The 13th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2012)
D. Arnold, J. Kaminski, F. Niccolucci, and A. Stork (Editors)

Including Links in Linked Data: CIDOC-CRM and the Fourth T. Berners-Lee Rule

A. D'Andrea

CISA, Università di Napoli "L'Orientale", Napoli, Italy

Abstract

The latest advancements in the semantic technologies are pushing many scholars towards new challenges, mainly in the field of data-sharing and data-integration. The introduction of the Linked Open Data (LOD) paradigm has opened new scenarios thanks to the four rules given by T. Berners-Lee that are contributing to create a common framework for the publication of contents on the WEB. This contribute tries to provide a state of art of the integration of CIDOC-CRM into Linked Data paradigm. It shows some possible research lines for the implementation of the fourth rule addressed to enrich content by including links to other data.

Categories and Subject Descriptors (according to ACM CCS): H.3.7 - Digital Libraries

1. Introduction

The archaeologist's job has gone through radical changes: from the first diaries or hand-written reports, to the use of the most modern and sophisticated digital technologies. Nevertheless the final goal of archaeological research remains basically unchanged: the recording and documentation of all possible information gathered during on-ground investigations.

In the last years several and different motivations have pushed many scholars to adopt a new strategy to manage archaeological datasets. The growth of on-field archaeological activities have yielded an exponential increase of data requiring new approaches and solutions in order to archive, safeguard and, finally, exploit this, apparently infinite, source of information. The explosion of a computerized approach to the archaeological collections and data has determined the creation of a number of management systems forms designed according to the formal structure of software like databases.

In order to guarantee homogeneity in data acquisition, archaeologists have spent most of their energies and efforts in the organization and optimization of the data collection procedures and at the same time in the creation of vocabularies and controlled word-lists. The end result has been the creation of information management systems, often encoded in conformity with the target of the archaeological project or conditioned by the background of the archaeologists or, finally, determined by local (national or regional) rules.

Some years ago, in the evaluation of the scientific approach of archeologists to the excavation, Carver [Car90] correctly pointed out this situation by highlighting the different approaches and data-collection by Processualist and Post-Processualist Archaeologists: the former analyze the stratigraphy like a deposit with a technical and "cold" approach; the latter, on the other hand, face the archaeological strata as a narrative already impressed in their mind. Integrating these different perspectives on research or the researchers' mental attitudes is not a simple task and neither is the "construction" of an informatics system able to foresee every recordable element/object, analysis or evaluation of the data, without diminishing the complexity of the archaeological research.

Thanks to the latest advancements in Information Technology, it is already possible to confront the problems regarding data integration in a new way based on the assumption that it is no longer necessary to force researchers to sacrifice their point of view and their approach. In the near future it will be possible to overcome this present fragmentation, ensuring at the same time a high level of data compatibility and sharing without modifying the data-model chosen by the archaeologist. To achieve this objective, standardization of the archaeological archives must reach a high level, in order to make them compatible and accessible without forcing archaeologists to choose a unique data-model. The third generation Web will allow to set up a new approach for data integration and retrieving by mapping different archives according to semantic and descriptive standards.

The developments in the semantic technologies are pushing many scholars towards new challenges, mainly in the field of data-sharing and data-integration. The introduction of the LOD paradigm has recently opened new scenario thanks to the four rules given by T. Berners-Lee that are contributing to create a common framework for the publication of contents on the WEB [Ber06].

The four recommendations are basically intuitive:

DOI: 10.2312/PE/VAST/VAST12S/009-012

1. Use URIs as names for things;
2. Use HTTP URIs so that people can look up those names;
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL);
4. Include links to other URIs.

As regards the first three rules the adopted conventions reflect a practical way to encode data for the retrieving process of the content. Less clear is the fourth that appears rather an invitation than a mandatory prescription. How should we consider it? Like a sort of recommendation in order to add generic links or an attempt to tie different specialized domain with the scope of enriching data?

The fourth rule can be considered as an encouragement with no relevance on the designing, implementing and publishing of LOD. On the contrary the missed application of the three previous rules can determine the incompatibility with the standard of LOD. It seems that the compliance with the fourth rule relies on the fate or on the feeling of the researcher. Notwithstanding most probably just through the links it is possible to assure a wide diffusion of the dataset and to contribute to enlarge the network by creating a network among domains and knowledge and not only among data. Only a correct way to consider the fourth rule can provoke a major awareness in the development of datasets encoded into LOD.

