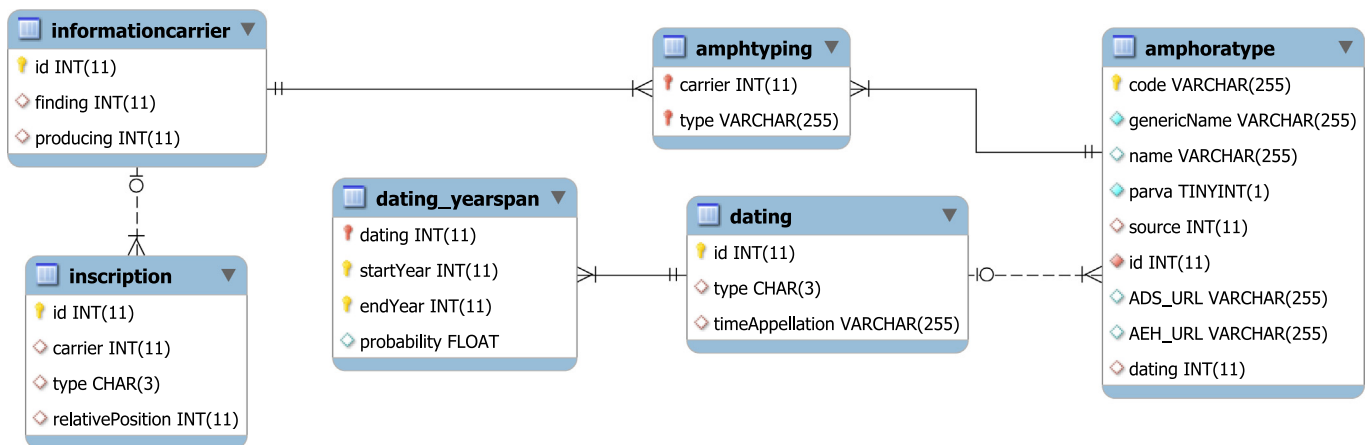


Fig. 6. A fragment of the relational specification of the EPNet database.

the Empire (Rome) on the control and management of long distance trade, rather than on the analysis of the role played by periphery and regional distribution.

Roman archaeology provides us with an incredible source of data and information about economic productions and transactions around modern Europe and the Mediterranean basin (see, e.g., Fig. 2). However, a scientific study of the mechanisms that have characterised these economic and political links is still missing. The main reason is the lack of formal approaches and methods in historical research. Specialists in history often do not even consider the possibility that their research can be scientifically supported and expressed using formal languages (codified using non-ambiguous languages capable of generating models that can be executed, by analytical or computational methods). However, ancient societies provide a great opportunity to evaluate diachronic real-world data with a virtual laboratory in which formal models can be built and different hypothesis and theories about the past explored (see Epstein, 2008).

2.2. Computer Science in EPNet

The computational infrastructure envisioned in the EPNet project takes the form of a 'Virtual Research Environment' offering: (i) a conceptual layer (in terms of an ontology) driving the access to datasets stored into fragmented, heterogeneous, and distributed digital repositories; (ii) a platform for sharing expert knowledge on characterisation, typology, and dating of Roman Empire epigraphies/artefacts; and (iii) dedicated data visualisation and data analytics tools. Taking into consideration the design and development of such a computational infrastructure, the conceptual modelling and knowledge representation effort in EPNet has been planned to address three main problems: (i) structuring and accessing large collections of data through the Web, (ii) providing a formally defined, unambiguous, framework for analysing the data and exporting them in a way that can be further manipulated by computer-based simulation and complex network analysis tools, and (iii) making each collection of data integrable with other complementary data sources. *Here is where interdisciplinarity and epistemological transparency require advanced semantic-aware data management methodologies and technologies, scholars need access to the data in a distributed worldwide-scaled environment, and semantic-based data access and integration becomes a 'sine qua non' requisite.*

Given the above requirements, semantic-based data management can account for discrete data in addition to qualitative interpretations, so as to answer broader questions about patterns in the historical record. In particular, it enables scholars to retrieve information in a domain-centred and scholar-friendly way, thus

supporting the identification of patterns and trends in this information and discover relationships between disparate pieces of it. Semantic technologies support EPNet in facing the main challenge of providing users with: (1) a running technology for accessing data in a way that is conceptually sound with their own domain knowledge; (2) a semantically transparent platform, ready to acquire and be complemented with new data from different sources; and (3) a theoretically grounded mechanism to homogenise information stored in different formats and according to different conceptualisations.

In particular, by adopting the OBDI approach and its supporting technology, an extensive amount of the EPNet data is to be connected and subsequently interpreted at a variety of levels, which will give new insights to the complexity of the Roman Empire exchange relations. A relevant portion of the research in humanities is still characterized by the absence of any specific technical knowledge about the usage of formal query languages, and of data federation and data virtualisation tools. In such a context, easing the access to existing relational databases by means of OBDI, is essential (1) for the generation of new knowledge, and (2) for the specification of values and parameters that will be manipulated in simulation experiments (which are among the planned research activities of EPNet).

3. Knowledge representation in EPNet

In this section, we present the EPNet Conceptual Reference Model (CRM), and the derived EPNet OWL 2 QL Ontology. Observe that, to allow for virtual integration based on rewriting SPARQL queries over the ontology to (efficiently executable) SQL queries over the data sources, it is necessary to use a lightweight ontology language such as OWL 2 QL.²⁰ We specifically built the ontology exposed to the user by relying on a historical domain vocabulary (linked to the existing abstract vocabularies, such as CIDOC CRM) that can be easily understood by historians, so as to ease the access to heterogeneous datasets. We split the creation of the conceptual layer into two steps: first we created a conceptual reference model that represents a complete model of the domain knowledge, but which is too expressive to be used as an ontology for query-answering purposes; such conceptual model is represented using the expressive conceptual modelling language ORM2 (c.f., Section 3.1). In a second step, we took the fragment of this ORM2 conceptual model that is relevant for query answering, and translated it to OWL 2 QL, at the same time extending it with concepts that are relevant for the user but not necessarily for

²⁰ Beyond this ontology language, rewritability of SPARQL queries into SQL queries is lost (Calvanese et al., 2013).

```

<record>
  <tmUri>http://www.trismegistos.org/text/154496</tmUri>
  <provinz>Britannia</provinz>
  <land>United Kingdom</land>
  <fo_antik_pleiades>"http://pleiades.stoa.org/places/17649950">London Mithraeum</fo_antik>
  <fo_modern_geonames>"http://www.geonames.org/2643741/city-of-london.html">London</fo_modern>
  <fundjahr>1889</fundjahr>
  <fundstelle>London</fundstelle>
  <aufbewahrung>London: Museum of London</aufbewahrung>
  <dekor/>
  <i_gattung>Building/dedicatory inscription</i_gattung>
  <denkmaltyp>Relief</denkmaltyp>
  <material>Marble, white</material>
  <hoehe>43.2</hoehe>
  <breite>55.9</breite>
  <metrik/>
  <sprache>Latin</sprache>
  <datierung_von>0175</datierung_von>
  <datierung_bis>0225</datierung_bis>
  <geographie>J</geographie>
  <beleg>provisional</beleg>
  <bearbeiter>Scott Vanderbilt</bearbeiter>
  <datum>2011-06-11</datum>
  <lit>
    <lit_line>RIB 0003.</lit_line>
  </lit>
  <kommentar>
    <komm_line/>
  </kommentar>
  <personen>
    <person n="1">
      <name>Ulpus Silvanus</name>
      <praenomen/>
      <nomen>Ulpus</nomen>
      <cognomen>Silvanus</cognomen>
      <supernomen/>
      <tribus/>
      <origo/>
      <geschlecht>M</geschlecht>
      <status>9</status>
      <beruf/>
      <l_jahre/>
      <l_monate/>
      <l_tage/>
      <l_stunden/>
    </person>
  </personen>
  <textus>Ulpus / Silva / nus / fac / tus / Ara / sione / emeri / tus leg (ionis) /
    II Aug (ustae) / votum / solvit</textus>
</record>

```

Fig. 7. Instance of the EDH XML Schema.

modelling the domain. Whereas the OWL 2 QL ontology allows us to answer user queries by automatically translating them into queries over the data sources (c.f., Section 5), the ORM2 model allows us to deploy the relational specification of the database (c.f., Section 4), and to ease the communication of computer scientists with domain experts.

3.1. The EPNet conceptual reference model

The specification of the EPNet CRM for the representation of epigraphic information and domain expert knowledge about Roman Empire Latin inscriptions is meant to unambiguously represent the way the data are understood by scholars, how they are connected together, and what their coverage is with respect to the literature of reference and the current research practices in the history of the Roman Empire. The CRM has been formally specified in the conceptual modelling language called ‘Object Role Modelling’ (ORM2) (Halpin and Morgan, 2010) by means of NORMA, a data modelling tool for ORM2.²¹

It is worth mentioning here that NORMA is equipped with a sound and complete reasoning module called ORMiE,²² providing automated reasoning services that facilitate the conceptual modelling activity. By relying on existing OWL 2 reasoners (e.g., HermiT, FaCT++), ORMiE provides: (i) *consistency checks* (where a schema is consistent if its classes/relations can be populated

without violating any of the constraints in the schema), (ii) *deduction of implicit constraints* (i.e., derived implicit ORM2 constraints, including inconsistent object types and fact types, are exposed to the modeller), and (iii) *automatic translation into OWL 2*. Previous results on the formal semantics of ORM2 and its sound and complete encoding into the description logic *ALCQI* provided the theoretical grounding for the design and implementation of ORMiE (Franconi et al. (2012)). Therefore, having specified the EPNet CRM in ORM2 gave us the advantage of relying on innovative technologies (i.e., NORMA) for deploying the new relational EPNet database and, thanks to ORMiE, for automatically translating the conceptual model into an OWL 2 QL ontology. Nonetheless, the graphical dress of ORM2, its intuitiveness and purely conceptual focus, make this language very well suited for knowledge engineering purposes, and especially for knowledge acquisition. The fact that a formally correct and complete translation into OWL 2 is also available and implemented, suggest the possibility to exploit ORM2 as the ultimate communication interface toward scholars without any formal background in conceptual modelling or knowledge representation. On the other hand, ontology editing tools like Protégé, while they can be used to directly specify OWL 2 ontologies, are not very intuitive for a non-expert audience. Negotiating with a domain expert a meaning (or, the definition of a concept) in front of the Protégé interface requires (from both sides) a deep understanding of the knowledge representation principles behind the tool, as well as of a non-intuitive vocabulary (e.g., the difference between ‘object properties’ and ‘data properties’, the fact that the relations are called ‘properties’), and of the overall organisation of a ‘knowledge

²¹ NORMA is an open source plug-in to Microsoft Visual Studio .NET freely downloadable from <http://www.ormfoundation.org/>.

