Using CIDOC CRM for dynamically querying ArSol, a relational database, from the semantic web

In the MASA Consortium context, the Laboratoire Archéologie et Territoires propose to open the ArSol database, a system for processing archaeological data, to the semantic web, using the CIDOC-CRM ontology and tools that implements Ontology-Based Data Access (OBDA) principles. After designing a set of mappings from ArSol fields to the CIDOC CRM ontology, we implement the software architecture to query ArSol from a SPARQL endpoint. We used -ontop-, a Protégé plugin, to design the OBDA mappings that are necessary for the SPARQL-to-SQL rewritings. Our final goal is to devise an application that will offer a single interface to query several distributed and independent archaeological databases, with heterogeneous structures, using CIDOC-CRM to relate them to each other. Querying ArSol in SPARQL via the CIDOC CRM is an important step towards this goal.

Olivier Marlet [1], Stéphane Curet [1], Xavier Rodier [1], Béatrice Bouchou-Markhoff [2]

[1] UMR 7324 - CITERES - Laboratoire Archéologie et Territoires, MSH Val de Loire
[2] Laboratoire d'Informatique, Université François Rabelais de Tours

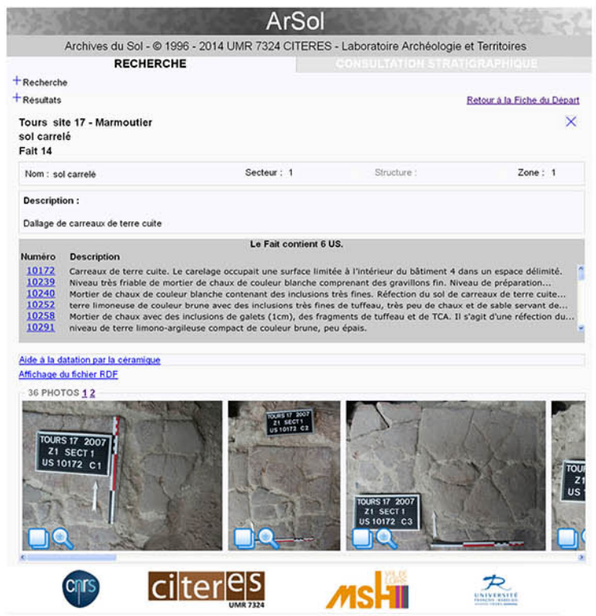
1 What is MASA consortium ?

The MASA Consortium ("Memory of Archaeologists and Archaeological Sites"), approved by the TGIR Huma-Num (Very large Facility in digital humanities), gathers several performers in French archaeological research. One of its missions is to propose solutions to sustain and homogenize archaeological information. Within this consortium, the Laboratory Archaeology and Territories in Tours took on the task of working on the issue of the archaeological data interoperability.

The objective is to develop a unified access to digital corpuses using procedures and common tools for documenting and archiving. Sharing, and particularly querying, all archaeological data involves being relieved of the constraints of specific software, structure, vocabulary, accessibility and language. The CIDOC-CRM ontology (international standard which structures digital data for Cultural Heritage) seems to be the most comprehensive model usable as a point of convergence for various archaeological databases. The goal is to enable researchers to make their system interoperable without changing thier database structure.

2 What is ArSol system ?

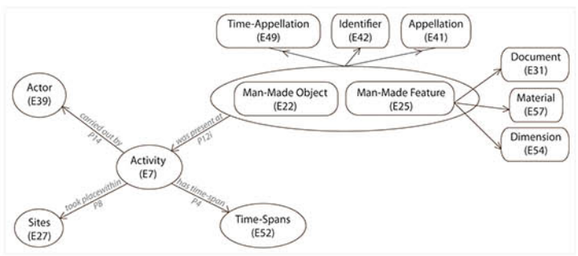
To work on interoperability, we have decided to work on the information system that we have been using for a long time: ArSol (Archives du Sol). This system has been developed with the 4D software since 1990 for all the stratigraphic excavations with a dual aims of data management and research.. First and foremost, to be interoperable, a database must necessarily be available on the web. The ArSol system has an online version allowing access to data from the Web.



