Roman Open Data: A semantic based Data Visualization & Exploratory Interface

Xavier Gimenez
SIRIS Lab, Research Division of SIRIS Academic

Alessandro Mosca
Free University of Bolzano

Bernardo Rondelli
SIRIS Lab, Research Division of SIRIS Academic

Guillem Rull
SIRIS Lab, Research Division of SIRIS Academic

1. From observation to explanation

Ancient History is based on abductive reasoning: it starts with a set of observations then seeks to find the most likely explanation for the observations. This process, unlike deductive reasoning, yields plausible conclusions but does not always verify it. Abductive conclusions can generate uncertainty or doubt, which is expressed in retreat terms such as "best available" or "most likely" explanation. In this framework, Ancient History strongly depends on the capacity of exploring observations, building the explanation and justifying it as the "most likely".

The EPNet project (Remesal et al. 2014, Remesal and Rondelli 2017) was born exactly to overtake this boundary. The main goal of EPNet was the use of formal tools for falsifying existing hypotheses concerning the roman economy. Over the last couple of centuries, several historians have developed a variety of hypotheses to explain the organization of the Roman Empire trade system. These hypotheses continue to be based on "the most likely" approach and are difficult to falsify.

Roman amphorae can be considered one of the best archaeological markers of Roman Empire economic production and transactions. They are a unique indicator of social relationships and chronological dynamics, because of their distribution throughout the Empire, the precise information contained in their epigraphy (similar to modern trademarks and labels) and their diachronic persistence.

Here we describe our contribution, within the EPNet Project, in providing historians with computational tools to compare, aggregate, measure, geo-localise, and search data about Latin inscriptions on amphoras for food transportation.

2. Open data, semantics and data visualization

At the end of the eighties, the CEIPAC research group began compiling a database of amphorae epigraphy. The database is based on a systematic analytical approach to the epigraphy of amphorae, initiated by H. Dressel, neglected for a long time, and then updated and improved first by E. Rodríguez Almeida and later by the CEIPAC group. This database currently contains more than 45,000 entries and represents an extraordinary source for providing observations about roman empire and supporting the generation of historical hypotheses.

A specific research activity within the EPNet project has been dedicated to the development of an open and semantic based new infrastructure of the CEIPAC database, integrating a series of other existing datasets (ADS – Roman Amphorae, Epigraphic database of Heidelberg, Pleiades graph of ancient places) using the ontology-based data access and integration approach (Calvanese et al. 2015, Mosca et al. 2015). A Data Visualization & Exploratory Interface has been then developed to explore in a more powerful and intuitive way all the data integrated.

We developed an ontology that models the domain of Roman amphora epigraphy. This ontology is defined according to the state-of-the-art formal ontological models and standards for representing cultural heritage objects; in particular, it is a specialization/extension of the CIDOC CRM[1] . Other standards that have been taken into account and served as inspiration are: FaBiO[2] for the bibliographic references, the Europeana Data Model[3], and the EAGLE Metadata Model and domain-centered vocabularies[4]. This knowledge representation effort addresses two main problems. Firstly, the ontology provides a formally defined, unambiguous, framework for exporting the data in a way that can be manipulated by computer-based simulations and complex-network analysis tools (*see the other papers in this volume*). And secondly, it allows the publication of the data on the web

---

1 http://www.cidoc-crm.org/
2 http://vocab.ox.ac.uk/fabio
3 http://pro.europeana.eu/edm-documentation
4 http://www.eagle-network.eu/about/documents-deliverables/

in an interoperable, standard-compliant way, following the Linked Data approach, which makes the project's data integrable with other data sources.

All the documentation for the ontology developed can be found in http://romanopendata.eu/sparql/doc. A technical and exhaustive description of the approach can be found in (Calvanese et al. 2015).

2.1 Data integration & management

We integrated the CEIPAC database with three complementary datasets:

- the ADS (Archeology Data Service) Roman Amphorae[5]
- the Epigraphic Database Heidelberg[6]
- the Pleiades gazetteer of ancient places[7]

The integrated system follows the Ontology-Based Data Access (OBDA) approach, in which the integrated data sources are mapped into a common domain ontology, so users can pose their queries on the system in terms of the homogeneous vocabulary defined by the ontology. In this way, users can access the data from all the integrated datasets without having to know the individual structure of each of them. To the users, all the data is structured according to the domain ontology of the integrated system. Data access is done via the query language SPARQL, a standard defined by the W3C, and one of the core components of the Linked Data framework. A SPARQL endpoint for the project is available at http://romanopendata.eu/sparql.

