Diego Calvanese, Alessandro Mosca, Jose Remesal, Martin Rezk, and Guillem Rull

A 'Historical Case' of Ontology-Based Data Access

***Proc. of Digital Heritage 2015 (DH 2015)*. 2015. To appear.**

A 'Historical Case' of Ontology-Based Data Access

Diego Calvanese*, Alessandro Mosca†, José Remesal‡, Martin Rezk*, and Guillem Rull‡

*KRDB Research Centre, Free University of Bozen-Bolzano, Italy, Email calvanese,mrezk@inf.unibz.it
†SIRIS Lab, Research Division of SIRIS Academic, Spain, Email a.mosca@sirisacademic.com
‡CEIPAC, University of Barcelona, Spain, Email remesal,grull@ceipac.ub.edu

***Abstract*—Historical research has steadily been adopting semantic technologies to tackle several recent problems in the field, such as making explicit the semantics contained in the historical sources, formalising them and linking them. Over the last decades, in social sciences and humanities an immense amount of new quantifiable data have been accumulated and made available in interchangeable formats, opening up new possibilities for solving old questions and posing new ones. This paper introduces a web-based platform to ease the access of scholars to historical and cultural data distributed across different data sources. The approach relies on the Ontology-Based Data Access (OBDA) paradigm, where the different datasets are virtually integrated by a conceptual layer (an ontology). This work is focused on investigating the mechanisms and characteristics of the food production and commercial trade system during the Roman Empire.**

***Index Terms*—e-Culture, History of the Roman Empire economics, Ontology-Based Data Access, Cultural Data Integration, Linked Open Data, Web-based query/answering system.**

I. Introduction

Historical research has steadily been adopting semantic technologies [1], [2], [3] to tackle several recent problems in the field, such as making explicit the semantics contained in the historical sources, formalising them and linking them [4]. Historians, especially in Digital Humanities, are starting to use historical sources to aggregate information about history. Moreover, the recent advances in computing and computational tools (from machine learning, to applied mathematical statistics, text mining and topic-modelling algorithms, and semantic technologies) make it feasible to meaningfully manipulate, manage, and analyse these datasets. An outcome of this, is that 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 [5].

Since a sustainable maturity in the development of Semantic Web and Linked Open Data technologies has been reached—think, e.g., of data exchange protocols, standardised knowledge representation languages, and common data formats[1]—a considerable number of public initiatives and projects have been funded to address the issue of building historical and cultural data, and making it public through the web. Among others, the following 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: EUROPEANA[2], ARIADNE[3], CULTURESAMPO[4], STICH@CATCH[5], MultimediaN N9C[6], CHIP[7], EAGLE[8], CIDOC CRM[9], GETTY Vocabularies[10], INCONCLASS[11], EPIDOC[12].

These projects can be characterised by one of the following two goals: *(i)* to explicitly 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; and *(ii)* to represent implementations of the envisioned applications.

A shortcoming of the existing models developed by the projects in the first category 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').

The emphasis of EPNet is on providing historians with computational tools to compare, aggregate, measure, geolocalise, and search data about Latin and Greek inscriptions on amphoras for food transportation. This approach relies on the Ontology-Based Data Access (OBDA) paradigm, where the different datasets are virtually integrated by a conceptual layer (an ontology).

Example 1.1: Suppose the user needs all the amphoras produced in 'La Corregidora' and its geo-coordinates. The EPNet dataset contains information about amphoras and some (potentially incomplete) information about geo-coordinates. On the other hand, the Pleiades dataset (http://pleiades.stoa.org) contains more complete geo-coordinates information but has no information about amphoras. There are hundreds of types of amphoras such as Dressel 1, Dressel 2-4, Leptiminus 1, each of them represented by an alphanumeric-numeric code, such as "DR1C-BTIR" in EPNet. Thus creating a query for this simple information need is not only extremely

[1] W3C Standards, see http://www.w3.org/

[2] http://www.europeana.eu
[3] http://www.ariadne-infrastructure.eu
[4] http://www.kulttuurisampo.fi
[5] http://www.cs.vu.nl/STITCH
[6] http://e-culture.multimedian.nl
[7] http://chip.win.tue.nl
[8] http://http://eagle-network.eu
[9] http://www.cidoc-crm.org
[10] http://www.getty.edu/research/tools/vocabularies
[11] http://www.iconclass.nl
[12] http://epidoc.sourceforge.net

complex, but requires the user to know the DB encoding of each type, the schemas in the datasources, and manually merge the information obtained from each of them. Ideally the user should 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 both datasets.