http://arsol.univ-tours.fr

3 CIDOC CRM ontology

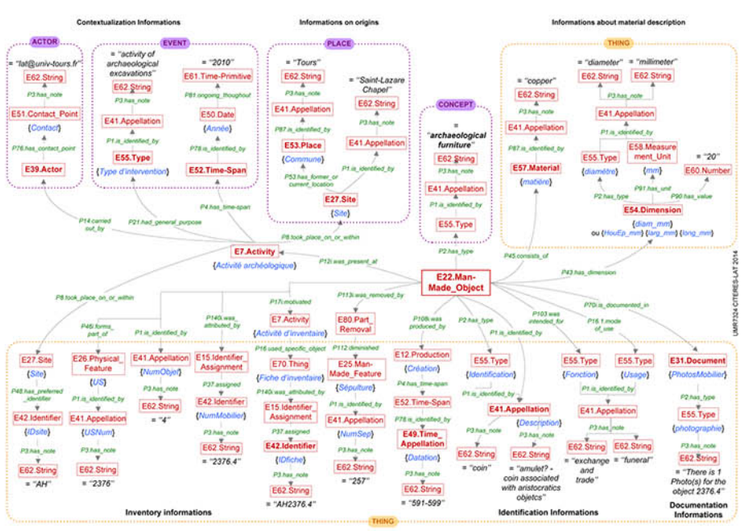
Initiated in 1994, this ontology was originally designed for museums, but its scope is so broad that we can find most of the elements and properties specific to archaeological data, in particular, entity "E22.Man-Made Object" corresponding to artefacts and entity "E25.Man-Made Feature" corresponding to features.



We see this ontology as a consensual description overlayer through which the query can pass, allowing communication with each database thanks to a system of equivalence, or "mapping". By using ontology, data become free from the database management system software and from the own database structure.

4 Mapping

The mapping consists in matching each field in the database with an entity of ontology respecting the whole hierarchical tree and the properties that organize the entities in relation to each other.

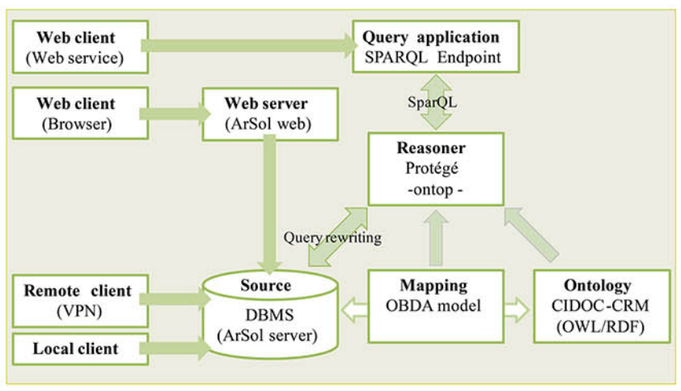


The RDF triplet "subject-predicate-object" becomes "source Entity-Property-Target Entity". The target entity at the end of one branch is the element that must be associated with the field in the database. That field is collected in the mapping using a SQL query in the database.

E22 Man-Made Object {Mobilier}. P16.1 mode of use: E55 Type {Usage}
E55 Type {Usage}. P2 has type: E55 Type
E55 Type. P1 is identified by: E41 Appellation
→ datatype = *"usage"*
et:
E55 Type {Usage}. P1 is identified by: E41 Appellation
→ value example = *"funéraire"*

5 Software architecture

For communicating the ontology with database using mapping as key match, a solution lies into the principles of OBDA systems (Ontology Based Data Access), recently introduced in the semantic web. With this OBDA layer, ontology is the access point to query the data, while the information remains in the databases used by researchers. OBDA system consists of: an ontology (semantic level), various data sources (in our case, ArSol), a set of mappings expressing the relationship between the data source and ontology, and an application layer to manipulate and query the system.