²² <https://visualstudiogallery.msdn.microsoft.com/186cfe15-daeb-4429-bc32-46957a0828b2>.

```

@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix spatial: <http://geovocab.org/spatial#> .

<http://pleiades.stoa.org/places/579885#this> a <http://geovocab.org/spatial#Feature>;
  rdfs:label "Athenae";
  spatial:C <http://pleiades.stoa.org/places/579888#this>;
  rdfs:comment "A_major_Greek_city-state";
  rdfs:seeAlso <http://dbpedia.org/resource/Athens>;
  geo:lat 37.970793; geo:long 23.719537;
  skos:altLabel "Athenae", "Athens";
  foaf:primaryTopicOf <http://pleiades.stoa.org/places/579885> .

```

Fig. 8. The title and description of the Athens resource are carried by `rdfs:label` and `rdfs:comment` properties, while attested forms of its names or their transliterations are carried by `skos:altLabel` properties. A single point representing the location of Athens is carried by `geo:lat` and `geo:long` properties. A link to the Pleiades web page on Athens, via which all of these triples are editable, is carried by the `foaf:primaryTopicOf` property and a link to the DBpedia resource on Athens (are thereby Wikipedia) is made with an `rdfs:seeAlso` property. The four used namespaces (prefixed `foaf`, `geo`, `rdfs`, and `skos`) are among the most widely used one in linked open data, and they ensure a good baseline for usability (see <http://pleiades.stoa.org/Members/sgillies/news-items/linked-data-for-pleiades-places>).

base' (which is usually not the main goal of a knowledge acquisition campaign).

The EPNNet CRM has been defined according to the state-of-the-art of formal ontological models and standards for representing the structure of cultural heritage objects and the relationships between them. In particular, in order to increase the interoperability of the EPNNet CRM and of the whole EPNNet dataset, with other similar initiatives and data sources, the main module of the model results in a specialisation/extension of the well-known CIDOC CRM, the most dominant ontology in the digital cultural heritage area (Crofts et al.). Having chosen to be compliant with the CIDOC CRM also paves the way for further integration with the Europeana Data Model (EDM) (Clayphan, 2012), and the datasets that constitutes its EAGLE pillar.

For the sake of maintenance of the model, and according to the specific nature of the involved information, the EPNNet CRM has been structurally organised into distinct interrelated modules. Moreover, according to the different aim of each module, we again relied on existing standards for recording, aggregating, and publishing information on the Semantic Web, such as: (i) FaBiO (the FRBR-aligned Bibliographic Ontology), for the representation of the bibliographic references documenting the main entities in the CRM; (ii) the EAGLE metadata model, used internally by the aggregator on the EUROPEANA-EAGLE Project server, and the PETRAE guidelines,²³ for the epigraphic information digital representation; (iii) EpiDoc, the subset of the Text Encoding Initiative's standard (TEI) for the representation of texts in digital form, initially developed for the publication of digital editions of ancient inscriptions; (iv) the 'Places Ontology',²⁴ the 'DBpedia – Place ontology section', and the 'W3C: Basic Geo (WGS84 lat/long) Vocabulary', for representing information about spatially located objects.

In the following, we describe the five main modules characterising the EPNNet CRM: (1) MAIN deals with the representation of the relevant domain entities (e.g., inscriptions, amphoric types, associated epigraphic information), their properties (e.g., finding place, letter dimensions, archaeometric characterisation), and mutual relationships (see Fig. 4). Epigraphic objects, and inscriptions in particular, are usually understood both (i) as *material objects, engraved on different types of supports (e.g., monuments, burials, amphoras) also known as 'information carriers', and (ii) as pieces of knowledge represented in textual form. Roman inscriptions that are associated with the production and distribution of edible items (e.g., wine, oil, 'garum' or fish sauce) are characterised by being engraved, by means of different techniques, on special potteries called*

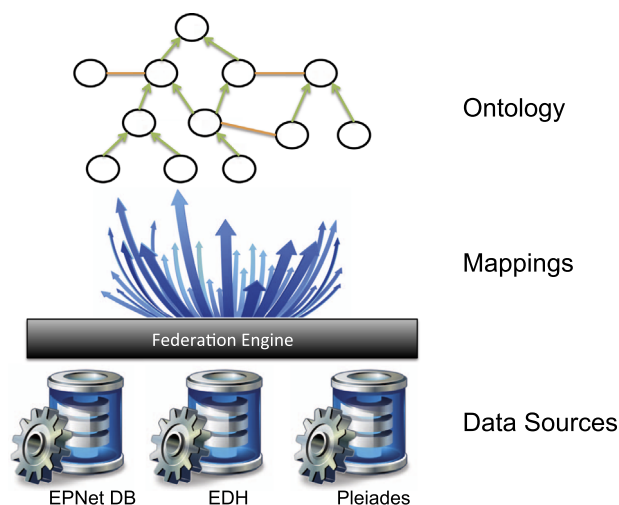


Fig. 9. Ontology-based integration of the three data sources in EPNNet.

'amphoras'. Amphoras are characterised by being instances of a specific AmphoricType (see Fig. 5)²⁵.

It is well-known that transcriptions play a central role in the study of inscriptions: transcriptions are used in studies and translations of their respective texts. Alternative transcriptions for the very same inscription may be the symptoms of the different scholarly interpretations that, if correctly retrieved together, can lead to new insights and the formulation of new historical hypotheses. The EPNNet CRM takes care of the transcriptions of the textual information brought by an inscription by means of specific concepts denoting: (i) the 'simple transcription' of the text into modern alphabets; (ii) the 'full transcription' of the conservation status of the text (specifying, for instance, the presence of backward or missing letters, erased but legible part of the text, etc.), and (iii) the EpiDoc compliant full transcription, having the goal of representing the original appearance of the inscription as accurately as possible. Inscriptions may mention the person names involved in the trade of the specific content of an amphora, and the professional activity of these persons, the sender and the receiver of a given item, their social status, the quantity and quality of the involved products, and the consular date. The EPNNet

²⁵ The catalogue of the amphoric types that is contemplated by the EPNNet CRM comes from the 'Roman Amphorae: a digital resource' project by the Archaeological Data Service Consortium (<http://archaeologydataservice.ac.uk/>), which offers a unique online introductory resource for the study of Roman amphorae: information about the 'distinctive features', 'data range', 'origin', geographic 'distribution', type of 'content', physical 'characteristics', 'petrology' analyses of the material composition of the amphoras, can be explored and retrieved.

²³ <http://petrae.huma-num.fr/>.

²⁴ <http://lov.okfn.org/dataset/lov/vocabs/place>.


```

SELECT ?x ?p ?o WHERE {
  { ?x rdf:type :InformationCarrier .
    ?x :hasProductionPlace ?pl .
    ?pl rdf:type :Place.
    ?pl dcterms:title "La_Corregidora".
    ?x ?p ?o .
  } UNION {
    ?x rdf:type :Amphora .
    ?x :hasProductionPlace ?pl .
    ?pl rdf:type :Place.
    ?pl dcterms:title "La_Corregidora".
    ?x ?p ?o .
  } UNION {
    ?x :hasAmphoricType ?z .
    ?x :hasProductionPlace ?pl .
    ?pl rdf:type :Place.
    ?pl dcterms:title "La_Corregidora".
    ?x ?p ?o .
  }
}

```

Fig. 10. Rewriting of the query in Example 4.

CRM contains specific concepts and relations to annotate all this information, thus allowing the final users to build complex queries involving them (e.g., “give me back all the inscriptions in the dataset talking about *olive oil* trade, and mentioning a person with the role of *negotiator*”).

Two main types of inscriptions have been considered in EPNet and consequently introduced in the EPNet CRM: the stamps, and the tituli picti. Stamps were superimposed to an amphora due to the control of the *annona* system during the long export of products within the borders of the Roman Empire: they normally registered personal names with one or more letters, names combined with other abbreviations, and special symbols (e.g., *ramus palmae*, *delfinus dextrorsum*). A ‘titulus pictus’ is a commercial inscription present on the surface of an amphora, specifying information such as the origin, the destination, and the type of product. Being able to query, compare, and aggregate information in a conceptually consistent way about both stamps and tituli picti is the way scholars support the interpretation of the meaning of the amphora texts, identify the production centres, and drawn new hypotheses on the chronology and features of the involved trade networks (Remesal Rodríguez, 2002; Remesal-Rodríguez, 2008).²⁶

(2) TIME offers a conceptual arrangement, driven by experts, of the different ways used to denote interval periods, dates, punctual instants of time, with respect to the given research domain. Inscriptions on amphoras for food transportation and market purposes, together with their associated information carriers, not only are *found* in a given moment in time by scholars, but they may talk about precise time periods by referring to them explicitly or by quoting the name of persons (e.g., the owners of the amphora content, the sellers, the traders, the producers of the amphora) involved in the exchange and transportation of the relative edible items. Nonetheless, the amphoras themselves are involved in a number of human activities (e.g., their production and potting), which need to be contextualised in time.

