Using ontologies as a faceted browsing for heterogeneous cultural heritage collections

Francesca Tomasi[1], Fabio Ciotti[2], Marilena Daquino[3], and Maurizio Lana[4]

[1] Department of Classical Philology and Italian Studies, University of Bologna, Italy
francesca.tomasi@unibo.it
[2] Department of Humanities, University of Roma Tor Vergata, Italy
fabio.ciotti@uniroma2.it
[3] Multimedia Research Resource Centre (CRR-MM), University of Bologna, Italy
marilena.daquino2@unibo.it
[4] Department of Humanities, University of Piemonte Orientale, Italy
maurizio.lana@uniupo.it

Abstract. In this paper we present a project regarding the possible use of multiple and interconnected OWL ontologies (GO, HiCo, and Proles) in order to explore the semantic content of heterogeneous digital collections (a digital library, a full-text scholarly edition, and a relational database) in the cultural heritage domain (Geolat, Vespasiano da Bisticci Letters, and Zeri photo archive). The aim is to discover knowledge by revealing, through facets, possible latent connections - or even contradictory statements - between data, moving from person, places and dates in an event-centric dimension determined by a context-oriented perspective.

Keywords: ontology matching, data modelling, facets, context, cultural heritage.

1 Introduction

Ontologies are an important tool for conceptualizing knowledge in a domain-oriented perspective, as in the cultural heritage scenery. But, in general, ontologies could be conceived and developed in order to potentially achieve distinct kind of goals, e.g.: to annotate full-texts according to a semantic model [1, 2]; to formalize concepts and constraints in a domain (e.g. digital editions) by using a data-centric approach [3, 4]; to export relational databases in open linked datasets, by converting tables and fields in classes and predicates [5]. Similarly, as known, collections in cultural heritage domain collect data from different areas in humanities (e.g. literature, art, history, etc.), were ontologies likewise could cover different roles and convey multiple functions.

As described later, our research groups developed, in these last few years, many different ontologies, and realized various heterogeneous collections in cultural heritage domain (see references section for relevant papers). It was then a challenging objective to think about a possible interconnection between these ontologies, having

the digital collections as a suite of complex objects, i.e. a knowledge base, to test the ontologies.

Starting from these considerations, the attempt here described is to reflect on possibilities offered by ontologies as cognitive tools. In fact, efforts in modelling could also be exploited for defining 'exploratory' methods. Collections can be browsed by using classes as conceptual categories, i.e. by following a taxonomical approach. Predicates define dynamic filters able to express multi-level relationships. Hierarchical, associative, and equivalence relationships between instances and therefore classes are able to define 'facets'.

In order to test this theoretical model, this feasibility study aims at creating a dialogue between different digital collections and to use ontologies, developed ad hoc for single collections, as a tool for browsing all the heterogeneous complex objects. Cultural objects are keys for creating relationships both at the conceptual and the physical level: original objects, digital objects, subjects of the objects, and interpretations of the objects.

The semantic environment we imagine focuses on specific potentially connected classes and predicates, which could be defined as the facets for browsing the whole collections. Facets help in searching for identities or affinities - but also potentially dissimilarities - between the involved complex objects. In particular, the scenario we imagine aims at defining facets starting from the most common categories in the cultural heritage domain:

- People. Each person has a specific role in context; i.e. the same person could cover different roles, depending on the context where the entity acts. People are agents: they reply to the question 'who';
- Dates. Dates give consistency to actions that involve people in a real or 'virtual' place (document/record/fragment). Dates define temporal entities: they reply to the question 'when';
- Places. Places are useful to identify real or virtual spaces in which events happen. Places are identifiable spaces: they reply to the question 'where'.
- Relationships between classes are defined by using an event-oriented approach: people act in a specific date and place, creating the event. Actions reply to the question: 'what'. The event is related to the context, i.e. the context determines the connection between classes.