Differently from providing access to virtual museums or digitalised collections, the OBDA implementation introduced in the paper, by means of state-of-the-art technologies and principles coming from the research area of Knowledge Representation [6], is meant to support scholars in *experimentally verifying* theoretical hypotheses, and in formulating new ones. Specifically, this paper provides the following contributions:

- The introduction of the EPNet Conceptual Reference Model.
- The specification of the relational schema which drove the deployment of the EPNet dataset.
- The EPNet ontology and mappings, and the ways they are used in the OBDA implemented system.
- The web-based implementation of the EPNet query/answering system.

The paper is organised as follows: Section II gives a brief introduction to the EPNet project. Section III concisely describes the different artefacts that we developed to build our data management system relying on ontologies. Section IV 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 is hyperlinked in the same section. Section V concludes the paper.

II. 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 interdependencies 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. The majority of them continue to be speculative and difficult to *falsify* [7], [8].

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 by *(i)* a large dataset of existing empirical data about Roman amphorae and their associated epigraphy that has been created during the last 2 decades (see, e.g., Fig. 1 and Fig. 2), and *(ii)* the front line theoretical research done by historians on the political and economic aspects of the Roman trade system.

The economy of the Roman Empire: an ongoing debate. A crucial aspect of any society is the production, supply and re-distribution of food. This topic has long been, and still remains, one of the open problems for sustainable decision

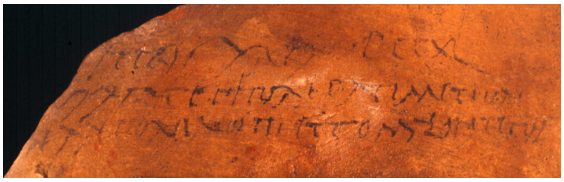

Fig. 1. *Titulus pictus* in 'delta' position over a Dressel 20 amphora

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, and *(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 or vice-versa, the scientific debate has focused mostly on the influence of the capital of the Empire (Rome) in the control and management of long distance trade, rather than on analysing 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 Fig. 1). 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 hypotheses and theories about the past explored (see [9]).

In this context, semantic-based technologies for data management, such as OBDA, can account for discrete data in addition to qualitative influences and interpretations, so as to answer broader questions about motives and patterns in the historical record. In particular, OBDA enables scholars to retrieve information stored in the EPNet dataset 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.

OBDA supports EPNet in facing the main challenge of providing users with: *(i)* a running technology for accessing data in a way that is conceptually sound with their own domain knowledge (see, the EPNet Conceptual Reference Model and the ontology introduced in the next section); *(ii)* a semantically-transparent platform, ready to acquire and be complemented with new data from different sources (domain-related historical datasets managed by research labs or promoted public); *(iii)* a theoretically grounded mechanism to ho-

Fig. 2. The result of a query over the stamp ACIRGI in the EPNet dataset

mogenise information stored in different formats and according to different conceptualisations (alternative representations of periods of time, for instance, or locations differently stored according to their ancient or modern name).

By means of the OBDA technology, extensive amounts of the EPNet data (see, e.g., Fig. 2) is to be connected and subsequently interpreted in a variety of levels that will give new insights to the complexity of the Roman Empire exchange relations. Moving beyond the limitations of a traditional relational DB is essential for the generation of new knowledge, and for the specification of values and parameters that will be manipulated in the simulation experiments.

III. Knowledge representation and data management in EPNet

In this section, we present the EPNet Conceptual Reference Model (CRM), the derived (logical) data model, and the EPNet ontology. The last part of this section is devoted to the introduction of the 'Pleiades' dataset whose data content has been integrated in the project dataset, thus *(i)* increasing the coverage of the data provided to the final users with respect to the domain of interest (completeness), and *(ii)* complementing the characterisation of the geographical entities already present in the initial dataset (accuracy).