²⁶ Notice that in specifying the EPNet CRM an *ad-hoc* notational convention has been adopted in order to distinguish: (i) *classes/properties that have been simply imported from CIDOC (the labelling format starting with ‘E’/‘P’ and followed by a number reports the original CIDOC name of the class/property at hand)*, from (ii) *classes/properties that are meant to be specialisations of CIDOC elements (the labelling ‘EP-E’/‘EP-P’ followed by a number indicates that the element is a proper EPNet specialisation of the CIDOC element labelled with ‘E’/‘P’ followed by the same number)*.

In historical research, domain experts are used to deal with different formats for temporal information, and these have been taken into consideration and homogenised in the implemented OBDI system (c.f., Section 5). This has guaranteed the possibility of maintaining the epistemological flexibility the scholars are used to when looking for specific data, while keeping the possibility of interchanging between the different representations, and translating one into another (e.g., to bidirectionally move from the string ‘Trajan Government’ to the numerical time-span 98AD–117AD). Notice also that the relation P4:hasTimeSpan has been reified in the model in order to be ‘qualified’ by a probability: this modelling choice allows for the representation of *uncertain*, as well as absolute dating. The possibility of specifying the exact technique used to associate a date to a given activity, and object in the dataset, has been also considered.

(3) SPACE is meant to deal with information concerning spatial arrangement and geographical localisation of the entities in the EPNet CRM. A heterogeneous set of entities in MAIN brings a characterisation in terms of their spatial localisation, from finding activities involved in the discovery of a given artefact, to the relative position of a section of a so-called ‘titulus pictus’ with respect to other stylistic, textual, and structural elements of the surface of an amphora. For this reason, the SPACE component of the CRM has been subdivided into two distinct sub-modules: (i) a ‘carrier-centred’ one, used to represent the spatial relationships between the structural and the epigraphic components of an amphora (e.g., relative position of an inscription with respect to the amphora hands) and, (ii) a geographic-oriented one, which provides the elements for the representation of the location of a carrier finding, its production and potting, the ‘function’ of this location (e.g., civil settlement, legionary camp, fort) and the latitude and longitude coordinates identifying it on a map.²⁷

(4) DOCUMENTAL is the EPNet CRM module devoted to the representation of the bibliographic information documenting the entities of interest (e.g., conference and workshop papers, books, web portals, and digital encyclopaedia). As regards to this module, the above mentioned FaBiO ontology has been implemented, and customised according to the project requirements.

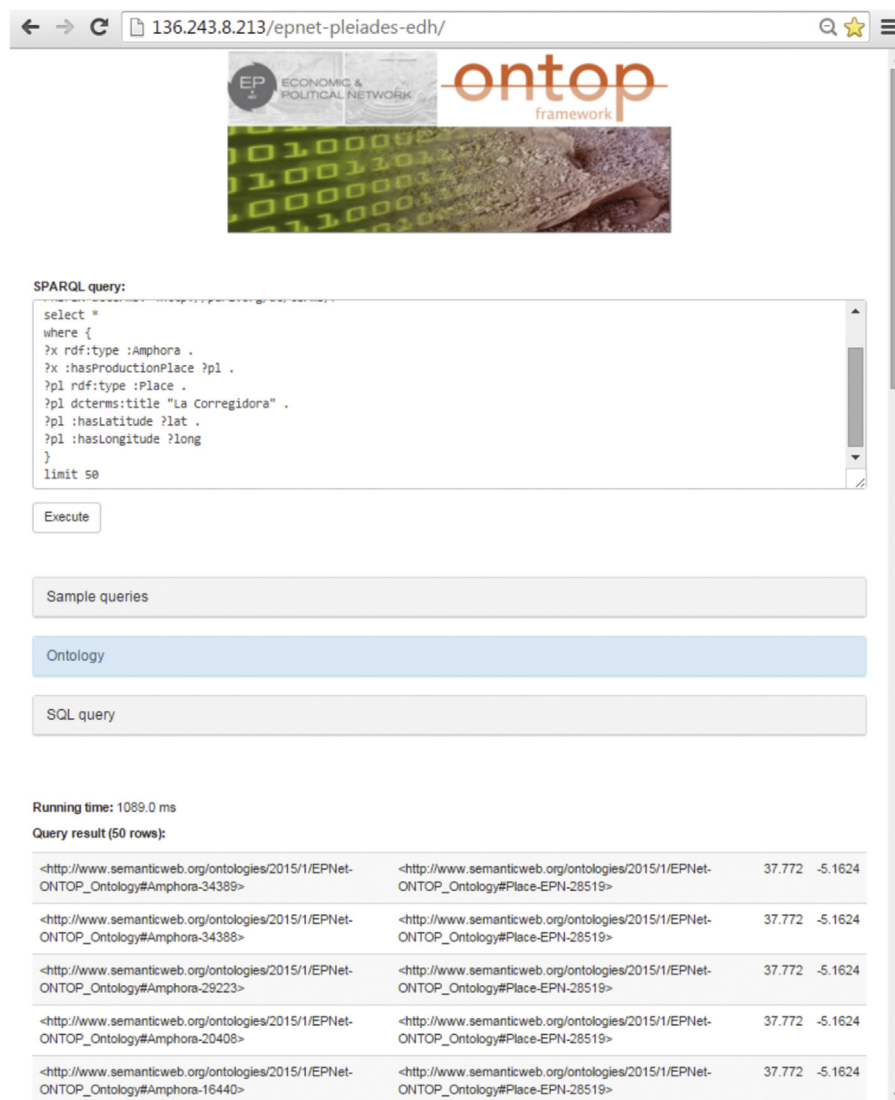
(5) UPPER TYPING is the module dedicated to collect all the taxonomical structures characterising the classes and properties in the MAIN CRM. Having all the taxonomical content of the EPNet CRM stored in a single conceptual place makes its management and successive extension extremely efficient and accurate. In this module, it is possible to find, e.g., specifications of the engraving techniques used to produce the inscriptions in the datasets, identified according to the state-of-the-art of the Roman Empire research literature.

The EPNet CRM consisting of the five modules introduced above, besides being formally correct, is comprehensive enough to host all the information and knowledge elicited with the domain experts, and represents a definitive improvement in data quality and granularity with respect to the previously adopted informal data structure descriptions that were available at the beginning of EPNet.

3.2. The EPNet ontology

In order to support the user with the possibility of accessing data through a domain-centred conceptual layer and terminology,

²⁷ The geographic module, complemented by information coming from different sources (e.g., the Pleiades and Pelagios – <http://pelagios-project.blogspot.com/> – datasets) offers the possibility to geo-localise the domain entities, as well as to make a distinction, and a semantically sound mapping, between historical (e.g., Roman provinces) and contemporary places. The role of the integration with the Pleiades dataset is further commented in Section 5.



The screenshot shows the OBDI web interface in a browser window. The address bar displays the URL `136.243.8.213/epnet-pleiades-edh/`. The page header includes the EP logo (Economic & Political Network) and the 'ontop framework' logo. Below the header is a banner image featuring binary code and a Roman amphora. The main content area contains a SPARQL query editor with the following query:

```

SPARQL query:
select *
where {
  ?x rdf:type :Amphora .
  ?x :hasProductionPlace ?pl .
  ?pl rdf:type :Place .
  ?pl dcterms:title "La Corregidora" .
  ?pl :hasLatitude ?lat .
  ?pl :hasLongitude ?long
}
limit 50

```

Below the query editor is an 'Execute' button. Underneath, there are tabs for 'Sample queries', 'Ontology', and 'SQL query'. The 'Ontology' tab is currently selected. Below the tabs, the execution results are displayed:

Running time: 1089.0 ms
Query result (50 rows):

| | | | |
|---|---|--------|---------|
| <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Amphora-34389> | <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Place-EPN-28519> | 37.772 | -5.1624 |
| <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Amphora-34388> | <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Place-EPN-28519> | 37.772 | -5.1624 |
| <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Amphora-29223> | <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Place-EPN-28519> | 37.772 | -5.1624 |
| <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Amphora-20408> | <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Place-EPN-28519> | 37.772 | -5.1624 |
| <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Amphora-16440> | <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#Place-EPN-28519> | 37.772 | -5.1624 |

Fig. 11. Screenshot of the OBDI web interface.

the relational specification introduced in the previous section has been automatically translated to an ontology, and enriched in its vocabulary. The resulting ontology is written in a formal language whose expressive power stays within the OWL 2 QL profile. This language allows for translating the input SPARQL queries into (possibly larger) SQL queries over the data sources that return all the facts that can be derived from the data, the mappings, and the ontology. This translation represents a form of reasoning over the ontology, whose axioms get compiled, together with the mappings, in the resulting SQL query. As a consequence, the ontology *modifies* and *extends* (by means of suitable class hierarchies, c.f., Example 4) the vocabulary of the database schema by re-introducing part of the domain-specific terminology extracted with the support of the domain experts. The ontology captures the domain knowledge by taking into consideration both the available data and the user requirements in terms of data accessibility and usage. The ontology contains 266 axioms, 234 classes, 35 object properties, and 21 data properties.