Even if we consider the fourth rule relevant for the diffusion and acceptance of LOD, it is no recommendation about its implementation. According to T. Berners-Lee it is enough to include links in the data, while there are neither explanations nor suggestions and, mainly, there is no standardized and agreed approach. In order to reach the scope of the inclusion of links into LOD a good suggestion could be provided by turning over the normal approach in a top-down view. Why must I include links in my LOD? Only because there is a rule, the fourth? Or I can exploit this rule in order to obtain a wider diffusion of my datasets? What is my objective? What knowledge or domains I can reach by means a good and structured level of links? What kind of search do I want or can I realize? Apparently the lack of information on the fourth rule impoverish the dataset by depriving it of every possible linkage with other digital archives.

Some recommendations could be given for the integration of CIDOC-CRM (*www.cidoc-crm.org/docs/cidoc_crm_version_5.0.4.pdf*) archives through a more appropriate exploitation of the fourth rule. After a short description of all major projects aimed at transforming datasets compatible with the CIDOC-CRM schema into LOD format, the paper will outline some possible strategies and tracks for a more useful approach to the LOD scenario by embedding different specialized ontologies, taxonomies and thesaurii into CIDOC-CRM archives.

2. State of Art

CIDOC-CRM has become a practical and conceptual schema for the linkage and integration among different datasets. Born as an ontological model basically for museum collections, CIDOC-CRM afterwards was used by different research teams for the merging of multiple digital archives containing disparate and not homogeneous archaeological collections (coins, statues, vessels, etc.).

The results obtained up to now in terms of semantic enriching of cultural archives are pushing numerous scholars to deepen and develop CIDOC-CRM by means of extensions and alignments. Recently thanks to the development of the semantic technologies for the WEB the CIDOC-ICOM group released recommendations on LOD for museums (*www.cidoc-crm.org/URIs_and_Linked_Open_Data.html*).

Starting from some philosophical remarks, the CIDOC-ICOM group proposed that:

1. When museum objects are referred to in Internet applications, it is necessary that the objects be uniquely identified by suitable URIs.

2. In order to avoid different institutions generating competing URIs for the same object, each object (or set of objects) should have one preferred authority that assigns the URI for the object. The URI authority for the object must be known to all interested parties or be easy to discover.

3. The most natural candidate for the URI authority for an object is the museum that curates the object, regardless of whether the museum intends to provide its own services on the Internet or not. This is because it is the only institution that can absolutely determine that two different museum object URIs actually describe the same thing.

Even if the CIDOC-ICON group recognised "*the accelerating trend for cultural resources to become the subject of digital resources*" the only recommendation given is related to the way for the assignment of the URI (rule n° 1). No further suggestions are given for rule n° 4 which is left in a state of uncertainty. Different (and minimalist) solutions have been recently proposed for the inclusion of other links in a dataset compatible with CIDOC-CRM schema.

In the VAST 2009 [JCHD09] a paper dealt with a model called CHoWDer (Cultural Heritage on the Web of Data) designed to illustrate how easily CH institutions could participate in the Linked Data initiative. They simply proposed to add a triple (for examples the following):

http://CHInstitution.org/MuseumBenaki

owl:sameAs

http://dbpedia.org/resource/Benaki_Museum

that didn't interfere with data, but that could allow to share an RDF graph containing RDF links; in this case all the

information stored in the Benaki Museum, in Athens. CHoWDer was not designed to answer all questions regarding linking CH to the Web of Data, but rather, to explore the possibilities of interlinking data. Through a simple triple, different RDFs could be linked giving the possibilities to discover other objects in the Benaki Museum belonging to other collections.

Another research [RMA09] exploits a mechanism for tagging - via an research application called *Tagg3D* - parts of a 3D geometry to CIDOC-CRM URIs in order to store entry points to a metadata repository. Like the previous project the architecture uses the D2R server to store the CIDOC-CRM.

While the first project uses the link principles according to the fourth rule in order to connect different datasets by means of the instruction *owl:sameAs*, the second one on the contrary uses a mechanism to generate a linkage between the metadata of the object and its 3D geometry available on on-line. As the CIDOC-CRM schema is aligned with COLLADA, the *Tagg3D* doesn't provide any linkage to other datasets, but simply assigns URI describing the geometry of the object.