Classes representing cultural objects have to be managed according to FRBR[1] model, following the idea of a hermeneutical approach based on a multiple level analysis. People, dates, places and events have all to be described as referred to the specific level of the cultural heritage object: work, expression, manifestation, item.

Starting from these assertions we choose, in detail, to put in dialogue three ontologies (alphabetical order)[2], potentially all related to the formalization of a wide and complex field in the cultural heritage domain (described in section 2):

---

[1] *Functional Requirements for Bibliographic Records* (FRBR), http://www.ifla.org/publications/functional-requirements-for-bibliographic-records

[2] In particular, as referred in the final references, *GO* is by M. Lana and F. Ciotti; *HICO* is by M. Daquino, S. Peroni, and F. Tomasi; *PRoles* is by M. Daquino, S. Peroni, F. Tomasi, and F. Vitali.

- *GO* [1, 2, 6], defines space and places in a geographical dimension;
- *HICO ontology* [3, 7], manages relationships between cultural objects and any entity considered an interpretation about the object itself;
- *Political Roles* (*PRoles*) *Ontology* [3, 4], defines people with a role in a space/time-indexed situation.

These three ontologies have to be experimented on three chosen collections (alphabetical order)[3] that represent three different, but related, projects, both for content and domain (described in section 3):

- the digital library: *Geolat*[4] [1, 2, 6].
- the digital edition: *Vespasiano da Bisticci's Letters*[5] [8].
- the relational database: *Zeri Photo Archive*[6] [5, 9].

In order to complete this project some operations will have to be further defined (described in section 4):

- ontology mapping (classes and predicates) for ontologies alignment;
- URI normalization and identity resolution recognition;
- definition of shared authorities for people, places, dates and event;
- XML markup enrichment in order to enable use of different ontologies on semi-structured contents lacking of semantics;
- elaboration of ad hoc queries to explore potentialities of the aligned models.

We will also have to define a possible framework to explore functionalities of our method, e.g. a semantic repository for semi-structured data, and an environment for the faceted browsing [see e.g. 10] that allows classes and predicates to be transformed in facets.

2 The ontologies

As we said, ontologies in general – and the chosen ontologies in particular – were developed for different use and for documenting different contexts. The project wants to explore connections between these ontologies, in order to use them as a mean for browsing information. The alignment will be necessary in order to guarantee the semantic portability of the resulting model (section 4), even at the moment the ontologies represent individual categories for the exploration. Each single ontology has already imported classes from other similar domains, and proceeded with the mapping

[3] In detail: *Geolat* is by M. Lana, D. Magro, F. Ciotti, with C. Meini, M. Benzi, and G. Vanotti, is funded by Fondazione Compagnia di San Paolo; *Vespasiano da Bisticci's Letters* is by F. Tomasi; *Zeri Photo Archive* is by C. Gognano, F. Mambelli, S. Peroni, F. Tomasi, and F. Vitali.

[4] Geolat, http://www.geolat.it

[5] Vespasiano da Bisticci, Letters, http://vespasianodabisticciletters.unibo.it

[6] Fondazione Zeri, Photographic Archive, http://www.fondazionezeri.unibo.it/catalogo

on the most common ontologies in the cultural heritage domain (especially EDM[7], CIDOC-CRM[8], DC[9] and DCMI terms[10]). The aim is to test these ontologies on the above-mentioned collections (better described in section 3) and to highlight possible re-usability of a single facet in a different context (i.e. another dataset, created with different purposes and a different data structure).

In detail, ontologies here chosen represent different parallel activities of our research group in digital humanities field. The domain is in fact the humanities; in particular the knowledge base on which we modelled concepts is represented by literary texts and documents, but also descriptive entries for cultural objects. The aim of the project is to reflect on the most common features of the analysis in a literary dimension: geographical and historical information (section 2.1); levels and methods of interpretation used by an editor of resources in the field of humanities (section 2.2); roles of people involved in actions described in the sources (section 2.3). Such models, if gathered, can represent a complete and well formalized point of view on approaches and methodologies used by humanists when dealing with several and heterogeneous cultural objects (i.e. texts and digital collections, works of art and related metadata). These ontologies are able to represent a huge part of the formal activities acted by editors considering their hermeneutical approach. Shared entities among ontologies are people, places, dates and events – i.e. all significant and minimal requirements for describing a complex scenario – and each of them further examine a specific issue, fundamental for an overall awareness of the domain of cultural objects description.

