Retrieving and Integrating Archaeological Data on the Web

The Herculaneum Case Study

Achille FELICETTI, PIN, University of Florence, Italy
Ilenia GALLUCCIO, PIN, University of Florence, Italy

Keywords: Archaeological Data, FAIR Principles, Ontologies, CIDOC CRM, Semantic Interoperability

Introduction

The first of the four FAIR principles, "Findability", essentially aims at answering one of the key problems of the digital world: the data live in isolated and fragmented ecosystems, so much so that in many cases it is very difficult to find them online. This is especially true for those produced by archaeological research. The FAIR recommendations are very useful for planning future actions, but how can we deal with the already published (and somehow already "compromised" in the FAIR perspective) legacy data? Some international initiatives, such as the ARIADNEplus project, are implementing policies to provide answers to this question, availing of the consensus and collaboration of large international institutions. However, a great deal of information on the Web nowadays remains out of the reach of any integration strategy. This paper sketches a possible solution to face this issue and proposes methods to integrate archaeological information dispersed online.

Archaeology and digital archaeology: a fragmented world

When dealing with archaeology, the dispersion we are witnessing in the digital world today finds an exact parallel (and in many cases an origin) in the real world. For centuries, the exploitation of archaeological material for private, propagandistic or commercial purposes has plagued archaeology and caused the dispersion of objects and materials, very often have ended up thousands of miles away from their original places of discovery. In the digital world, this has resulted in the online publication of data and metadata regarding this archaeological material in a myriad of different formats, languages and structures. This poses enormous obstacles to scholars who want to use the Web for their research and inevitably complicates the study of artefacts and monuments that are stylistically similar or coming from common archaeological contexts. The case we have chosen for our paper is one of the most emblematic in this regard: the artefacts originating from the archaeological excavations of the ancient city of Herculaneum.

Herculaneum: the dispersion of an archaeological treasure

Destroyed together with Pompeii by the eruption of Vesuvius in 79 AD, Herculaneum remained in oblivion for sixteen centuries until it was discovered by chance at the beginning of the 18th century during the excavation of a well. From that moment on, the site witnessed an uninterrupted series of excavations, during which an unimaginable amount of archaeological material came to light and its objects and monuments, either legally acquired or illegally stolen, witnessed over the centuries a great dispersion over the world. A lot of them ended up in private collections and in many of the great national museums of Europe and the United States. In the last decades, most of these institutions have published their catalogues online and created metadata for each of the objects in their possession. However, it is currently impossible to use Google or any other web tool to query this multitude of archives to retrieve, for example, all the bronze statues found during a specific excavation or all the marble artefacts manufactured in Herculaneum before its destruction. A system capable of providing adequate answers to such questions would constitute an incomparable resource for scholars.

“Semantic Herculaneum”: the reassembly of a digital treasure

One of the possible solutions to build such a system consists in finding, harvesting and aggregating dispersed archaeological information on a single platform to make them interoperable and usable in an integrated way. The use of ontologies and semantic tools is essential in this context to provide data with a high level of standardisation and to preserve the digital provenance, i.e. the connection with the information on the original websites, along the whole process. To test the feasibility of our approach, the web catalogues

of the Getty Museum, the Staatliche Museen zu Berlin, the Museo Archeologico Nazionale di Napoli and the British Museum were chosen for the large number of Herculaneum artefacts they currently host. A series of special ad hoc scripts have been developed to exploit the specific query facilities of each museum's website for extracting relevant information reported within the results' web pages.

Despite the diversity of data structures and encodings, strong similarities have been identified among retrieved information, especially with regard to the way in which basic data such as dimensions, materials, shapes and provenance of objects is presented. CRMarchaeo, the extension of the CIDOC CRM devoted to the description of objects and phenomena of the archaeological world, was the ontology chosen as the common language of the integrated archive. CRMarchaeo is an event-based model capable of describing artefacts in great detail by associating them with the complex web of activities in which they have been involved in their lifetime. Metadata generated using CRMarchaeo in a formal and standard language can thus narrate the history of the artefacts from their production, use and discovery up to their acquisition, status and current location.

The information extracted from the pages of the original websites has been encoded in CRMarchaeo format through a series of mapping and conversion operations resulted in a set of RDF triples. The 3M tool developed by FORTH has been used for this operation. The data extraction, mapping and conversion platform we set up allow the entire process to be re-executed in an automated way to obtain fresh and updated data if required. This way, the metadata of about 500 Herculaneum artefacts of different epochs and nature (bronzes, mosaics, marble statues, inscriptions, papyri) have been integrated without loss of information and preserving digital provenance of data (thus fostering data "Accessibility" recommended by FAIR).

In the resulting semantic graph, managed by means of the BlazeGraph database, all data are encoded in a unifying language, coexist in a single structure and are perfectly interoperable, regardless of their digital origin (see Fig. 1). Appropriate semantic search interfaces have been implemented to query the aggregated data using ResearchSpace, a tool developed by the Mellon Foundation for the British Museum that makes the graph browsable along the axes of different properties and offers interfaces for most of the advanced semantic queries defined for CIDOC CRM information, such as those based on the Fundamental Categories and Properties theorized by Tzompanaki and Doerr (2012).

ResearchSpace also allows sophisticated data analysis on integrated information, including the ability to derive knowledge on whether and how various artefacts have interacted with each other over time, and to investigate for spatial and temporal relations between them (see Fig. 2). It also allows arranging the artefacts on timelines or maps, to obtain spatial distribution and to group them according to specific criteria, such as excavation or production events, i.e. objects present in a specific area or place (like a *villa* or a *domus*) at a given time or coming from a specific *atelier*.