The Conceptual Reference Model (CRM). The specification of the EPNet CRM for the representation of epigraphic information and domain expert knowledge about Roman Empire Latin inscriptions was meant to unambiguously represent the way the data are understood by scholars, how they are connected, and what their coverage is with respect to the literature of reference and 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), and by means of NORMA, a data modelling tool for ORM2[13]. Nonetheless, the CRM has been defined according to the state-of-the-art 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 CRM, and of the whole EPNet dataset, with other similar initiatives and data sources, the main section of the model results in a specialisation/extension of the well known CIDOC Conceptual Reference Model, the most dominant ontology in cultural heritage.

For the sake of model maintenance, and according to the specific nature of the involved information, the CRM has been structurally organised into distinct interrelated subsections. Moreover, according to the different aim of each subsection, we again relied on existing standards for recording and publishing information on the Semantic Web, such as FaBiO (the FRBR-aligned Bibliographic Ontology - http://vocab.ox.ac.uk/fabio) for the bibliographic references documenting the entities in the CRM. The following are the five main sections of the EPNet CRM:

Main deals with the representation of the main 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. 3).

Time offers a conceptual arrangement, driven by the experts, of the different modalities used to denote interval periods, dates, and punctual instants of time, w.r.t. the given research domain. As explained in Section IV-A, the different formats the domain experts are used to deal with temporal information have been homogenised in the implemented OBDA system, in order to maintain the epistemological flexibility they provide in looking for specific data, while keeping the possibility to interchange between them and translate one into another (e.g., to move from the string '`Trajan Government`' to the corresponding numerical time-span '`98-117`').

Space is meant to deal with information concerning space and geographical localisation of the entities in Main CRM. A heterogeneous set of entities in Main CRM brings a characterisation in terms of space, from finding activities involved in the discovery of an artefact, till the relative position of an inscription with respect to other stylistic and structural elements of an amphora. The Space section has been, for this reason, divided into two distinct subsections: *(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 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. The geographic part of the model, complemented by information coming from different sources (see final paragraph of this section), 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.

Documental is devoted to the representation of the bibliographic information documenting the entities of interest (e.g., conference and workshop papers, books, web portals and digital encyclopedia).

Upper Typing is simply a collection of all the taxonomic structures characterising the entities in the Main CRM. Having all the taxonomies collected in a single place makes their management and successive extension a lot easier also for scholars with no technical background.

The CRM model made of the five sections introduced above, besides being formally correct and consistent, is comprehensive enough to host all the information and knowledge elicited from the domain experts, and represents a definitive

[13] NORMA is an open source plug-in to Microsoft Visual Studio .NET, freely downloadable from http://www.ormfoundation.org/.

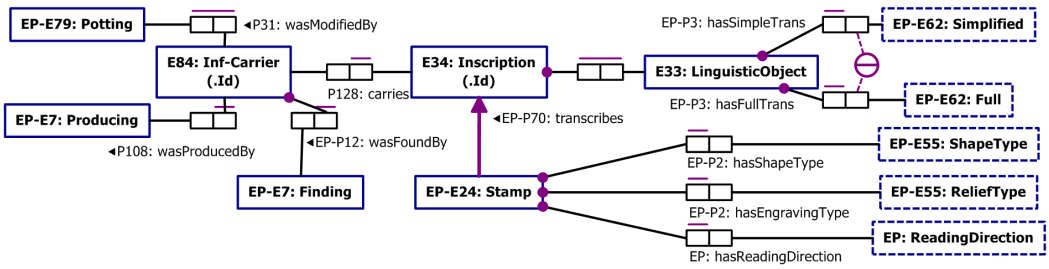

Fig. 3. A fragment of the EPNet CRM, where `Inscriptions` are related with the activities `Producing`, `Potting`, and `Finding`. `Stamps` are inscriptions characterised, among other, by their `Relief`, `Shape` type, and `ReadingDirection`. The model, written in ORM2, also shows that inscriptions are directly connected with 'simplified' and 'full' transcriptions, bringing information about their translation into contemporary languages, and their conservation status, respectively. The pink coloured symbols indicate cardinality constraints that have been superimposed to the schema, while the arrow stands for the usual *is-a* relation.

improvement in quality and granularity w.r.t. the previously adopted informal data structure descriptions we faced at the beginning of EPNet.