Both papers don't give any information about a standard application of the fourth rule, neither they provide solutions for those given by the CIDOC-ICOM group. They seems on-going experimental researches aimed at finding solutions and designing software for generic issues concerning the adoption of LOD principles rather than projects addressing to the implementation of real dataset, digital archives or Semantic engines.

LODAC (Linked Open Data for ACademia) (*www.museumsandtheweb.com/mw2011/papers/building_linked_data_for_cultural_information_*) is a Japanese project which aims to share and publish a wide range of data using LOD. Since Japanese museums have developed unique collection systems it is difficult to retrieve relevant information by searching multiple museum databases. To solve this problem, the Agency for Cultural Affairs has directed the development of a common index search system called a "Cultural Heritage Online". The Tokyo National Museum published "Structured Model for Museum Object Information" which is based on the concept of CIDOC CRM. This model is seldom applied to museums in Japan as it is not easy for curators or museum staff. Recently, a growing number of organizations is building crossover search systems to enable multiple museum database searches. To overcome this fragmentation the LODAC project is an attempt to solve multiple searches by implementing a LODAC-Museum with Semantic Web and LOD technology. The prototype system uses existing museum collection data, thesauri, and other types of information. LODAC-Museum generates datasets which are scraped from individual museum websites, mapped to RDF in CIDOC CRM and then transformed into Linked Open Data format.

LODAC Museum tackled several challenges including:

1. Aggregatation of cultural heritage and museum information from all museums in Japan
2. Integratation of natural language search and provide reliable SPARQL endpoint

As regards the fourth Berners-Lee rule only future plans foreseen to link the prototype to generic LOD contents such as dbpedia.jp

Similarly the CLAROS project (*explore.clarosnet.org*) uses an extension of CIDOC-CRM as a glue to enable simultaneous searching of major collections in university research institutes and museums. The CLAROS Data service (*data.claros.org*) provides an interface for the data of the CLAROS Project, and complements the CLAROS Explorer. This service provides metadata about archaeology and art in machine-readable formats such as RDF, JSON and KML. The user can use the Objects and People views to start exploring. As for the previous projects no further links have been designed and implemented to join CLAROS project to other archaeological datasets. The only integrated source is the "*Lexicon of Greek Personal Names*" (LGPN) that was established to collect and publish all ancient Greek personal names, drawing on the full range of written sources from the 8th century B.C. down to the late Roman Empire.

The last two projects, analysed in this contribution, implemented only the first three rules leaving unsolved the fourth one. The LOD available on the website of the Archaeology Data Service (archaeologydataservice.ac.uk/research/stellar/) are a direct result of the STELLAR project, a joint project between the University of Glamorgan, the ADS and English Heritage. LOD is used as a method of exposing, sharing, and connecting data via URIs on the Web rather than a procedure to integrate different domains including links. Similarly way all British Museum URIs start with HTTP and are within the domain *collection.britishmuseum.org*. There is an access to the collection data available through the Museum's web, but in a computer readable format. The use of the W3C open data standard, RDF, allows the Museum's collection data to join and relate to a growing body of linked data published by other organisations.

3. A new strategy for the fourth rule

All the projects described so far, use the LOD paradigm only to convert their digital archives, compatible with the CIDOC-CRM schema, into a computer readable format. No particular suggestion or recommendation have been provided for the implementation of fourth rule. While CIDOC-CRM is utilized as the glue for the integration of disparate resources on the base of a standardized and accepted model, the Linked Data are used to allow the publishing and sharing of content on the Web.

Is it possible to outline a different scenario more aligned with the intentions of T. Berners-Lee by enriching data with multiple links to other datasets?

CIDOC-CRM is based on a conceptual schema focused and centred on the event considered as the main feature

able to link different entities and features. Basically the most important element of the CIDOC-CRM schema are: Who (Actor), Where (Place), What (Objects) and When (Time).

If we try to assign to each basic feature a specialized domain we can find a good way to implement the fourth rule and also to guarantee a level of compatibility of the archives with other domain.