RDF, the standard used to encode information, allows them to be easily integrated in other CIDOC CRM-enabled environments (such as the ARIADNEplus infrastructure or the PARTHENOS Data Cloud), exported in other formats, and published as Linked Open Data. The Interoperability and Reusability of the new archive (corresponding to the "I" and "R" of the FAIR principles) are thus guaranteed.

Conclusions

This experiment produced extremely encouraging results and has shown that the retrieval and aggregation of information published on the web can be a viable way to establish interoperability, even over non- or dis-integrated and hardly accessible data. The CRMarchaeo model has shown its full potential in providing a comprehensive language for dealing with the multiple facets of the various encodings and data formats; the tools used to manage and query the integrated semantic information avail advanced queries and in-depth analysis of the integrated data. In conclusion this activity has shown in practice how FAIR principles can be propagated even to un-FAIR data present online through the use of ontologies and standard tools.

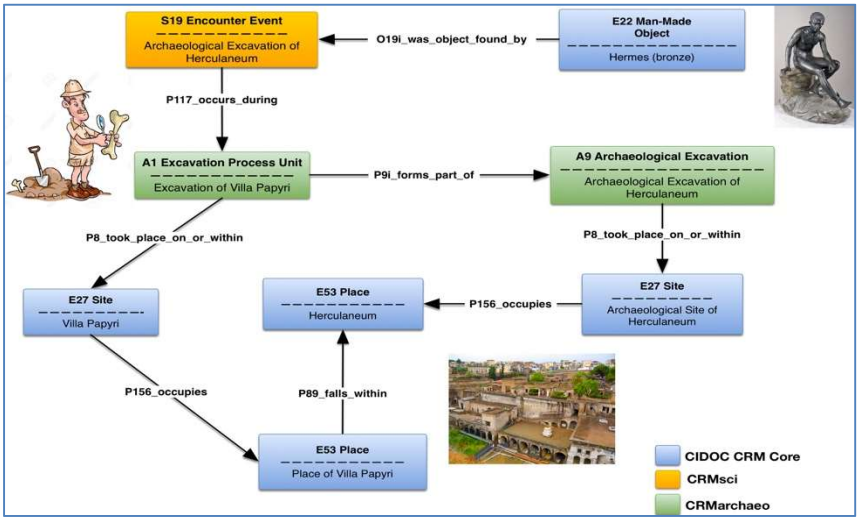

Fig. 1. CIDOC CRM encoding of information concerning a bronze artefact coming from the Herculaneum site.

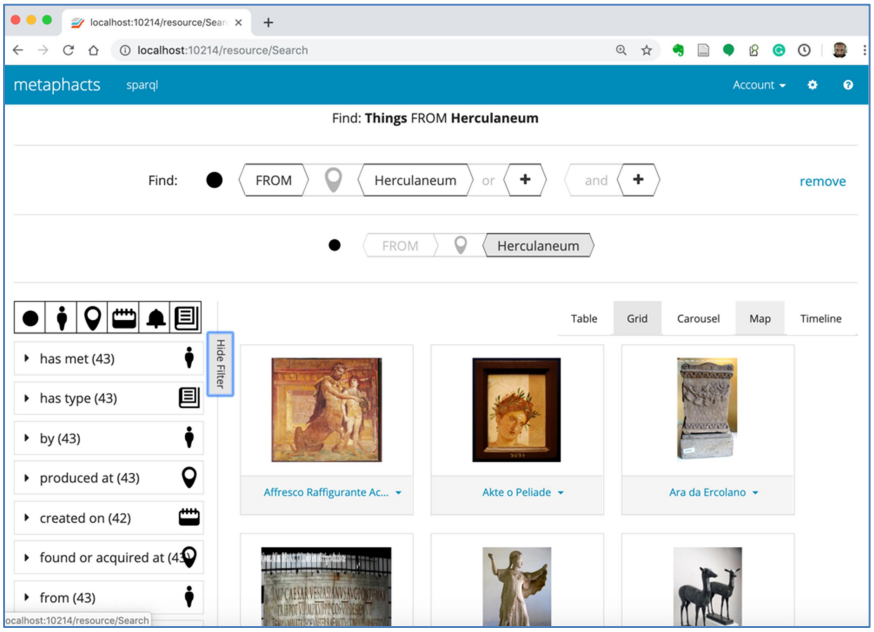

Fig. 2. The semantic interface configured for the Herculaneum data showing various ResearchSpace query features.

References

ARIADNEplus Project: https://ariadne-infrastructure.eu

Doerr, M., Felicetti, A., Hermon, S., et al (2019). 'Definition of the CRMarchaeo, an extension of CIDOC CRM to support the archaeological excavation process'. ICOM CIDOC Conceptual Reference Model Special Interest Group, 2016. Technical report 1.4.8. Prato, Italy, PIN S.c.R.L. http://cidoc-crm.org/crmarchaeo (Accessed: 22 July 2019)

Felicetti, A., Gerth, P., Meghini, C. and Theodoridou, M. (2016). 'Integrating Heterogeneous Coin Datasets in the Context of Archaeological Research', Extending, Mapping and Focusing the CIDOC CRM. Paola Ronzino and Franco Niccolucci (eds.). CRMEX 2015. Workshop, 19th International Conference on Theory and Practice of Digital Libraries (TPDL 2015), Poznan, Poland, September 17, 2015. http://ceur-ws.org/Vol-1656/paper2.pdf (Accessed: 22 July 2019)

BlazeGraph Semantic Database: https://www.blazegraph.com (Accessed: 22 July 2019)

ResearchSpace Knowledge System: https://www.researchspace.org (Accessed: 22 July 2019)

Tzompanaki, K. and Doerr, M. (2012). 'Fundamental categories and relationships for intuitive querying CIDOC-CRM based repositories', Technical Report TR-429, ICS-FORTH, April 2012.