The EPNet dataset. While the CRM represents the knowledge of the domain, it does not specify how to store the actual data. Data storage greatly 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 our data, so we must provide a relational specification that complements our 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 on 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. 4. Tables are depicted as boxes, with their name at the top (e.g., `inscription`) and the list of attributes following (e.g., `id`, `carrier`). Each attribute consists of a name and a data type (e.g., `id INT(11)`, which indicates that the identifier of an inscriptions is an integer number). In particular, notice the tables `informationcarrier`, `amphoratype`, and `amphtyping`, 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 its producing and finding activities (detailed data about these activities is stored in separate tables); `amphoratype` records the information of each kind of amphora; and `amphtyping` links amphora identifiers with the corresponding type identifier(s) (could be more than one if the exact type of an amphora 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 (see for example the lines connecting `informationcarrier`, `amphtyping`, and `amphoratype`).

The EPNet ontology. In order to support the user with the possibility of accessing data through a domain-centred conceptual

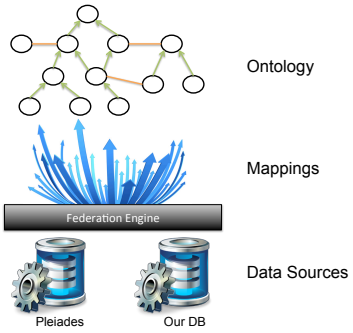

Fig. 5. EPNet system and Pleiades

layer and terminology, the relational specification introduced in the previous section has been encoded into an ontology. The resulting ontology, written in a formal language whose expressivity stays within the OWL 2 QL profile[14]), *modifies* and *extends* (by means of suitable concept hierarchies, see Example 4.2) 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, at the same time, the available data and the user requirements in terms of data accessibility and usage.

In the majority of the current projects in cultural heritage and humanities dealing with semantic technologies, the conceptualisation of the domain is expected to expose data structures suitable for a generic audience (from tourists visiting a museum or searching on the Web their favourite piece of art, till public administrations willing to open up their cultural resources and historic properties). Instead, the EPNet ontology has 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 [10]. 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* and (binary) *relations* the experts

[14] http://www.w3.org/TR/owl2-profiles/

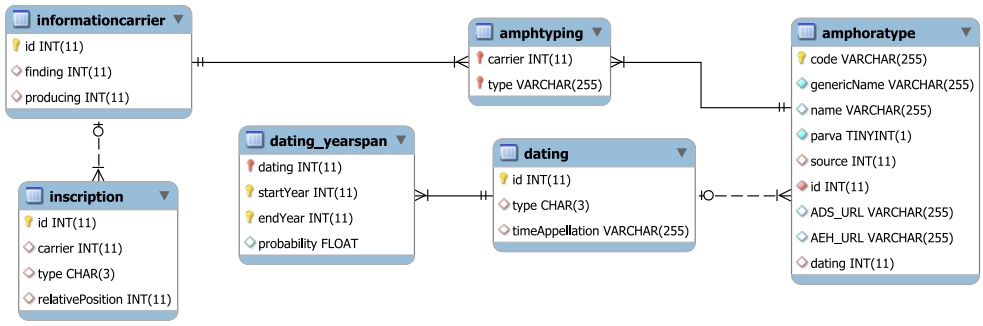

Fig. 4. A fragment of the relational specification of our database

make use of in conceptually classifying the entities of their research domain. As an example[15], consider the following axioms:

```
          :Stamp rdfs:subClassOf  :Inscription .
:TitulusPictus rdfs:subClassOf  :Inscription .
       :Amphora rdfs:subClassOf  :InfCarrier .
     :carriedBy rdfs:domain  :Inscription .
     :carriedBy rdfs:range  :InfCarrier .
    :producedAt rdfs:domain  :InfCarrier .
    :producedAt rdfs:range  :TimeSpan .
       :hasName rdf:type  owl:DatatypeProperty .
```

They say that the concepts `:Stamp` and `:TitulusPictus` are both subconcepts of `:Inscription` (see also Fig. 3), while `:Amphora` is a specialisation of `:InfCarrier`. The `:carriedBy` relation links inscriptions with their informational carrier and, similarly, the domain and range of the `:producedAt` relation are specified to be the `:InfCarrier` concept and the `:TimeSpan` in which the existence of the carrier is historically attested. The last axiom is for characterising `:hasName` as a datatype property, i.e., a property whose range is a specific datatype (`:String` in this case).