Differently from most current projects in cultural heritage and humanities, where the conceptualisation of the domain is expected to expose data models that are suitable for a generic audience (from tourists visiting a museum or searching on the Web their favourite piece of art, to public administrations willing to open up their cultural resources and historic properties), the EPNet CRM and the derived ontology have been specified in collaboration with experts of the history of the Roman economy with the main aim of: (i) supporting them in measuring aggregate changes over decades and centuries, (ii) trying out historical hypotheses across the time-scale of centuries, and (iii) systematically collecting information to question standard narratives (Guldi and Armitage, 2014). In that sense, the characteristic trait of the EPNet ontology, and of the domain knowledge encoded in the EPNet CRM, is that of being 'functional to research'.

The EPNet ontology contains axioms that provide formal definitions for the concepts (represented by OWL 2 classes) and relations (represented by OWL 2 properties) that experts make use

Ontology

Classes:

- Thing
 - Alphabet
 - AmphoraSection
 - Body
 - Foot
 - Handle
 - Neck
 - Rim
 - Shoulder
 - AmphoricType
 - Africana
 - Agora
 - Alia
 - Almagro

```

SQL query

SELECT *
FROM (
SELECT
1 AS "xQuestType", NULL AS "xLang", ('http://www.semanticweb.org/ontologies/2015/1/EPNet-O
NTOP_Ontology#Amphora-' || REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(RE
PLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(CAST(QV
IEW1."id" AS VARCHAR(10485760)),' ','%20'),!','%21'),'@','%40'),'#','%23'),'$', '%24'),'
&','%26'),' ','%42'),'(','%28'),'(',')','%29'),'[','%5B'),'']','%5D'),' ','%2C'),' ','
'%3B'),' ':'','%3A'),'?'','%3F'),'=','%3D'),'+','%2B'),'','','%22'),'/','%2F')) AS "x
",
1 AS "plQuestType", NULL AS "pllang", ('http://www.semanticweb.org/ontologies/2015/1/EPNet
-ONTOP_Ontology#Place-' || REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(RE
PLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(CAST(QV
IEW3."location" AS VARCHAR(10485760)),' ','%20'),!','%21'),'@','%40'),'#','%23'),'$', '%
24'),'&','%26'),' ','%42'),'(','%28'),'(',')','%29'),'[','%5B'),'']','%5D'),' ','%2C')
',' ','%3B'),' ':'','%3A'),'?'','%3F'),'=','%3D'),'+','%2B'),'','','%22'),'/','%2F'))
AS "pl",
6 AS "latQuestType", NULL AS "latLang", CAST(QVIEW5."latitude" AS VARCHAR(10485760)) AS "1
at",
6 AS "longQuestType", NULL AS "longLang", CAST(QVIEW5."longitude" AS VARCHAR(10485760)) AS
"long"
FROM
InformationCarrier QVIEW1,
Producing QVIEW2,
Activity_Location QVIEW3,
ProductionPlace QVIEW4,
GeographicLocation QVIEW5
WHERE
QVIEW1."id" IS NOT NULL AND
(QVIEW1."producing" = QVIEW2."id") AND
(QVIEW1."producing" = QVIEW3."activity") AND
(QVIEW3."location" = QVIEW4."id") AND
QVIEW3."location" IS NOT NULL AND
(QVIEW3."location" = QVIEW5."id") AND
(QVIEW5."name" = 'La Corregidora') AND
(QVIEW5."longitude" IS NOT NULL AND QVIEW5."latitude" IS NOT NULL) AND

```

Fig. 13. SQL query that is actually executed on the datasets as reported by the web interface.

Query result (only first 50 rows shown):

| | |
|--|---|
| <http://136.243.8.213/obdasystem#Amphora-548> | <http://136.243.8.213/obdasystem#Place-27680> |
| <http://136.243.8.213/obdasystem#Amphora-549> | <http://136.243.8.213/obdasystem#Place-27680> |
| <http://136.243.8.213/obdasystem#Amphora-550> | <http://136.243.8.213/obdasystem#Place-27680> |
| <http://136.243.8.213/obdasystem#Amphora-551> | <http://136.243.8.213/obdasystem#Place-27680> |
| <http://136.243.8.213/obdasystem#Amphora-552> | <http://136.243.8.213/obdasystem#Place-27680> |
| <http://136.243.8.213/obdasystem#Amphora-553> | <http://136.243.8.213/obdasystem#Place-27680> |
| <http://136.243.8.213/obdasystem#Amphora-554> | <http://136.243.8.213/obdasystem#Place-27680> |
| <http://136.243.8.213/obdasystem#Amphora-1618> | <http://136.243.8.213/obdasystem#Place-27680> |

Fig. 14. Results of the user's query execution shown in the web interface.

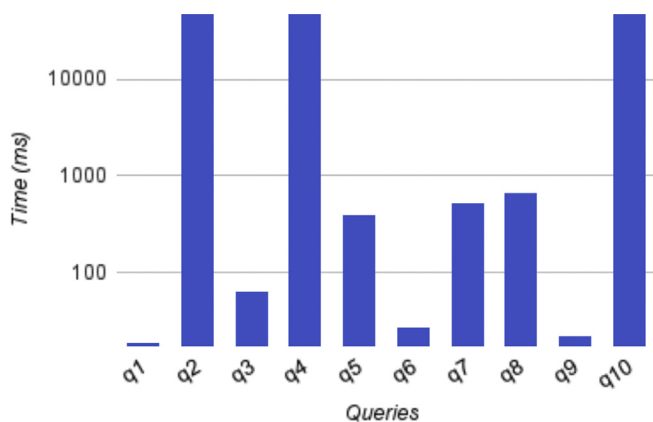


Fig. 15. Query Execution Time.

of in conceptually classifying the entities of their research domain. As an example,²⁸ the axioms below say that the classes :Stamp and :TitulusPictus are both subclasses of :Inscription (see Fig. 4), while :Amphora is a specialisation of :InformationCarrier. The :carriedBy property links :Inscription with their informational carrier and, similarly, the domain and range of the :producedAt property are specified to be respectively the :InformationCarrier class and the :TimeSpan in which the existence of the carrier is historically attested. The last axiom below is for characterising :hasName as a datatype property, i.e., a property having a range in a specific datatype (:String in this case).

```
:Stamp rdfs:subClassOf :Inscription.
:TitulusPictus rdfs:subClassOf :Inscription.
:Amphora rdfs:subClassOf :InformationCarrier.
```

```
:carriedBy rdfs:domain :Inscription.
:carriedBy rdfs:range :InformationCarrier.
:producedAt rdfs:domain :InformationCarrier.
:producedAt rdfs:range :TimeSpan.
:hasName rdfs:type owl:DatatypeProperty.
```

In addition, in order to expose the user to a domain-oriented vocabulary that is easily understandable by scholars, specific axioms have been added to the ontology. For instance, the following axiom introduces in the ontology the new property :engravedOn, by asserting that it generalises the :carriedBy property between inscriptions and their informational carriers:

```
:carriedBy owl:subObjectPropertyOf :engravedOn.
```

In this way, the final users can query the datasets by relying on newly introduced terms. We introduce also the following axiom, which will be used in Section 5:

```
:hasAmphoricType rdfs:domain :Amphora.
```

Notice that the expressive power of OWL 2 QL allows for the specification, among others, of disjointness constraints between classes, in this way supporting data consistency checking that can be automatically performed by means of traditional reasoning techniques (see Section 5.3). Being able to apply consistency checks over the project data is of particular interest in such a context, considering that the data are often collected by non-experts and manually entered into a DB system without the support of any specific data entry interface.

4. The datasets

In this section, we introduce the EPNet, Pleiades, and 'Epi-graphic Database Heidelberg' datasets, whose data content has been integrated in the EPNet project, thus increasing the coverage

²⁸ A more comprehensive picture of the ontology can be found at <http://136.243.8.213/obdasystem/>, where a simple user interface has been implemented with the only aim of testing the system and its basic query functionalities.

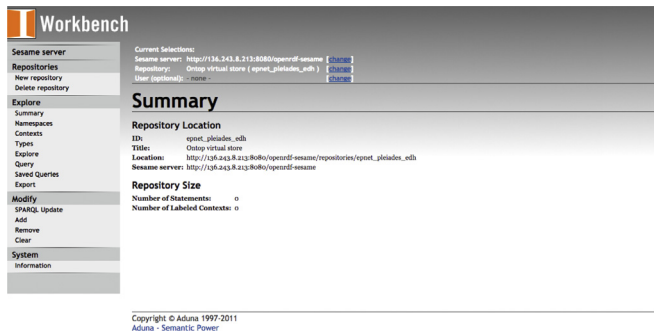


Fig. 16. -ontop- as a SPARQL Endpoint.

of the data provided to the final users with respect to the domain of interest (*completeness*), and complementing the characterisation of the geographical entities already present in the EPNet dataset (*accuracy*).

4.1. The EPNet relational repository

Although the EPNet CRM represents the experts' knowledge of the domain, it does not specify how to store the actual data. Data storage substantially depends on the underlying technology, i.e., different technologies store data in different ways, which results in a specification that is tied to the particular technology being used. Since the knowledge of the domain is independent of any particular technology, it is a common practice to specify data storage separately from it.