2.1 GO

GO is a Geographical Ontology that is now developed inside the Geolat project (see section 3.1). This last has the aim to annotate a digital library of latin texts (digilibLT, www.digiliblt.uniupo.it), where annotation is realized using GO. This ontology is build ad hoc for Geolat, reusing data offered by Pleiades gazetteer[11] and Pelagios[12], and establishing relationships with other relevant geographical ontologies. GO is structured as a two-tier model: a T-box modelling geospatial classes of locations, their properties and their relationships and an A-box with geospatial information about individual places and location. Four modules are defined in GO: GO-TOP: general module for top-level concepts; GO-FAR: For Ancient Resources; GO-PHY: Physic geography (e.g. mountains, rivers); GO-HUM: geosocial module describing Human Artefacts (e.g. cities) and social structures (e.g. formally defined regions, territories).

---

[7] Europeana Data Model (EDM). Documentation, http://pro.europeana.eu/page/edm-documentation

[8] CIDOC Conceptual Reference Model (CRM), http://www.cidoc-crm.org/

[9] Dublin Core (DC), http://dublincore.org/

[10] Dublin Core Metadata Initiative (DCMI) Terms, http://dublincore.org/documents/dcmi-terms/

[11] Pleaides, http://pleiades.stoa.org/

[12] Pelagios, http://pelagios-project.blogspot.it/

2.2 HICO

HICO is an OWL 2 DL ontology created in order to define a formalized, shared and exchangeable model for describing the context of cultural heritage objects and the workflow for stating authoritative assertions about such information. When formalizing any sort of assertions that can be questionable, we are stating something about an object, i.e. extracting something from its content. Each assertion is a subjective authors' "reading" which involves a specific layer of the source wherefrom the reading belongs to, e.g. the text, or better the expression of the source of interest. This compels a multilayers representation of the text – or of the cultural object of interest –, necessary for clearly defining an interpretative process as meta-contextual level for provenance of assertion. It's important to have in mind that a text is a multi-faceted object that has to be treated at different levels of analysis. HiCo ontology, according to FRBR model, attempts to solve part of this issue, and considers other and more specific questions related to the interpretative process. So, HiCo starts from the first formalization of provenance statements in PRoles (2.3) and extends the analysis of required entities involved in the interpretative process, defining a more detailed workflow.

2.3 PRoles

We have developed the PRoles Ontology[13] to explore new possibilities in the representation of authority records. The aim was to formalise complex relationships, such as agents' political roles and events in which these agents are involved, with a specific role, as attested in full-text sources. Initial targets were indeed related to sources dealing with historical events and relevant related politicians, in order to describe useful information in archival and historical context. At the same time this approach was adopted in order to define a deeper analysis of relations among people (corporate bodies, families and persons as in EAC-CPF[14] definition) and documents where they are cited. Except specific features related to this restricted domain of interest, PRoles imports and extends two other models – PRO ontology[15] and N-ary Participation pattern[16] – which allow to describe a wide range of information extracted from full-text sources stakeholders can be interested to.

Although PRO ontology had been thought in principle for an application in the publishing domain, it has been developed so as to accommodate any kind of role, regardless the domain of interest. In particular, PRO defines a class to specify roles an agent can holds, *pro:Role*, and a class for representing role attributions as reified relationships, i.e. individuals of the class *pro:RoleInTime*, which allows to describe agents' having a role in a precise interval and within a particular context (such as in some organisation or place, on a document or with respect to other agents).