In addition, in order to expose the user to a domain-oriented vocabulary, special axioms have been added to the ontology. For instance, the following axiom introduces a new relation in the ontology by saying that `engravedOn` generalises the `carriedBy` relation between inscriptions and their informational carriers:

```
:carriedBy rdfs:subObjectPropertyOf  :engravedOn .
```

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

Pleiades. Pleiades[16] 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' [11], Pleiades is co-organised by the Institute for the Study of the Ancient World (NYU) 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 dataset has been selected in order to complement the EPNet dataset. In particular, it provides a number of geographic entities that is strictly greater than those present in the project dataset (e.g., specific municipalities and Roman provinces are present in EPNet only if they are a finding, producing, or potting place). The integration with Pleiades supports EPNet in tracing trade routes and economic connections on the Roman Empire territory in a more precise way and over a satisfactory picture of the past anthropic environment. If a location is present in both the Pleiades and the project DB but missing some attributes in the latter (e.g., the place has no geo-coordinates), the system is able 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.

IV. OBDA IN EPNET

Since the mid 2000s, *Ontology-Based Data Access* (OBDA) has become a popular approach to tackle the problems mentioned in Section II. An overall architecture of the EPNet OBDA setting is shown in Fig. 5. In OBDA, a conceptual layer is given in the form of an ontology that defines a shared vocabulary, models the domain, hides the structure of the data sources, and can enrich incomplete data with background knowledge. In our setting, the ontology is the one presented in Section III. Then, queries are posed over this high-level conceptual view, and the users no longer need an understanding of the data sources, the relation between them, or the encoding of the data. Queries are translated by the OBDA system into queries over the data sources.

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. Intuitively, the mappings expose the data in the database as Resource Description Framework (RDF) triples. RDF is a World Wide Web Consortium (W3C) specification for data interchange on the Web. This standard is based

[15] 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.

[16] http://pleiades.stoa.org

upon the idea of making statements about resources in the form of *subject–predicate–object* expressions. These expressions are known as *triples* in RDF terminology. Example of such triples are

```
http://epnet-url.org/1 rdf:type :Amphora
http://epnet-url.org/1 :producedIn
                    http://epnet-url.org/place/5
```

respectively stating that the element represented by the URI http://epnet-url.org/1 is an amphora, and that it was produced in the place represented by the URI http://epnet-url.org/place/5.

Intuitively, each of the mapping assertions that generates these triples (in OBDA) consist of a *source*, which is an SQL query retrieving values from the database, and a *target*, defining RDF triples with values from the source.

$$\underbrace{\textit{(subject predicate object)}}_{\text{target triple}} \longleftarrow \underbrace{\textit{SQL Statement}}_{\text{source query}}$$

Subjects and objects in RDF triples are resources (individuals or values) represented by URIs or literals. They are generated using *templates* in the mappings. For instance, the URI template `:Amphora-{ic_id}`, where `ic_id` is an attribute in some DB table, generates the URI `:Amphora-1>` when `ic_id` is instantiated to '1'. In addition, the colon symbol ':' represents the default URI string, in this example `http://epnet-url.org`, hence the generated URI is actually `http://epnet-url.org/Amphora-1`. Let us illustrate this with a further example.

Example 4.1 (EPNet Mappings): The following mapping populates the class `:Dressel1` (which is a subclass of `:Amphora`):

```
:Amphora-{ic_id} rdf:type :Dressel1    ⟵
    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='DR1'
```

Observe that this is a rather complex SQL query that joins information from three different tables. This complexity is hidden to the users by the simple concept `:Dressel1`. ∎

The ontology together with the mappings and the database exposes a *virtual* RDF graph, which can be queried using *SPARQL*, the standard query language in the Semantic Web community.

Example 4.2: Assume that the users need all the amphoras produced in "La Corregidora" and its geo-coordinates. This can be translated to the following SPARQL query using the vocabulary from the ontology.

```
SELECT * WHERE {
    ?x rdf:type :Amphora .
    ?x :producedIn ?pl .
    ?pl rdf:type :Place .
    ?pl :hasName "La Corregidora" .
    ?pl :hasLatitude ?lat .
    ?pl :hasLongitude ?long
}
```

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 and Pleiades.