In EPNet, we use a relational database management system (RDBMS) to store the data, so we must provide a relational specification that complements the EPNet CRM. A RDBMS structures data in the form of tables (a.k.a., relations), so a relational specification has to indicate which are the tables that form the database and which are their attributes (a.k.a., columns). It is important to note that the data currently available in the project does not cover the entirety of the domain's knowledge represented in the CRM, but rather a subset of it. Consequently, our efforts in providing a relational specification have focused so far on this specific subset of the domain. Due to space reasons, only a small fragment of this relational specification is shown in Fig. 6. Tables are depicted as boxes, with their name at the top (e.g., *inscription*), followed by the list of attributes (e.g., *id*, *carrier*, ...). Each attribute consists of a name and a data type (e.g., *id* *INT(11)*, which indicates that identifiers of inscriptions are integer numbers). In particular, notice the tables *informationcarrier*, *amphotyping*, and *amphoratype*, which we will be using in the examples in the following section. *informationcarrier* stores data about amphorae, such as an identification number and a reference to both their producing and finding activities (detailed data about these activities is stored in separate tables). *amphoratype* records the information of each kind of amphora. *amphotyping* links amphora identifiers with the corresponding type identifier(s) (these could be more than one, if the exact amphora's type could not be identified but was narrowed down to a small set of possible types instead). Relationships between tables are depicted in the specification as lines connecting them. Consider, for example, the line connecting *amphotyping* with *informationcarrier*, which represents the fact that the value for the *carrier* attribute of an entry in *amphotyping* must be the value of the *id* attribute of some entry in *informationcarrier*; similarly for the line connecting *informationcarrier* and *amphoratype* with respect to values of their *type* and *code* attributes, respectively.

4.2. Epigraphic database heidelberg

The Epigraphic Database Heidelberg EDH²⁹ (Feraudi-Gruenais, 2010) is a funding member of both the Electronic Archive of Ancient Greek and Latin Epigraphy (EAGLE) and of the Europeana Best Practice Network for Ancient Greek and Latin Epigraphy (EAGLE BPN).³⁰ The primary responsibility of EDH as an epigraphic archive in the EAGLE consortium is to collect all Latin and bilingual (Greek and Latin) inscriptions from the provinces of the Roman Empire. This is accomplished with the development and maintenance of four databases: one for transcriptions, one for photos of inscriptions³¹, a bibliographic database³² for all publications cited in the inscriptions database, and a geographic database. EDH inputs data from the entries in the *Année Épigraphique*³³ and checks all readings, providing not just a reproduction of published text but also a new edition that follows simple and functional encoding standards for what concern the text. It provides also a very thorough set of structured information and metadata about the support of the inscription, the text and its typology, and the contents and themes. It allows for complex searches across the more than 70.000 records and for rich information to be returned to the user, with photos where available, and links to translations when these are present in the EAGLE MediaWiki³⁴.

In the context of the EAGLE BPN project, a consortium of all major epigraphic databases in Europe and abroad,³⁵ EDH leads the networking, harmonization and content curation actions which aim at providing an equal standard of data about inscriptions to the aggregator which prepares data for harvest from Europeana.³⁶ Among other harmonization and content curation tasks, EDH uses a simplified TEI/EpiDoc specification to encode all data about inscriptions and supports the direct production of EpiDoc editions of the texts, conversion to EpiDoc and also crosswalk from EpiDoc to its own data model.

The EDH model is structured to contain all information relevant to epigraphic studies: the text in a diplomatic and in a critical edition, reference bibliography for the reading and interpretation of the inscription, provenance, original place of location, material, type of inscription and topic or context, execution technique, year of finding, language, dimensions, date, a commentary, and structured information for each person attested in the document, including praenomen, nomen, cognomen, tribus, origo, gender, status, occupation, and age of death. EDH hosts also a local stable identifier for each inscription, the Trismegistos identifier of the text,³⁷ an authoritative identifier and, when available, a Pleiades id for the place of origin of a monument. In Fig. 7, we show an example instantiation of the EDH model. This structure is represented following the Text Encoding Initiative specification EpiDoc (Epigraphic Documents in TEI XML) Elliott et al. (2007), which has been specifically developed for documents from the ancient world, like inscriptions and papyri. This schema has been developed to meet the needs of epigraphists and has met their wide acceptance by now, becoming the de facto standard for digital editions of ancient texts. EDH has been involved in the EpiDoc consortium since its very early stages, and has initially produced mappings and transformations to crosswalk EpiDoc XML into EDH.

²⁹ <http://edh-www.adw.uni-heidelberg.de/>.

³⁰ More information on the network and the research carried out can be found in Orlandi et al. (2014).

³¹ <http://edh-www.adw.uni-heidelberg.de/foto/suche/>

³² <http://edh-www.adw.uni-heidelberg.de/bibliographie/suche/>

³³ <http://www.anneeepigraphique.msh-paris.fr/>.

³⁴ <http://www.eagle-network.eu/wiki/>.

³⁵ Founded by the European Commission ICT-PSP from 2013 to 2016, Grant Agreement n. 325122.

³⁶ <http://www.europeana.eu/portal/>.

³⁷ <http://www.trismegistos.org/>.

One of the basic principles of the working method of EDH is that inscription readings are not simply accepted from the editions and secondary literature. As much as possible, these readings are verified at least on the basis of drawings or photographs. EDH focuses on the recollection of epigraphs engraved mainly on monuments. This kind of epigraphic data is usually rich in information about distinguished personalities of the Roman society and their social connections. In this sense, EDH is a nice complement for the EPNet dataset, as epigraphs on amphorae are brief and do not give much details on the role played by the persons being mentioned. Being able to correlate the names found on amphorae with the "social networks" described in the monumental epigraphy should put us in a better position to understand the trade routes and the agents involved.

4.3. Pleiades

Pleiades³⁸ is an open-access digital gazetteer for ancient history. It provides stable Uniform Resource Identifiers (URIs) representations for tens of thousands of geographic entities. Built on the Classical Atlas Project (1988–2000), which produced the 'Barrington Atlas of the Greek and Roman World' (Princeton, 2000), Pleiades is co-organised by the Institute for the Study of the Ancient World and the Ancient World Mapping Center (UNC Chapel Hill). Pleiades is beginning to expand beyond its classical Greco-Roman roots and is establishing lines of interoperability with a number of other web-based resources treating the geographical, textual, visual, and physical culture of antiquity. The Pleiades places dataset consists of more than 210,000 subject–predicate–object statements about more than 34,000 ancient places. It is written in a single text file using the terse Turtle RDF (Resource Description Framework) format (see Section 5). Data about the city of Athens, for instance, are represented in Fig. 8.

The Pleiades dataset has been selected in order to complement the EPNet dataset. In particular, it provides a set of geographic entities that strictly contains those present in the EPNet dataset (e.g., specific municipalities and Roman provinces are present in EPNet only if they are a finding, producing, or potting place). The

```
http://project-url.org/1 rdf:type :Amphora .
http://project-url.org/1 :hasProductionPlace http://project-url.org/place/5 .
```

integration with Pleiades supports EPNet in tracing trade routes and economic connections on the Roman Empire territory in a more precise way, and provides a satisfactory picture of the past anthropic environment. If a location is present in both the Pleiades and the EPNet DB but missing some attributes in the latter (e.g., the place has no geo-coordinates), one would like to identify the missing attributes, catch their associated values, and with them augment the entry in EPNet, thus increasing the overall accuracy and completeness of the stored data.

5. Ontology-based data integration in EPNet

Since the mid 2000s, *Ontology-Based Data Access and Integration* (OBDA/OBDI) has become a popular approach for providing an integrated, uniform access to heterogeneous data sources. This paradigm allows us to successfully address all the

issues mentioned in Section 2. The overall architecture of an OBDI system is shown in Fig. 9, where as data sources we have considered the ones adopted in the EPNet project and discussed in Section 4. In OBDI, a conceptual layer is given in the form of an ontology (c.f. Section 3), which captures knowledge about the domain of interest, and provides a high-level conceptual view of the underlying data sources in terms of a shared vocabulary. The ontology is connected to the data sources through a declarative specification, given in terms of *mappings* that relate symbols in the ontology (classes and properties) to (SQL) views over data. In this way, through the mappings, users can query the data sources using the shared ontology vocabulary, without the need of understanding the precise structure of the sources, the relations between them, or the encoding of the data. Making use of the mappings, the OBDI system translates the user queries into SQL queries formulated over the sources, while at the same time exploiting the domain knowledge encoded in the ontology to overcome incompleteness in the data and enrich query answers.

The mappings expose the data in the data sources as Resource Description Framework (RDF) triples. RDF³⁹ is a World Wide Web Consortium (W3C) specification for data interchange on the Web. This standard is based upon the idea of making statements about resources in the form of expressions with three components: (*subject predicate object*). In RDF terminology, these expressions are known as *triples*. The subject and object in an RDF triple are *resources*, i.e., individuals represented, e.g., by an URI (Uniform Resource Identifier), or values represented by literals (belonging to predefined datatypes, such as strings or integers), while the predicate must be an individual. URIs can be abbreviated with prefixes such as:

```
@prefix <http://example.org/book/>.
@prefix myPre:<http://example.org/book/myPre/>.
```

Then, statements such as :1 and myPre:2 respectively represent the URIs:

```
http://example.org/book/1
http://example.org/book/myPre/2
```

The prefix consisting only in a colon (:) is called default prefix. Examples of RDF triples are the following:

which respectively state that the element represented by the URI <http://project-url.org/1> is an amphora, and that it was produced in the place represented by the URI <http://project-url.org/place/5>.

Each *mapping* consists of a *source query*, which is an SQL query formulated over the data sources, and a *target triple pattern*, which contains *URI templates* that refer to the answer variables (i.e., attributes) of the source query:

```
subject predicate object ← Source SQL query
      Target triple pattern
```

When the source query is evaluated over a data source, each answer tuple is used to instantiate the URI templates in the target triple pattern, and thus generates an RDF triple. We illustrate this with a simple example.