To facilitate the addition of new data to the system, a web-based data entry interface has been designed and is available at http://romanopendata.eu/dataentry. Authorized users can login into this web application, fill the corresponding forms with the new epigraphic data, and submit it for review. An administrator user that is also an expert in the domain will later verify the quality of the data and, if it qualifies, accept it for publication. New records are added as part of the CEIPAC dataset for Roman amphora epigraphy.

2.2 Data visualization

Data visualisations and graphics are fundamental to studying complex subject matter. However, beyond acknowledging this value, scientists rarely consider how visualisations can enable discovery, create engaging and robust reporting, or support online resources. Producing accessible and unbiased visualisations from complicated, uncertain data requires expertise and knowledge from science, computing, and design (McInerny, 2014), however it represents a powerful solution for supporting scientific research both in terms of data explorations and argumentation. In the last years the use of data visualizations to support scientific research has increased, as well as the literature that emphasizes the research about how people perceive, read (and misread) and interprets different kinds of visualizations and which types of plots are most effective and easiest to decipher. Nowadays, there is a good amount of literature depicting best practices about how to create effective visualization ('Visualization Analysis & Design', Tamara Munzner 2014). On this domain, and as a result of the growth of influence that Data Visualization is acquiring, it is noteworthy to mention some examples,

---

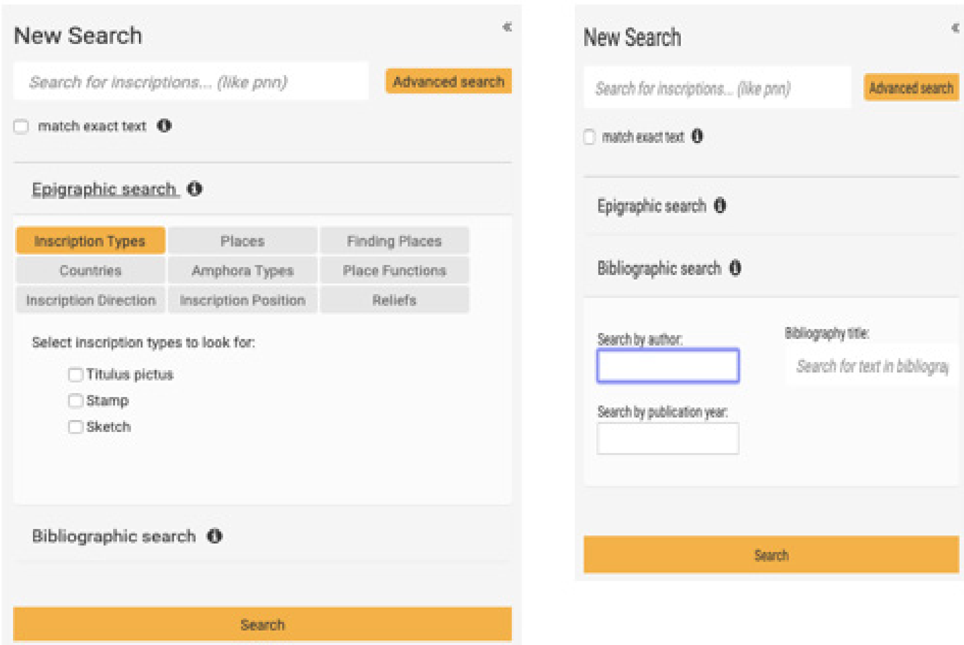
[5] http://archaeologydataservice.ac.uk/archives/view/amphora_ahrb_2005/
[6] https://edh-www.adw.uni-heidelberg.de
[7] https://pleiades.stoa.org/

such as the publication of The Atlas of Science: Visualizing what we know (Porter and Börner 2011), which offers a compendium of the evolution of science through the use of visual maps.

In the context of the EPNet project, in order to facilitate the exploration and navigation of the integrated data for users that are not proficient with the SPARQL query language, we have developed an interactive, online data visualization tool, which is available at http://romanopendata.eu.