There are several OBDA systems in both, academia and industry [12], [13], [14], [15]. We work with Ontop [12], [16], [17], [18], a mature open-source system, which is currently being used in a number of projects. Ontop allows the users to *materialize* virtual RDF graphs, generating RDF triples that can be used with RDF triplestores, or alternatively the graphs can be kept *virtual* and queried only during query execution. The virtual approach avoids the cost of materialization and can profit from more than 30 years of maturity of relational systems (efficient query answering, security, robust transaction support, etc.). To answer queries in the virtual approach by exploiting the information given by the ontology, Ontop relies on *query-rewriting*. To illustrate this let us come back to Example 4.2. When the user queries the class `:Amphora`, Ontop uses the ontology to infer that all the elements that belong to one of the subclasses (e.g., `:Dressel1`) also belong to the class `:Amphora`. Intuitively, Ontop rewrites the query in Example 4.2 creating a union for each subclass of `:Amphora`:

```
SELECT * WHERE {
  { ?x rdf:type :Amphora .
    ?x :producedIn ?pl .
    ?pl rdf:type :Place.
    ?pl :hasName "La Corregidora".
    ?pl :hasLatitude ?lat.
    ?pl :hasLongitude ?long .
  } UNION {
    ?x rdf:type :Dressel1 .
    ?x :producedIn ?pl .
    ?pl rdf:type :Place.
    ?pl :hasName "La Corregidora".
    ?pl :hasLatitude ?lat.
    ?pl :hasLongitude ?long .
  } UNION {
    ?x rdf:type :Leptiminus1 .
    ...
  }
}
```

Ontop is available as a Protégé 4 plugin, a SPARQL endpoint through Sesame Workbench, and a Java library supporting OWL API and Sesame API.

A. EPNet Data Integration

Ontop allows for virtual data integration. In this approach the data remain in the sources and are accessed at query time. Ontop does not modify the underlying databases, which is a requisite in this use case, neither does it require complex extract, transform, load processes. The classes and properties in the ontology, cluster different fragments of the databases into a homogenized well defined set of triples.

Ontop does not integrate the databases at the SQL level. For that it relies on a standard federation engine such as Teiid[17] or Exareme [19]. Any of these engines will expose a set of schemas containing the tables from each of these datasources. Ontop does the semantic integration and homogenization over these federated databases. Here we will discuss the integration of EPNet and Pleiades focusing on space and time periods. The integration starts in the ontology, where concepts cover

[17] teiid.jboss.org/

information contained in both datasets. The :Place concept, for instance, is characterised in the ontology by having a given function (e.g., :ProductionPlace, :CivilSettlement, :LegionaryCamp), it is linked through :hasLatitude and :hasLongitude relations to its geo-coordinates, and :fallsWithin or :isContainedIn other known places. Then the information from both datasets get connected through properties. In our running example we find:

- :producedIn, connecting amphoras in EPNet and places in EPNet and Pleiades, and
- :hasLatitude, connecting places in EPNet and Pleiades with latitude coordinates in both datasets.

Space. Both EPNet and Pleiades have information regarding places, settlements, geo-coordinates, etc. However, Pleiades is more complete space-wise, moreover it contains a kind of settlement that is missing in EPNet. If a place is not in the EPNet dataset, we completely rely on the data from Pleiades (name(s), geo coordinates, and kind of settlement). If the place is in EPNet (with comparison done by name), then we keep the existing EPNet data, and add the kind of settlement (which is not in EPNet). Moreover, if the existing data is incomplete (e.g., missing coordinates), we fill it with Pleiades data.

To cluster all the information about places in both dataset into a single well-defined concept :Place we use mappings. Here we present a simplified version of the mappings for the sake of presentation:

Pleiades:

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

EPNet:

```
:Place-{gl_id} rdf:type :Place  ⟵
    SELECT gl.id AS gl_id
      FROM GeographicLocation gl
```

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

Time. Regarding the time periods, EPNet and Pleiades specify time periods using list of integers, for instance: $[(98, 117), (130, 140)]$ to state that an object or a place existed either in the period 98 AD – 117 AD, or 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 this, 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 any of these 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 (e.g., xsd:gYear, xsd:dateTime, xsd:period). However, adding them would simply require the small effort of adding a few mappings.