Although PRO provides a first formalization of relationships (describing someone

[13] Political Roles Ontology (Proles), http://www.essepuntato.it/2013/10/politicalroles

[14] Encoded Archival Context - Corporate Bodies, Persons, and Families (EAC-CPF), http://eac.staatsbibliothek-berlin.de/

[15] Publishing Roles Ontology (PRO), http://purl.org/spar/pro

[16] Nary Participation, http://ontologydesignpatterns.org/wiki/Submissions:Nary_Participation

holding a role within a particular context), the description of information that can be extracted from full-text often needs an additional level of contextualisation, like describing agents participating to events (located in time and in space) with a particular role. In order to enable such descriptions, another model was reused to include agents with relationships in events, i.e., the N-ary Participation ontological pattern. This pattern describes, mainly, individuals of the class *nary:NaryParticipation*, which allows modelling any object as a participant in an event, i.e., an agent who participates for a specific period of time in an event, holding a time-indexed political role and relating with other objects (agents, places, sources, etc.).

3 The cultural heritage collections

Resources on which the project is working represent three different complex object collections, or even three different models: a digital library, a digital edition, a relational database. All these collections have a technological common base: they are available as XML datasets, and all the information regarding people, places, dates and events have been identified with URIs. The project is to expose all the collections as RDF datasets in a LOD[17] perspective.

The idea is to further analyse and collect such resources, in order to study how objects can be browsed through the use of the aligned ontologies, exploring possible new information generated by alignment itself.

These issues are addressed meanwhile datasets are refined for their publication. In this phase, rethinking in a wide perspective their initial conceptualization and formalization (i.e. the adoption of an ad hoc ontology to model information), will enable more information discoverability whatever will be the implementation. The choice of gathering below described projects is justified by the wide field of research they aim to describe, as together they cover a heterogeneous range of information which could be of interest for other research groups and feasible in other use cases.

As we said above, with respect to the choice of the ontologies, also the collections presented here are the result of our research group activities. These collections cover multiple aspects of humanistic interest: literary texts (section 3.1), manuscripts and archival documents (section 3.2), and finally photographs depicting works of art (section 3.3). These data exemplify suitable layers of analysis in a domain-oriented perspective, representing different typologies of humanistic data. At the same time this choice is able to describe three different ways of knowledge transmission in a digital environment: a library, an edition and a catalogue. In order to test the ontologies by creating a common semantic model starting from people, dates, places and events, all the collections should reply to the questions: who, where, when, and what.

3.1 Geolat

The aim of Geolat project is to make accessible the Latin literature through a query interface of geographic / cartographic type. The work starts from a digital library that

[17] Linked Open Data (LOD), http://linkeddata.org/

when completed will contain works of Latin literature from its origins to the end of the Roman Empire (conventional date, the 476 d. C.). This stage involves the integration of various already existing repository of Latin texts of high philological quality, which will be integrated starting from their already existing TEI/XML encoding. In a second phase the works so collected are analyzed at morphological level by means of a parser (Lemlat of ILC in Pisa) so as to associate with each word its analysis / morphological description, including a first-level identification of proper names done with NER. A third level of modelling will be tied to the logical relationship between textual references (and their annotations by an encoder) and their referent in the GO ontology (section 2.1).

3.2 Vespasiano da Bisticci's Letters

A digital annotated (XML/TEI) collection of letters from the XV century, sent/received to/by the florentine copyist Vespasiano da Bisticci. The collection is available in a web environment that focuses on: people mentioned in the documents; classical latin and greek manuscripts requested/copied/proposed to/by Vespasiano da Bisticci school and attested in the letters. The original letters are archival documents and manuscripts codices, held by cultural European institutions. All the letters were transcribed, annotated and commented from philological, lexical, historical and prosopographycal points of view.

The purpose of the digital edition is to identify persons related to manuscripts, in order to expose a datasets of people related to manuscripts, these ones described by technical words. A first experiment of using HICO (section 2.2) is a starting point for exploring the potentiality of the ontology in this context, highlighting the network of relations around the cultural object, such as a letter of the collection.