This tool innovates by coupling the semantic web approach with data visualization technologies, for supporting data-driven explorations and pattern detections. This means that all data shown in the tool is obtained in real time by querying the integrated system via the provided SPARQL endpoint. The generated queries depend on the user's selections within the tool: users are indirectly making their own queries by clicking on items and selecting filters, without having to be aware of the SPARQL query language at all.

In order to offer the maximum flexibility in terms of search capabilities, the user interface offers a set of filtering options that covers most of the relevant attributes that characterizes the data. The filters are divided into two categories (epigraphic and bibliographic attributes) that can be combined when performing and advanced search, as show in the figure below:

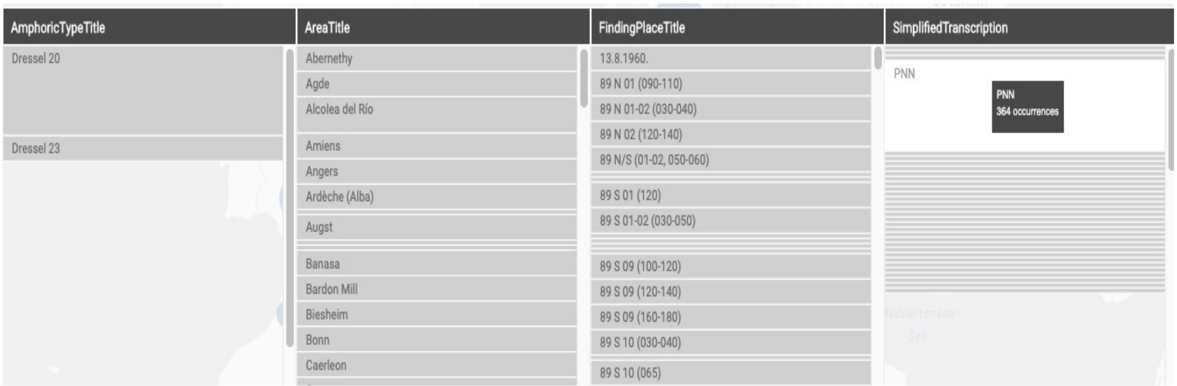


The interface relies on some well-know foundational visualization design patterns, by offering an interconnected set of views (linking and brushing paradigm), each of which facets the data in different dimensions, and at the same time, all reacting to possible filter or brushing actions done by users from any view.

The tool also promotes 'progressive disclosure', by showing first aggregated views of the data and then providing a clear path to unveiling details, accessible on demand:

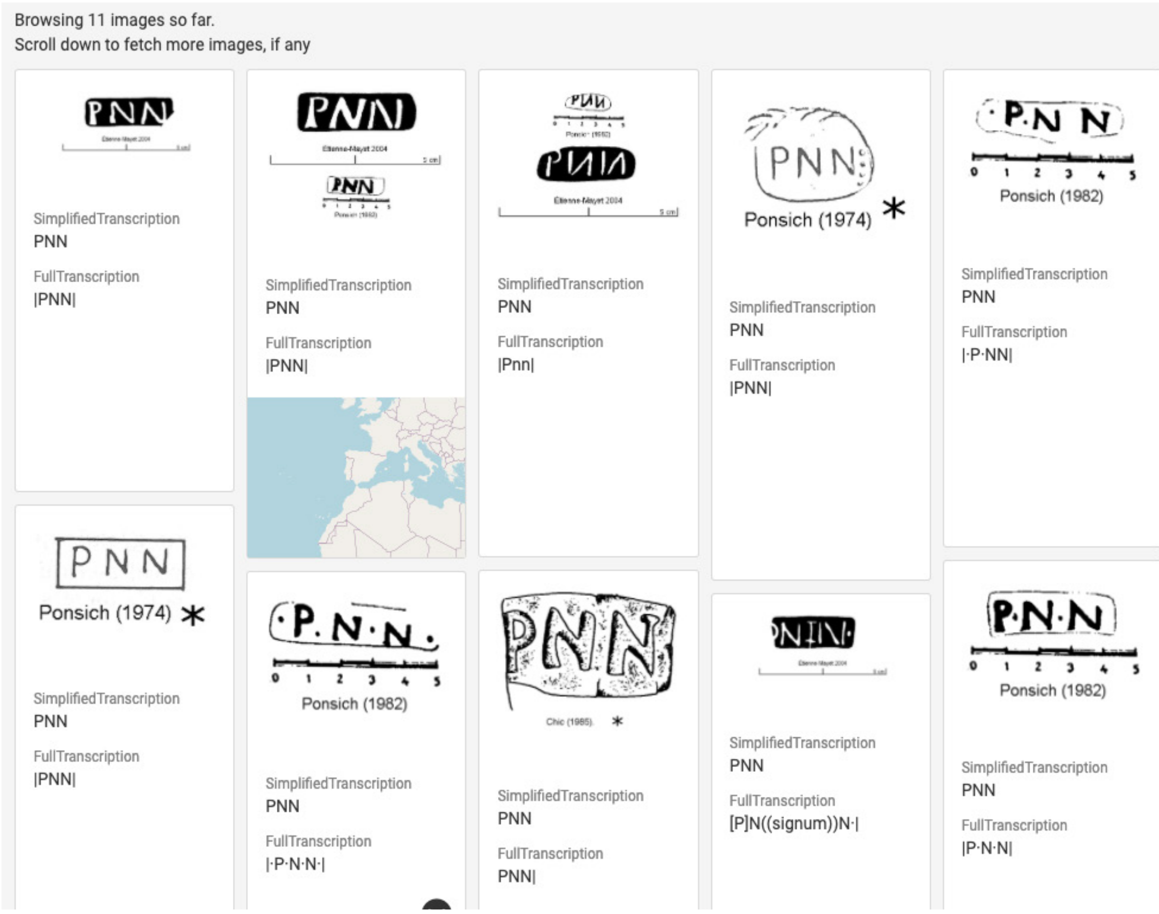
Example of the user interface showing aggregated data from 3 different query results, and then providing affordances to drill down to more granular data.

The tool is equipped also with more advanced filtering systems, like Elastic lists, which offers mechanisms for browsing multi-facetted data, by enabling the selection of differents attributes on the data and generating combinations of attributes that shape the data according the selection.



Example of an elastic list showing the distribution of occurrences by 4 differents attributes and enabling combinations of selection in order to filter the data by multiple criteria.

Access to individual datum is always available from the different existing views (geographically, timeline, and tabular representations), as shown below:

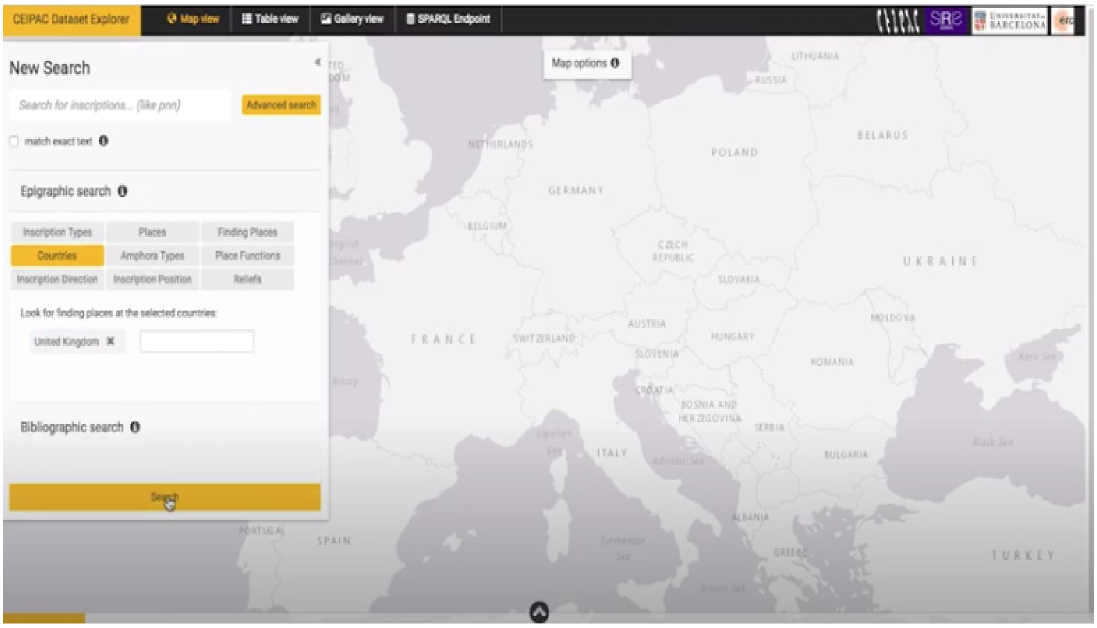


In order to facilitate the use of the visualization tool, a series of demos and tutorials have been realized and accessible from the home page.