B. Ontop Data Consistency

A logic based ontology language, such as OWL, allows ontologies to be specified as logical theories, this implies that it is possible to constrain the relationships between concepts, properties, and data. In OBDA, inconsistencies arise when the data in the sources together with the mappings violate the constraints imposed by the ontology, and it is of interest to check whether such violations occur. The following are some important types of constraints:

- *Disjointness*, stating that the intersection between two classes or between two properties should be empty. For instance, the classes :MilitarCamp and CivilSettlement must not have elements in common.
- *Functionality of properties*, stating that no individual can be related to more than one element through a functional property. For instance, the property :hasShape is functional since every amphora must have a unique shape.

Notice that disjointness can be expressed in the OWL 2 QL profile of OWL 2, while functionality cannot. However, both types of constraints can be checked by Ontop by posing suitable queries over the ontology, and checking whether the answer to such queries is non-empty.

C. User Interface

A preliminary user interface for testing the OBDA functionalities in EPNet is available online[18]. It provides users with a text area where to write SPARQL queries (e.g., the query in Example 4.2) using the vocabulary of the ontology discussed in Section III (for the convenience of the user, a summary of the ontology is provided by the interface; see Fig. 6). Following SPARQL syntax, users need to begin their queries with a prefix declaration, which in our case is:

```
PREFIX : <http://136.243.8.213/obdasystem#>
PREFIX rdf:
       <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
```

After executing the query, the interface shows the SQL query that was sent to the underlying RDBMS (Fig. 7), and the result of the query in tabular form (Fig. 8).

[18] 136.243.8.213/obdasystem

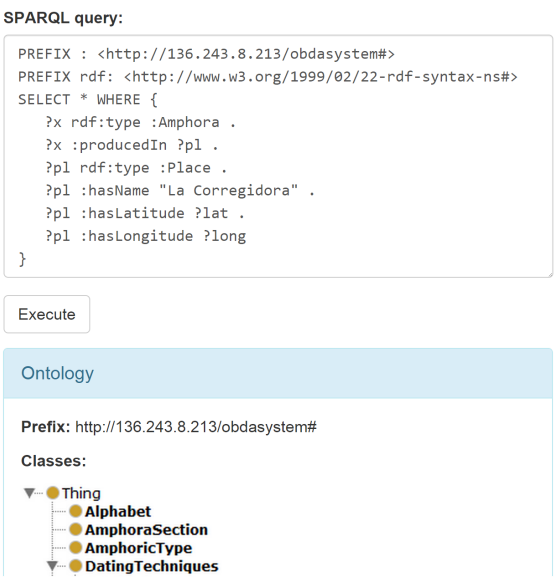

Fig. 6. Screenshot of a user's query in the OBDA web interface

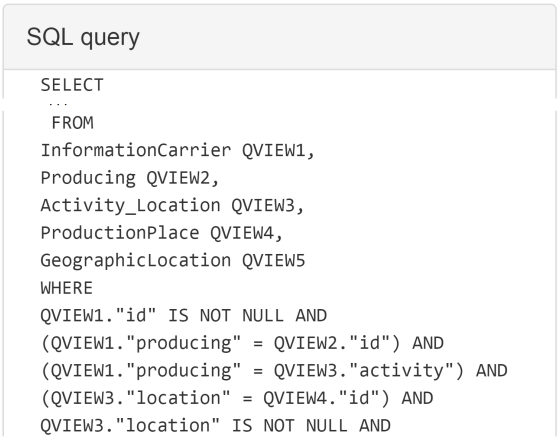

Fig. 7. SQL query that is actually executed on the EPNet dataset

V. Concluding Remarks and Future Work

This paper presents the design and implementation of the OBDA approach in the context of the EPNet project. The OBDA technology helped us to deal in an efficient and sound way with data access, integration, and consistency issues. The integration with a greater number of available datasets, from different scholars and research initiatives, has been already planned. EPNet will also explore the application of text mining techniques to automatically extract information from the epigraphic corpus (e.g., person names, professions, places), thus going beyond the 'syntactical' descriptions of the conservation status of the inscriptions themselves, and fruitfully complementing the information already present in the project dataset.

Acknowledgments

The work is partially funded by the ERC Advanced Grant n. ERC-2013-ADG 340828 and the EU under the large-scale integrating project (IP) Optique (Scalable End-user Access to Big Data), grant n. FP7-318338.