Example 2 (URI template instantiation). Consider a source query with an answer variable *ic_id*, and a target triple pattern

³⁸ <http://pleiades.stoa.org>

³⁹ <http://www.w3.org/RDF/>

containing a URI template of the form `:Amphora-{ic_id}`, where answer variables to be instantiated are enclosed in `{...}`. Then, if the evaluation of the source query returns, e.g., a tuple in which `ic_id` is instantiated to '1', the URI template will generate the URI `:Amphora-1`. Recall that the colon symbol ':' represents the default URI prefix, in this example `http://project-url.org/`, hence the generated URI will actually be `http://project-url.org/Amphora-1`.

The following example illustrates a mapping that is used to populate a class of the ontology by exploiting this mechanism.

Example 3 (*Mapping to populate a class*). The following mapping populates the class `:Dressell` (which is a subclass of `:AmphoricType`), considering that an RDF triple in which the predicate is `rdf:type` states the membership of the individual representing the subject in the class representing the object.

```

:Amphora-{ic_id} rdf:type :Dressell ←
SELECT ic_id AS ic_id, t.code AS t_code
FROM InformationCarrier ic JOIN AmphTyping amt ON amt.carrier=ic_id
JOIN AmphoraType t ON t.code=amt.type
WHERE amt.type='DRI'

```

Observe that the SQL query in this mapping is rather complex, since it joins information extracted from three different tables. This complexity is hidden to the user, who is exposed only to the simple concept `:Dressell`.

The standard mapping language is called R2RML,⁴⁰ however here we have used a simplified syntax for the sake of presentation. The RDF graph exposed by the ontology, and defined from the data through the mappings, can now be queried, by adopting SPARQL, the standard query language for the Semantic Web. Answering such queries will involve reasoning over the axioms of ontology and the mappings. We illustrate this by means of an example.

Example 4. Consider Example 1, where the user is interested in retrieving all the information carriers produced in 'La Corregidora' and all the information available about them. This can be expressed by means of the following SPARQL query using the vocabulary of the ontology.

```

SELECT ?x ?p ?o WHERE {
  ?x rdf:type :InformationCarrier .
  ?x :hasProductionPlace ?pl .
  ?pl rdf:type :Place.
  ?pl dct:terms:title "La_Corregidora".
  ?x ?p ?o .
}

```

Observe that users do not need to know the particular codes of the amphoras, nor they need to manually integrate the information coming from EPNet, EDH, and Pleiades. The first four triples in this query obtain all the objects (amphoras, stones, monuments) `?x` produced in 'La Corregidora'. Observe that here we re-used the vocabulary coming from the Dublin Core Metadata Initiative (`dct:terms:title`). The last triple (`?x?p?o`) will return all the available information for the individual instantiating variable `?x`. We illustrate how reasoning comes into play when determining the answers to this query: consider an individual a_1 produced in 'La Corregidora' that has an amphoric type. Then a_1 and all its information should be returned as an answer, since one can *infer* that a_1 is of type `:InformationCarrier`. Indeed, the domain of `:hasAmphoricType` is `Amphora`, which in turn is a subclass of `:InformationCarrier`. It is worth noticing that one could also

return the geo-coordinates of 'La Corregidora' by including in the query the properties `:hasLatitude` and `:hasLongitude`.

Several OBDI systems are currently available (Rodríguez-Muro and Rezk, Bishop et al., 2011; Sequeda et al., 2014; Civili et al., 2013). We work with `-ontop-` (Rodríguez-Muro and Rezk, Rodríguez-Muro et al., 2013; Calvanese et al., 2016), a mature open-source system, which is currently being used in academia and in industry (Giese et al., 2015; Calvanese et al., 2015; Kharlamov et al., 2015, 2014; Lopez et al., 2015; Bouchou and Niang, 2014; Rahimi et al., 2014). `-ontop-` allows users to answer queries in two modalities: in the *materialized* approach, the system effectively generates the RDF graph defined by the ontology, the mappings, and the data, and then such graph is used to answer queries, as in an RDF triplestore; alternatively, the RDF graphs can be kept *virtual*, and the SPARQL user queries are answered by translating them to SQL queries over the data sources, making use both of the axioms in the ontology and of the mappings. The virtual approach avoids the cost of materialization and allows one to profit from more than 30 years of maturity of relational database systems (including efficient query answering, robust transaction support, security, etc.). `-ontop-` supports the virtual approach with all major RDBMSs, such as Oracle, IBM DB2, Microsoft SQL Server, PostgreSQL, and MySQL.

To answer queries in the virtual approach exploiting the information in the ontology, `-ontop-` relies on *query-rewriting* (Calvanese et al., 2007). We illustrate this on our running example.

Example 5. Consider the SPARQL query in Example 4. To determine the instances of the class `:InformationCarrier`, `-ontop-` relies on the ontology to infer that all instance of the classes `:Monument` and `:Amphora` also belong to `:InformationCarrier`, and hence that the same holds also for all elements in the domain of `:hasAmphoricType`. This inference is then used to rewrite the query in Example 4 into a union of triple patterns, where each triple pattern represents a conjunction of conditions. The union contains one triple pattern for each subclass of `:InformationCarrier` and for each property that has `:InformationCarrier` (or one of its subclasses) as domain or range. This is illustrated in Fig. 10.

It is important to observe that to rewrite a query, `-ontop-` does not need to reason with the data, but only with the axioms in the ontology. This is what makes this approach scalable.

`-ontop-` is available as a plugin for Protégé 5, one of the most widely adopted editors for OWL ontologies, as a SPARQL endpoint through Sesame Workbench, and as a Java library supporting the standard OWL API and Sesame API.

5.1. Integration of project data

As mentioned, `-ontop-` allows for virtual data integration, where the data remain in the sources (in our case, EPNet, EDH, and the Pleiades dataset) and are accessed at query time. To do so, `-ontop-` does not need to modify the underlying databases, which is a requisite in the use case we are considering, and it also, does not require complex extract, transform, and load processes.

`-ontop-` does not integrate the database at the SQL level, and relies for that on a standard federation engine, such as Teiid⁴¹ or Exareme (Kilapi et al., 2015). Such engines expose a relational schema (or a set of relational schemas) containing all tables from the federated data sources. `-ontop-` does the semantic integration and homogenization over these federated databases. Here we

⁴⁰ <http://www.w3.org/TR/r2rml/>

⁴¹ <http://teiid.jboss.org/>

discuss the integration of EPNet, EDH, and Pleiades, focusing on space and time periods. The integration starts in the ontology, whose concepts cover the information contained in the three datasets. The `:Place` concept, for instance, is characterised in the ontology by having a given function or role (e.g., `:Production-Place`, `:CivilSettlement`, or `:LegionaryCamp`), it is linked through the `:hasLatitude` and `:hasLongitude` properties to its geo-coordinates, and `:fallsWithin` or `:isContainedIn` other known places. Then the information from the three datasets get connected through properties. Three properties that are relevant for our running example are:

- `:hasProductionPlace`, connecting information carriers from EPNet and EDH to places in Pleiades;
- `:hasLatitude`, connecting places in Pleiades with latitude coordinates in the three datasets; and
- `:hasLongitude`, connecting places in Pleiades with longitude coordinates in the three datasets.

Space: All the datasets, EPNet, EDH, and Pleiades, have information regarding places, settlements, geo-coordinates, etc. However, Pleiades is more complete in this regard, and also contains kinds of settlements that are missing in EPNet and EDH. If a place is not in the EPNet dataset, we completely rely on the data from Pleiades (name(s), geo coordinates, and the kind of settlement). If a place is in EPNet (where the comparison is done by name), then we keep the existing EPNet data, and add the kind of settlement (which is not available in EPNet), obtaining it from Pleiades. We operate in a similar way if the existing data is incomplete (e.g., due to missing coordinates).

To cluster all the information about places in all datasets into a single well-defined concept `:Place`, we make use of suitably defined mappings. For the sake of presentation, we show here a simplified variant of the mappings:

```
EPNet: :Place-{gl_id} rdf:type :Place ←
      SELECT gl_id AS gl_id
      FROM GeographicLocation gl

EDH: :Place-EDH-{edh_place_nid} rdf:type :Place ←
      SELECT ep.numberId AS edh_place_nid
      FROM edh.findingPlaces efp JOIN edh.places ep ON efp.spot=ep.id

Pleiades: pleiades:{path} rdf:type :Place ←
      SELECT pp.path AS path
      FROM pleiades.places pp JOIN pleiades.names pn ON pn.pid=pp.id
```

Observe that URIs here also encode provenance information, namely, “pleiades” and colon (EPNet default URI). This can help the user to assess where the information is coming from.

Time: Regarding time periods, EPNet and Pleiades specify time periods using lists of integers. For example, [(98,117), (130,140)] is used to state that an object or a place existed either in the period 98 AD – 117 AD, or in the period 130 AD – 140 AD. Besides these numeric values, users often are interested in using governments as time periods. For example, instead of using 98 AD – 117 AD, they prefer to use the term “Trajan Government”. To achieve that, we add a mapping defining the term “Trajan Government” as follows:

```
:Amphora-{ic_id} :producedAt :Trajan-Government ←
      SELECT ic.id AS ic_id
      FROM <complex join>
      WHERE startYear <= 117 AND endYear >= 98
```

Now the user can query the amphoras in production during this period using one of the following two equivalent queries:

```
SELECT * WHERE {
  ?x rdf:type :Amphora .
  ?x :producedAt :Trajan-Government .
}
```

and

```
SELECT * WHERE {
  ?x rdf:type :Amphora .
  ?x :producedAt ?y .
  ?y rdf:type :YearSpan .
  ?y :startsAt ?s .
  ?y :endsAt ?e .
  FILTER (?s <= 117 && ?e >= 98)
}
```

Neither of these formats for time follows the standard formats (xsd:gYear, xsd:dateTime, xsd:period, etc.). However, adding them would simply require the small effort of adding a few mappings.