The category Actor-Who including people or group of people (modern or ancient) and institutions can be easily described linking the source to FOAF (Friend-of-a-Friend) (*www.foaf-project.org*) The FOAF project describes people, the links between them and the things they create and do. It began early in 2000 as an "*experimental linked information project"* and currently is a small but shapely piece of the wider Semantic Web project. FOAF is a simple technology that makes it easier to share and use information about people and their activities to transfer information between Web sites, and to automatically extend, merge and re-use it online. By linking the entity Actor to FOAF is possible exploit further links. For instance if I search a particular archaeologist responsible for the cataloguing of a collection or for the excavation of an important archaeological site I can find further information about other researches or activities in the same o in other field.

The category Place-Where including names of ancient or modern places and sites can be linked to specialized geographical database like Geonames (*www.geonames.org*) that covers all countries. The GeoNames database contains over 10,000,000 geographical names corresponding to over 7,500,000 unique features. Beyond names of places in various languages, data stored include latitude, longitude, elevation, population, administrative subdivision and postal codes. Those data are accessible free of charge through a number of Web services and a daily database export. The Web services include direct and reverse geocoding, finding places through postal codes, finding places next to a given place. Scholars interested in other information related to the place or site where they are currently investigating, through Geonames could discover further relevant data. Alternatively s an alternative it is possible to use DBPedia (*wiki.dbpedia.org*) which is more generic as it is a community effort to extract structured information from Wikipedia and to make this information available on the Web. This link is particularly efficient to describe modern places like museums, etc. If you are interested in the name of ancient places you can use also Pleiades (*pleiades.stoa.org*) that gives scholars and students the ability to use, create, and share historical geographic information about the ancient world in digital form. At present, Pleiades has extensive coverage for the Greek and Roman world, and is beginning to expand into Ancient Near Eastern, Byzantine, Celtic, and Early Medieval geography.

The Object-What is a more problematic category because a unique database containing all objects and artefacts discovered and exhibited doesn't exist. In this case it would be more appropriate to use a thesaurus encoded in SKOS or a generic source illustrating categories of ancient object or, preferably, to link the data to an available digital archive encoded into LOD like those of the British museum or CLAROS project already described. This is one of the scopes of the LOD i.e. reusing the sources in order to avoid a duplicate description. Nevertheless this is a common practice among the archaeologists that use to refer to a previous published object to catalogue those coming to their investigations. Images of photos could be published in Flickr (*www.flickr.com*) and linked to the objects giving a fast and a simple way to compare and catalogue objects.

Less simple is to express links for the last category Time-When. Perhaps would to be a good approach try to refer to the time by using the entities and properties that are in more foundational ontologies like Proton (*proton.semanticweb.org*) and Dolce (*www.loa.istc.cnr.it/DOLCE.html*).

4. Conclusions

FOAF, GEONAMES, DBPedia, PLEIADES, FLICKR are databases containing an impressive number of records which are continuously incremented. Linking digital archives to these on-line sources can ensure a real enrichment for our resources and mainly the possibility to discover new information. This is the only way for the implementation of the fourth rule and to guarantee the survival of our data and our archives.

The LOD have been used so far to publish data on the Web according to a machine readable standard, but no one has still explained which is the best way to aggregate the different archives in a really integrated world of data. The results reached are encouraging, but new and further effort should be focused on the implementation of the fourth T. Berners-Lee rule. Only through this approach all the principles of LOD paradigm could be observed.

References

[Ber06] BERNERS-LEE T.: Linked Data—Design Issues. www.w3.org/DesignIssues/LinkedData.html, 2006.

[Car90] CARVER M.O.H.: Digging for data: archaeological approaches to data definition, acquisition and analysis. In *Lo scavo archeologico: dalla diagnosi all'edizione*, R. FRANCOVICH AND D. MANACORDA Eds., 1990, pp 45-120.

[JCHD09] JANKOWSKI J, COBOS Y, HAUSENBLAS M., DECKER S.: Accessing Cultural Heritage using the Web of Data. In *10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage 2009 (VAST09)* St. Julians, Malta, 2009.

[RMA09] RODRIGUEZ-ECHAVARRIA K., MORRIS D., ARNOLD D.: Web based presentation of semantically tagged 3d content for public sculptures and monuments in the UK. In Web3D 16-17 June 2009, Darmstadt, Germany.