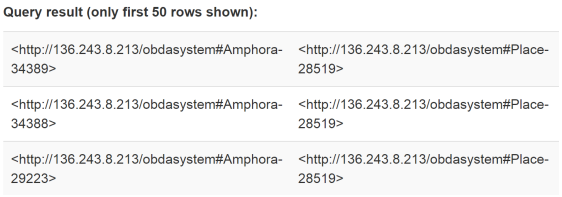

Fig. 8. Results of the user's query execution.

References

[1] P. Hitzler, M. Krötzsch, and S. Rudolph, *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, 2009.

[2] N. Shadbolt, T. Berners-Lee, and W. Hall, "The semantic web revisited," *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, 2006.

[3] J. Domingue, D. Fensel, and J. A. Hendler, *Handbook of Semantic Web Technologies*, ser. Handbook of Semantic Web Technologies. Springer, 2011, no. Volumes 1–2.

[4] A. Merono-Penuela, A. Ashkpour, M. van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, and F. van Harmelen, "Semantic technologies for historical research: A survey," *Semantic Web J.*, pp. 1–26, 2014.

[5] P. Raghavan, "It's time to scale the science in the social sciences," *Big Data & Society*, vol. 1, no. 1, 2014.

[6] F. van Harmelen, V. Lifschitz, and B. Porter, *Handbook of Knowledge Representation*. Elsevier, 2007.

[7] P. Garnsey and C. Whittaker, *Trade and Famine in Classical Antiquity*, ser. Supplementary volume - Cambridge Philological Society. Cambridge University Press, 1983.

[8] E. Cascio and D. Rathbone, *Production and Public Powers in Classical Antiquity*, ser. Supplementary volume - Cambridge Philological Society. Cambridge Philological Society, 2000.

[9] J. M. Epstein, "Why model?" *J. of Artificial Societies and Social Simulation*, vol. 11, no. 4, 2008.

[10] J. Guldi and D. Armitage, *The History Manifesto*. Cambridge University Press, 2014.

[11] R. J. A. Talbert, Ed., *Barrington Atlas of the Greek and Roman World*. Princeton University Press, 2000.

[12] M. Rodriguez-Muro and M. Rezk, "Efficient SPARQL-to-SQL with R2RML mappings," *Journal of Web Semantics*, 2015.

[13] B. Bishop, A. Kiryakov, D. Ognyanoff, I. Peikov, Z. Tashev, and R. Velkov, "OWLIM: A family of scalable semantic repositories," *Semantic Web J.*, vol. 2, no. 1, pp. 33–42, 2011.

[14] J. F. Sequeda, M. Arenas, and D. P. Miranker, "OBDA: query rewriting or materialization? In practice, both!" in *Proc. of the 13th Int. Semantic Web Conf. (ISWC)*, vol. 8796. Springer, 2014, pp. 535–551.

[15] C. Civili, M. Console, G. De Giacomo, D. Lembo, M. Lenzerini, L. Lepore, R. Mancini, A. Poggi, R. Rosati, M. Ruzzi, V. Santarelli, and D. F. Savo, "MASTRO STUDIO: managing ontology-based data access applications," *Proc. of the VLDB Endowment*, vol. 6, no. 12, pp. 1314–1317, 2013.

[16] G. Xiao, M. Rezk, M. Rodriguez-Muro, and D. Calvanese, "Rules and ontology based data access," in *Proc. of the 8th Int. Conf. on Web Reasoning and Rule Systems (RR)*, ser. Lecture Notes in Computer Science, vol. 8741. Springer, 2014, pp. 157–172.

[17] R. Kontchakov, M. Rezk, M. Rodriguez-Muro, G. Xiao, and M. Zakharyaschev, "Answering SPARQL queries over databases under OWL 2 QL entailment regime," in *Proc. of the 13th Int. Semantic Web Conf. (ISWC)*, ser. Lecture Notes in Computer Science, vol. 8796. Springer, 2014, pp. 552–567.

[18] D. Calvanese, M. Giese, D. Hovland, and M. Rezk, "Ontology-based integration of cross-linked datasets," in *Proc. of the 14th Int. Semantic Web Conf. (ISWC)*. Springer, 2015.

[19] H. Kllapi, P. Sakkos, A. Delis, D. Gunopulos, and Y. E. Ioannidis, "Elastic processing of analytical query workloads on IaaS clouds," arXiv.org e-Print archive, CoRR Technical Report abs/1501.01070, 2015. [Online]. Available: http://arxiv.org/abs/1501.01070