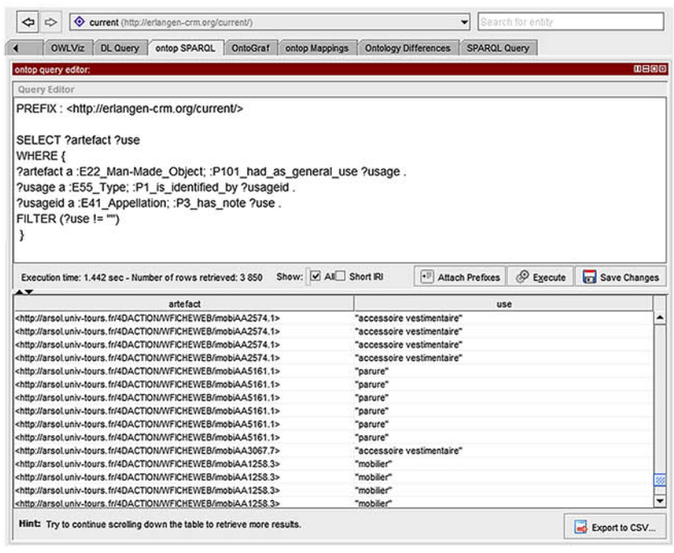
6 Protégé and -Ontop-

We use Protégé (free software dedicated to ontologies) and -ontop-, an extension developed at the University of Bozen-Bolzano (Italy). It first allows us to specify the matches between ontology and database, then to evaluate SPARQL queries about ontology, whose results are searched for in the database. The integrated user's interface of Protégé helps us to make the OBDA file containing statements of correspondences between ontology and data source.

```
[PrefixDeclaration]
:          http://erlangen-crm.org/current/
owl:       http://www.w3.org/2002/07/owl#
rdf:       http://www.w3.org/1999/02/22-rdf-syntax-ns#
xsd:       http://www.w3.org/2001/XMLSchema#
xsp:       http://www.owl-ontologies.com/2005/08/07/xsp.owl#
ecrm:      http://erlangen-crm.org/current/
rdfs:      http://www.w3.org/2000/01/rdf-schema#
skos:      http://www.w3.org/2004/02/skos/core#
swrl:      http://www.w3.org/2003/11/swrl#
swrlb:     http://www.w3.org/2003/11/swrlb#
protege:   http://protege.stanford.edu/plugins/owl/protege#
[SourceDeclaration]
sourceUri         jdbc4D
connectionUrl     jdbc:odbc:BaseArsol
username          LAT
password          ********
driverClass       sun.jdbc.odbc.JdbcOdbcDriver
[MappingDeclaration] @collection [[
mappingId         MAPID-1
target            <http://arsol.univ-tours.fr/4DACTION/WFiCHEWEB/imobi{CoSiUSObj}> a ecrm:E22_Man-Made_Object ;
                  ecrm:P101_had_as_general_use <http://arsol.univ-tours.fr/4DACTION/WFiCHEWEB/imobi{CoSiUSObj}/usage>.
                  <http://arsol.univ-tours.fr/4DACTION/WFiCHEWEB/imobi{CoSiUSObj}/usage> a ecrm:E55_Type ;
                  ecrm:P1_is_identified_by <http://arsol.univ-tours.fr/4DACTION/WFiCHEWEB/imobi{CoSiUSObj}/usageid>.
                  <http://arsol.univ-tours.fr/4DACTION/WFiCHEWEB/imobi{CoSiUSObj}/usageid> a ecrm:E41_Appellation ;
                  <http://arsol.univ-tours.fr/4DACTION/WFiCHEWEB/imobi{CoSiUSObj}/usageid> ecrm:P3_has_note {G_usage} . source
                  SELECT CoSiUSObj, G_usage FROM af4mobilie
]]
```

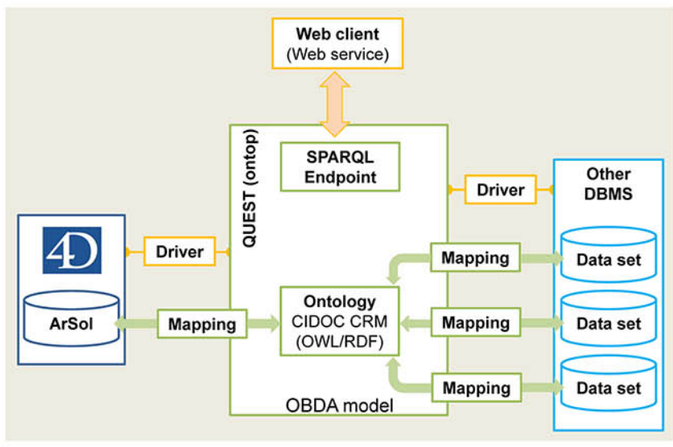
7 SPARQL endpoint

To offer online query for web applications, which is called "SPARQL endpoint", we need then an HTTP server like Jetty (Java) with the web application Sesame-Workbench that contains Quest, -ontop- version independent of Protégé and ensuring the evaluation of SPARQL queries. Ontology is loaded as an XML-OWL file. The data source is defined by indicating the required drivers to allow Java to communicate with the source database software.



8 Interoperability

Now, to give full meaning to interoperability, we expand our number of data sources, doing the same process with others archaeological databases. The only constraint is that each database should have a minimum of descriptor fields, which is the case for most databases used in archaeology, and that it must be web-accessible.



The SPARQL endpoint is essential for the web services that will query ArSol but we also have to allow humans to access ArSol and other databases via this semantic level. So we will develop a web interfaces for users, with the SPARQL endpoint.