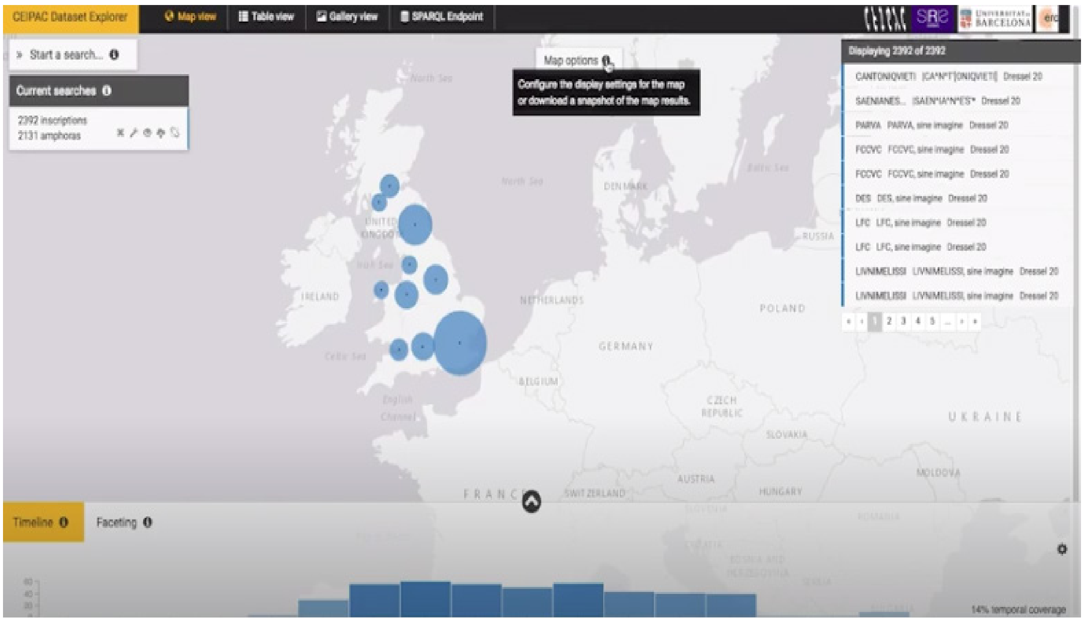
A use-case example:

To demonstrate some of the capabilities of the tool, here the series of actions performed by a user aiming at downloading data from “*searching amphoras nearby the Hadrian’s Wall”:*