5.2. Linking entities in different data sources

When integrating several datasets, complementary information about the same entity might be distributed over several data sources, and moreover the same entity might be represented using different identifiers in each data source. The first important issue that comes up is that of *entity resolution*, which requires to understand which records actually represent the same real world entity.

In this work we are concerned with linking places and time periods, which are the two categories shared by the three datasets. To understand which records in the DB represent the same place we compared their names. To understand which DB records represent the same time period, it is necessary to homogenize first the type of the DB records (e.g., strings vs. integers) and then the format for the time interval (e.g., start and end date). Observe that these changes do not correspond to actual modifications in the external datasets, but are just used to understand which data items correspond to each other, and thus should be linked.

Traditional relational data integration techniques use extract, transform, and load (ETL) processes to address this problem. These techniques usually (require to) choose a single representation of the entity, merge the information available in all data sources, and then answer queries on the merged data. Unlike the traditional approach, we *virtually merge* the data, by defining the mappings in such a way that they consistently generate only one URI per real world entity, *without* changing the original datasets.

Example 6. In the case of places, consider the following query, which looks for all the inscriptions on amphorae found in the city of Rome:

```
SELECT ?x ?t WHERE {
  ?x rdf:type :Inscription .
  ?x :engravedOn ?y .
  ?y rdf:type :Amphora .
  ?y :hasFindingPlace ?z .
  ?z dct:title "Roma" .
  ?x :isTranscribedBy ?u .
  ?u :hasTranscription ?t .
}
```

Such a query translates into an SQL query that accesses the three underlying datasets, given that they all store information about places:

```
SELECT ...
FROM edh.inscriptions QVIEW1,
     edh.findingPlaces QVIEW2,
     edh.places QVIEW3,
     edh.places_not_in_corpus2 QVIEW4,
     pleiades.places_not_in_corpus2 QVIEW5,
     pleiades.places QVIEW6,
     pleiades.names QVIEW7
WHERE ...
UNION SELECT ...
FROM Inscription QVIEW1,
     InformationCarrier QVIEW2,
     Finding QVIEW3,
     Activity_Location QVIEW4,
     GeographicLocation QVIEW5,
     LinguisticObject_Inscription QVIEW6,
     LinguisticObject QVIEW7
WHERE ...
UNION ...
```

Similarly, for the case of time periods, consider the following query that looks for inscriptions dated in a time period that overlaps with 70 AD – 75 AD:

```
SELECT ?x WHERE {
  ?x rdf:type :Inscription .
  ?x :engravedOn ?y .
  ?y rdf:type :Amphora .
  ?y :hasProductionDate ?z .
  ?z :startsAt ?sy .
  ?z :endsAt ?ey .
  FILTER (?sy <= 75 && ?ey >= 70)
}
```

This query results in an SQL query that accesses both the EPNet and the EDH datasets, since these are the two sources that store production dates:

```
SELECT ...
FROM edh.inscriptions QVIEW1
WHERE ...
UNION SELECT ...
FROM Inscription QVIEW1,
     LinguisticObject_Inscription QVIEW2,
     InformationCarrier QVIEW3,
     Producing QVIEW4,
     Dating QVIEW5,
     Dating_YearSpan QVIEW6
WHERE ...
UNION ...
```

5.3. Data consistency

OWL 2 and its profile OWL 2 QL are grounded in description logic, and an ontology expressed in these languages encodes knowledge about the domain of interest in terms of a logical theory. Such theory constrains the possible instantiations of the classes and properties of the ontology. In an OBDI scenario as the one we are considering here, the data retrieved from the sources by means of the mappings might violate the constraints expressed in the ontology. Such violations are an indication of potential problems, either with the data, or with the specification of the ontology and/or the mappings. In any case, it is desirable to detect violations of ontology constraints, so as to be able to intervene and possibly correct them.

-ontop- provides means to express (and detect the violation of) two types of constraints⁴²:

- *Disjointness* is used to express that two classes (or two properties) should not have elements (or pairs of elements) in common. For instance, this should hold for the two classes : MilitarCamp and CivilSettlement.
- *Functionality of Properties* is used to impose that a property (or its inverse) cannot relate an individual to more than one individual. For instance, the property :hasShape is functional since every amphora can have at most a single shape.

Interestingly, for the lightweight ontology language OWL 2 QL, checking the consistency of the ontology, the mappings, and the data can be reduced to answering an SQL query over the data (Calvanese et al., 2007). For instance, to check that : MilitarCamp and :CivilSettlement are disjoint, one can make use of a SPARQL query asking for the existence of an individual that belongs to these two classes simultaneously. The answer to that query is obtained by rewriting it to an SQL query and evaluated it over the sources, and if that answer is not empty, then the OBDI setting is inconsistent. Analogously, one can check if a property such as :hasShape is functional.

5.4. User interface

A preliminary user interface for testing the OBDI functionality in EPNet is available online.⁴³ The screenshot in Fig. 11 shows an overview of the interface, while details are shown in screenshots in Figs. 12–14.

The interface provides users with a text area where to write SPARQL queries (e.g., the query in Example 4) using the vocabulary of the ontology discussed in Section 3 (for the user's convenience, a summary of the ontology is provided by the interface; see Fig. 12). In the future we plan to provide a visual query interface to help the users to build their queries without the need of knowing SPARQL.

Following SPARQL syntax, users need to begin their queries with a prefix declaration, which in our case is:

```
PREFIX : <http://www.semanticweb.org/ontologies/2015/1/EPNet-ONTOP_Ontology#>
PREFIX rdf : <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX dcterms : <http://purl.org/dc/terms/>
```

After executing the query, the interface shows the SQL query that was sent to the underlying RDBMS (Fig. 13), and the result of the query in tabular form (Fig. 14). In the user interface, we provide a catalogue made of the following 10 sample queries:

- q₁: Amphoras produced in 'La Corregidora' and its geo-coordinates.
- q₂: Inscriptions on amphoras found in the city of Rome.
- q₃: Inscriptions on amphoras dated in a period that overlaps with 70 AD – 75 AD.
- q₄: Inscriptions on amphoras found in the city of 'Mainz' containing the text 'PNN'.
- q₅: All the properties associated with the amphoras that have been produced in the archaeological site named 'La Catria'.
- q₆: Known materials used to produce some informational carrier.
- q₇: Distinct 'Civil Settlements' where amphoras of type 'Dressel' have been found.
- q₈: Inscriptions having a TEI/EpiDoc conformant transcription, found in the city of 'London', and produced during the

⁴² OWL 2 QL allows for expressing disjointness assertions between classes and properties, but it does not support functionality of properties.

⁴³ <http://136.243.8.213/epnet-pleiades-edh/>.

'Trajan Government'.

q_9 : Inscriptions having a transcription conformant to the CEIPAC notation, found in the city of 'London', and produced during the 'Trajan Government'.

q_{10} : Inscriptions having any transcription, found in the city of 'London', and produced during the 'Trajan Government'.

Evaluation: The execution times of the queries in the catalogue are reported in Fig. 15. It is worth noting that most of the queries run in less than 1 s, and all queries run under 50 s, which can be considered to be quite fast, especially since we are integrating three different datasets. Three queries are the most challenging ones: q_2 , q_4 and q_{10} . To explain their relatively bad performance, let us consider q_4 . The generated SQL query returns +2000 answers, and the translation contains the union of 8 select-project-join queries. Each sub-query contains the join of 4–7 tables from the three different datasets (such as Inscriptions from EDH, Places from Pleiades, and InformationCarrier from EPNet). Moreover, for each inscription, the SQL performs complex string operations to check the occurrence of the text 'PNN'. In light of the complexity of the SQL query required to answer this information need, it is not surprising that it takes 47 s. On the other hand, q_1 , is translated into the union of 4 select-project-join queries, where each sub-query contains the join of 5–9 tables from the three different datasets.

5.5. SPARQL End-point

Besides our web-based platform, we also provide a standard SPARQL HTTP endpoint⁴⁴ (shown in Fig. 16), which can be used by other systems to query our integrated data. Other systems can use this URI (without "/summary") to send queries, and get in response the computed answers. More precisely, a SPARQL endpoint is a (referenceable) entity over which SPARQL queries can be posed and that is compliant with the SPARQL Protocol for RDF specification.

6. Conclusion

In this paper we have presented the knowledge representation effort within the EPNet project, which has led to the development of an ontology for the integration of three data sources adopted for historical research. The integration has been carried out by adopting the OBDI paradigm, and has been supported by the -ontop- system. This enabled us to deal in an efficient and sound way with data access, integration, and consistency issues. In order to develop an EPNet CRM and the associated ontology, we specialised existing standards in the domain of epigraphy. In this direction, we also relied on the EAGLE Metadata Model, not only because it is already based on solid modelling standards such as CIDOC CRM and EpiDoc, but also because it paves the way to the possibility of a future integration in the EAGLE federation of epigraphy databases.

In addition to easing data integration, the knowledge representation effort has also produced a formal, yet easy to read, specification in the form of an ORM conceptual schema, which facilitates communication between the different partners in the project (especially those not versed in computer science), and sets a common ground for discussing future research questions.

As further work, we are planning to extend our data integration effort to incorporate additional datasets. Also, we plan on improving our system's user interface by applying data visualization techniques, so users can explore and understand the data in a

more convenient way than the traditional tabular lists. Finally, we expect to make the URIs in our linked data point to this interface.

Acknowledgements

This research has been partially supported by the EU ERC Advanced Grant EPNet (*Production and distribution of food during the Roman Empire: Economics and Political Dynamics*, grant agreement ERC-2013-ADG 340828, and by the EU large-scale Integrating Project Optique (*Scalable End-user Access to Big Data*), grant agreement n. FP7-318338.

References

- Baader, F., Calvanese, D., McGuinness, D., Nardi, D., Patel-Schneider, P.F. (Eds.), 2007. *The Description Logic Handbook: Theory, Implementation and Applications*, 2nd Ed. Cambridge University Press, Cambridge, United Kingdom.
- Bishop, B., Kiryakov, A., Ognyanoff, D., Peikov, I., Tashev, Z., Velkov, R., 2011. OWLIM: a family of scalable semantic repositories. *Semantic Web J.* 2 (1), 33–42.
- Bouchou, B., Niang, C., 2014. Semantic mediator querying. In: *Proceedings of the 18th International Database Engineering & Applications Symposium (IDEAS)*, ACM Press, pp. 29–38. <http://dx.doi.org/10.1145/2628194.2628218>.
- Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R., 2007. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *J. Autom. Reason.* 39 (3), 385–429.
- Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rodríguez-Muro, M., Rosati, R., 2009a. Ontologies and databases: The *DL-Lite* approach. In: Tesararis, S., Franconi, E. (Eds.), *Reasoning Web. Semantic Technologies for Information Systems – 5th International Summer School Tutorial Lectures RW Lecture Notes in Computer Science*, vol. 5689. Springer, pp. 255–356.
- Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R., Ruzzi, M., 2009. Using OWL in data integration. In: *Semantic Web Information Management – A Model-Based Perspective*, Springer, 2009, pp. 397–424.
- Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R., 2013. Data complexity of query answering in description logics. *Artif. Intell.* 195, 335–360.
- Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodríguez-Muro, M., Xiao, G., 2016. Ontop: answering SPARQL queries over relational databases, submitted for publication.
- Calvanese, D., Giese, M., Hovland, D., Rezk, M., 2015. Ontology-based integration of cross-linked datasets. In: *Proceedings of the 14th International Semantic Web Conference (ISWC)*, vol. 9366 of *Lecture Notes in Computer Science*, Springer, pp. 199–216.
- Civili, C., Console, M., De Giacomo, G., Lembo, D., Lenzerini, M., Lepore, L., Mancini, R., Poggi, A., Rosati, R., Ruzzi, M., Santarelli, V., Savo, D.F., 2013. MASTRO STUDIO: managing ontology-based data access applications. *Proceedings of the VLDB Endow.* 6 (12), 1314–1317.
- Clayphan, R., 2012. Europeana data model. In: *DCMI-UK Seminar: Five Years On*.
- Crofts, N., Doerr, M., Gill, T., Stead, S., Stiff, M., Definition of the CIDOC conceptual reference model, ICOM/CIDOC Documentation Standards Group. CIDOC CRM Special Interest Group 5.
- Das, S., Sundara, S., Cyganiak, R., Sep. 2012. R2RML: RDB to RDF mapping language, W3C Recommendation, World Wide Web Consortium, available at (<http://www.w3.org/TR/r2rml/>).
- Domingue, J., Fensel, D., Hendler, J.A., 2011. *Handbook of Semantic Web Technologies*. Springer, Berlin, Germany.
- Elliott, T., Bodard, G., Milonas, E., Stoyanova, S., Tupman, C., Vanderbilt, S., 2007. EpiDoc guidelines: Ancient documents in TEI XML, available online at (<http://www.stoa.org/epidoc/latest/>).
- Epstein, J.M., 2008. Why model? *J. Artif. Soc. Soc. Simul.* 11 (4).
- Feraudi-Gruenais, F. (Ed.), 2010. *Latin on Stone: Epigraphic Research and Electronic Archives*. Lexington Books, Lanham, Maryland, United States.
- Franconi, E., Mosca, A., Solomakhin, D., 2012. ORM2: Formalisation and encoding in OWL 2. In: *On the Move to Meaningful Internet Systems: OTM 2012 Workshops*, vol. 7567 of *Lecture Notes in Computer Science*, Springer, pp. 368–378.
- Garnsey, P., Whittaker, C., 1983. *Trade and famine in classical antiquity, supplementary volume – Cambridge philological society*. Cambridge University Press, Cambridge, United Kingdom.
- Giese, M., Soylu, A., Vega-Gorgojo, G., Waaler, A., Haase, P., Jiménez-Ruiz, E., Lanti, D., Rezk, M., Xiao, G., Özçep, Ö.L., Rosati, R., 2015. Optique: Zooming in on big data. *IEEE Comput.* 48 (3), 60–67.
- Guldi, J., Armitage, D., 2014. *The History Manifesto*. Cambridge University Press, Cambridge, United Kingdom.
- Halpin, T., Morgan, T., 2010. *Information Modeling and Relational Databases*. Morgan Kaufmann, Burlington, Massachusetts, USA.
- Hitzler, P., Krötzsch, M., Rudolph, S., 2009. *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, United Kingdom.
- Kharlamov, E., Solomakhina, N., Özçep, Ö.L., Zheleznyakov, D., Hubauer, T., Lamarter, S., Roshchin, M., Soylu, A., Watson, S., 2014. How semantic technologies

⁴⁴ http://136.243.8.213:8080/openrdf-workbench/repositories/epnet_pleiades_edh/summary.

- can enhance data access at Siemens Energy. In: Proceedings of the 13th International Semantic Web Conference on (ISWC), vol. 8796 of Lecture Notes in Computer Science, Springer, pp. 601–619.
- Kharlamov, E., Hovland, D., Jimenez-Ruiz, E., Lanti, D., Lie, H., Pinkel, C., Rezk, M., Skjeveland, M.G., Thorstensen, E., Xiao, G., Zheleznyakov, D., Horrocks, I., 2015. Ontology based access to exploration data at Statoil. In: Proceedings of the 14th International Semantic Web Conference on (ISWC), vol. 9367 of Lecture Notes in Computer Science, Springer, pp. 93–112.
- Kilapi, H., Sakkos, P., Delis, A., Gunopulos, D., Ioannidis, Y.E., 2015. Elastic processing of analytical query workloads on IaaS clouds, CoRR Technical Report. [arxiv:1501.01070](http://arxiv.org/abs/1501.01070), arXiv.org e-Print archive, available at (<http://arxiv.org/abs/1501.01070>).
- Krötzsch, M., Simancik, F., Horrocks, I., 2012. A description logic primer, CoRR Technical Report. [arxiv:1201.4089](http://arxiv.org/abs/1201.4089), arXiv.org e-Print archive, available at (<http://arxiv.org/abs/1201.4089>).
- Lo Cascio, E., Rathbone, D.W., 2000. *Production and Public Powers in Classical Antiquity*, Supplementary volume. Cambridge Philological Society, Cambridge, United Kingdom.
- Lopez, V., Stephenson, M., Kotoulas, S., Tommasi, P., 2015. Data access linking and integration with DALI: Building a safety net for an ocean of city data. In: Proceedings of the 14th International Semantic Web Conference (ISWC), vol. 9367 of Lecture Notes in Computer Science, Springer, pp. 186–202.
- Merono-Penuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F., 2015. Semantic technologies for historical research: a survey. *Sem. Web J.* 6 (6), 539–564.
- Orlandi, S., Santucci, R., Casarosa, V., Liuzzo, P.M. (Eds.), 2014. *Information Technologies for Epigraphy and Cultural Heritage – Proc. of the First EAGLE International Conference*, Studi Umanistici, EAGLE BPN. Sapienza Università Editrice, Roma, Italia.
- Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R., 2008. Linking data to ontologies. *J Data Sem X*, 133–173.
- Raghavan, P., 2014. It is time to scale the science in the social sciences, *Big Data & Society* 1 (1).
- Rahimi, A., Liaw, S., Taggart, J., Ray, P., Yu, H., 2014. Validating an ontology-based algorithm to identify patients with Type 2 Diabetes Mellitus in electronic health records. *Int. J. Med. Inform.* 83 (10), 768–778.
- Remesal Rodriguez, J., 2002. *Baetica and Germania. Notes on the concept of provincial interdependence in the Roman Empire. The Roman Army and the Economy*. Amsterdam, 293–303.
- Remesal-Rodriguez, J., 2008. *Provincial interdependence in the Roman Empire: an explanatory model of Roman economy*. *Bar Int. Ser.* 1782, 155.
- Rodriguez-Muro, M., Rezk, M., Efficient SPARQL-to-SQL with R2RML mappings, *J. Web Semantics* Available online at (<http://www.sciencedirect.com/science/article/pii/S1570826815000153>).
- Rodriguez-Muro, M., Kontchakov, R., Zakharyashev, M., 2013. Ontology-based data access: Ontop of databases. In: Proceedings of the 12th International Semantic Web Conference (ISWC), vol. 8218, Springer, pp. 558–573.
- Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenko, B., van Ossendrup, J., Siebes, R., Taekema, J., Wielemaker, J., Wielinga, B., 2006. *Multi-mediaN e-culture demonstrator*. In: Proceedings of the 5th International Semantic Web Conference (ISWC), vol. 4273 of Lecture Notes in Computer Science, Springer, pp. 951–958.
- Sequeda, J.F., Arenas, M., Miranker, D.P., 2014. OBDA: Query rewriting or materialization? In practice, both!. In: Proceedings of the 13th International Semantic Web Conference on (ISWC), vol. 8796 of Lecture Notes in Computer Science, Springer, pp. 535–551.
- Shadbolt, N., Berners-Lee, T., Hall, W., 2006. *The Semantic Web revisited*. *IEEE Intell. Syst.* 21 (3), 96–101.