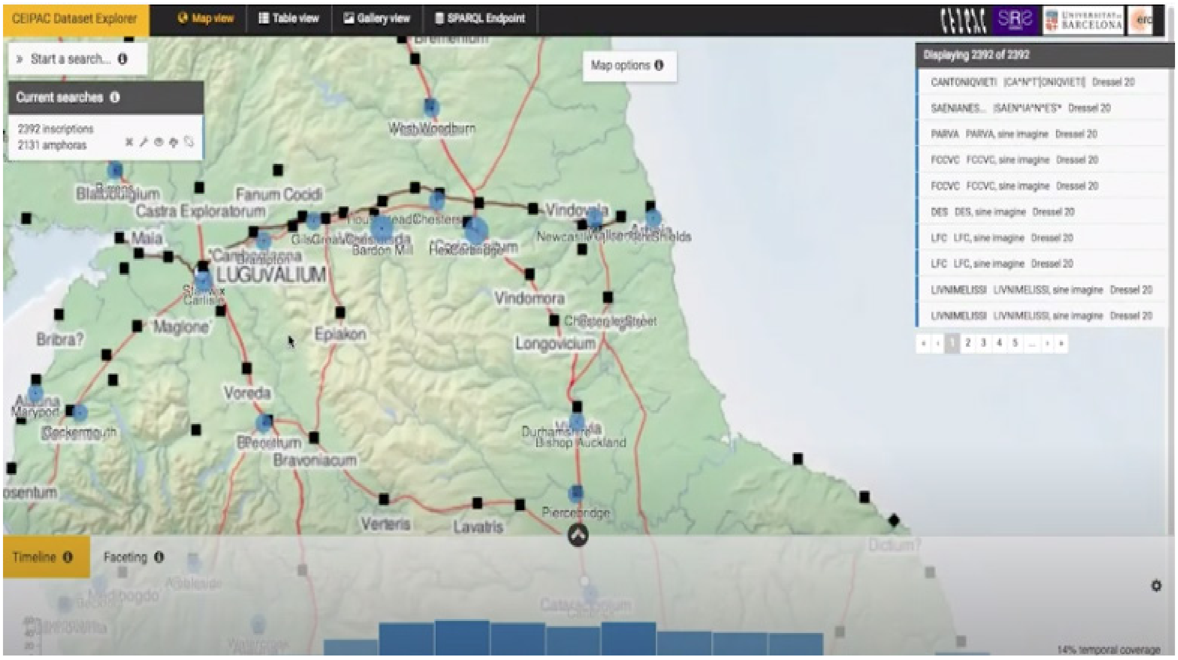
1. By using the advanced search, filter results only for amphoras found within United Kingdom:



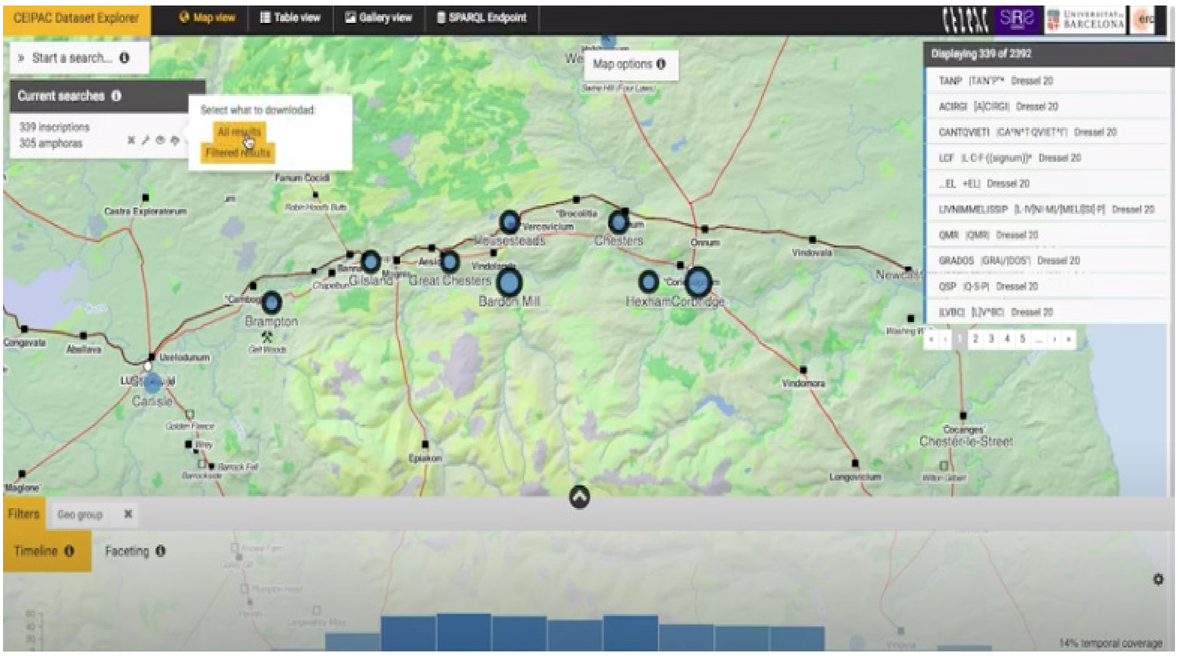
2 . Once the results are obtained, the user can change the type of map by selecting the roman Empire ancient map.



3. The user performs zooming and panning option to frame the area of interest (North of Britain)



4. After selecting places of interest around the Hadrian's Wall (MaryPort, Carlisle, Bardon Mill (Vindolanda), Corbridge, ...), the user can download the data filtered by the geographical selection performed by the user.



3. From technology change to research-culture change

Over the last ten years, the amount of data available in the so-called linked data cloud has continued to grow very rapidly. Although a huge amount of data is now also available in the cultural heritage and digital humanities sectors, there are still few examples of reuse of such data, especially in the field of ancient history and archaeology. In other words, there exists an increasing number of research projects working for developing innovative technological solutions for data integration and access, however few scholars and research groups within the humanities are really using them.

This shortage requires more attention and understanding in order to take advantage from the research efforts realised in the last years. For fully exploiting the enormous potential offered by semantic technologies and linked data it will be necessary to face the challenge of contributing not only to a technological change but also to a new academic and scientific culture: a culture that needs to prioritize the processes of data management and data sharing. In sectors such as digital humanities, still dominated by a strong tradition from the two-dimensional world of print, facing these challenges will be more difficult and more urgent than in other sectors.

In this perspective, the first lesson learnt from the EPNet is that publishing data in open format, using the most advanced semantic technologies is not enough if you don't train scholars and students in a) learning how to use these technologies and b) how to contribute to publishing their data in the same format.

The second lesson is the urgent need to change from a culture of controlling data for individual benefit to the capacity of linking and contributing to data sharing and integration. The questionable practice of publishing the explanation before the access to the full datasets of observations is an issue that the humanities in general, and ancient history in particular, should seriously deal with. The attitude of not publishing full datasets for not creating the conditions (full access to the data) for someone else to be able to propose a better or before explanation should be strongly fought in the academic environment, as it happens in other scientific disciplines.

The tension between promoting scientific goals in spite of protecting the privacy interests of study participants should be overcome, as this happened and is happening in other scientific fields during the last decades. Science is becoming more and more collaborative and without embracing the open science paradigm, humanities and ancient history will be even more relegated to the periphery.

Projects like EPNet demonstrate that it is much easier to dedicate effort to improve technological solutions (e.g from a classical relational database to a fully integrated semantic platform), but definitely harder changing the research culture in which these solutions are built and exploited. The challenge is how to facilitate this research-culture change and for this reason future investments should be made in that direction.

References

Calvanese, D., Liuzzo, P., Mosca, A., Remesal, J., Rezk, M. and Rull, G., 2016. Ontology-based data integration in EPNet: Production and distribution of food during the Roman Empire. Engineering Applications of Artificial Intelligence, 51, pp.212-229.

Calvanese, D., Mosca, A., Remeseal, J., Rezk, M., Rull, G.. A 'Historical Case' of Ontology-Based Data Access, Proceedings of the 2015 Digital Heritage International Congress, Vol. 2, pp. 291-298, IEEE, Granada, 2015, ISBN: 978-1-5090-0047-0

McInerny, G.J., Chen, M., Freeman, R., Gavaghan, D., Meyer, M., Rowland, F., Spiegelhalter, D.J., Stefaner, M., Tessarolo, G. and Hortal, J., 2014. Information visualisation for science and policy: engaging users and avoiding bias. Trends in ecology & evolution, 29(3), pp.148-157.

Mosca, A., Remeseal, J., Rezk, M., Rull, G.. Knowledge Representation in EPNet. In Morzy, Tadeusz, Patrick Valduriez, and Ladjel Bellatreche, Eds. New Trends in Databases and Information Systems: ADBIS 2015. Poitiers, France, September 8-11, 2015. Communications in Computer and Information Science Proceedings, Vol. 539, pp. 427-437. Springer, 2015.

Porter, M.A. and Boerner, K., 2011. Atlas of Science Visualizing What We Know. Science, 331(6018).

Remesal, J., Díaz-Guilera, A., Rondelli, B., Rubio, X., Aguilera, A., Martín-Arroyo, D., Mosca, A., Rull, G.. The EPNet Project. In Information Technologies for Epigraphy and Cultural Heritage. Proceedings of the First EAGLE International Conference (Paris, September 29-30 – October, 1, 2014), Roma 2014, pp. 455-464.

Remesal, J., Rondelli, B, The EPNet Project: a non-conventional framework for falsifying historical theories. In Economía romana. Nuevas perspectivas / The Roman economy. New perspectives, Barcelona 2017, pp. 119-124.