3.3 Zeri Photo Archive

The Zeri Photo Archive, a rich digital catalog today considered one of the most important repertories of Italian art on the web**,** is an archival collection of photographs of paintings stored in a traditional RDBMS.

We started a project for converting data to LOD by adopting a layered conceptualization (namely, CIDOC-CRM) as both a descriptive and a conceptual model. We first proceeded to reengineer the Entity/Relationship model provided by the database tables, which structures data according to the Scheda F[18] (Italian for F entry, a description standard issued by the Central Institute for Cataloguing and Documentation [ICCD] for the cataloguing of photographic materials – where F stands for "Fotografia", photograph in Italian), into an OWL 2 DL ontology, so as to obtain a first version of an ontology, that we call FEO, F Entry Ontology [5].

This was the first phase of a complete reconversion project, that will see the transformation of the data currently stored in the database into RDF statements compliant to a new ontology we are developing, and the use of automatic and semi-automatic tools to generate links to existing datasets. The ontology itself is being iteratively enhanced following modifications of the ICCD Scheda F and of CIDOC-CRM, mak-

[18] Scheda F, http://www.iccd.beniculturali.it/index.php?it/387/beni-fotografici

ing sure that the whole conceptual organization and entity naming of the existing model are affected as little as possible.

4 The faceted browsing environment

The final idea is to use the potentiality of each ontology for exploring the semantic content of all the cultural collections. The aim is to test if ontologies developed for a specific domain could be suitable in other context. At the same time the idea is to reveal possible connections, but also eventual contradictory statements, between data in the collections. If an user searches for a person with a role, connected to a specific place, and a specific date, or an aggregation of all (person, place and date) in an event-oriented perspective, we assume we are able to understand the relationships between the collections: i.e. the same person covering the same role in all the collections, or the same place described in all the collections with different functions.

In this regard, we adopt the term "facet" both as a point of view on digital collections, i.e. as semantic lenses [10] for further clarifications on which quality standards should data achieve, and a way for browsing results of search (categories and filters as multi-level relationships).
In order to achieve these tasks the steps of the projects could be outlined in:

- ontology matching, in order to verify a semantic integration - mostly needed because of the use of different names for describing the same concept and to understand possible different conceptualizations for the same topic -, and to have a common ontology for browsing all the datasets;
- URI attribution, when needed, and normalization of the naming, i.e. URI design related issues and identity resolution among datasets;
- choice of shared authority files for people and places (es. LC authorities[19], VIAF[20], Geonames[21], but also Freebase[22] and Dbpedia[23]);
- XML markup enrichment, when needed, in order to let the ontologies able to extract information in a different context than the starting one;
- definition of ad hoc query, starting from:
  - people's role detectable across collections;
  - description of the same places and relationships involving them;
  - definition of interpretations related to actions (someone do something) and attributions (someone asserted something);
  - events having in common a place, a date and/or a person;
- selection of a faceted environment, in order to expose results of analysis and deploy knowledge discovered [11].

[19] Library of Congress (LA) authorities, http://authorities.loc.gov/
[20] Virtual International Authority File (VIAF), https://viaf.org/
[21] Geonames, http://www.geonames.org/
[22] Freebase, https://www.freebase.com/
[23] DbPedia, http://wiki.dbpedia.org/

5 Conclusions and future works

These steps are required to outline features of each dataset and to define which semantic enrichment is needed in order to benefit of ontology matching: an accurate analysis of the use cases will define more precisely when a model can be reused with minimal efforts (mostly in ontology matching phase) and when a specific-domain model should be considered only to address a restricted research topic. This will produce both an evaluation of ontologies and a refinement of higher-quality data.

Since first trials, it seems obvious that some specific-domain issues have to be explicitly further formalized, representing a great effort in terms of semantic enhancement of starting data, e.g. attestation of roles where none are formally expressed.

As it is conceived, HiCO ontology could represent, without particular further work, a superstructure to describe how places, events, roles and relations described in datasets are bounded to cultural objects, dealing with provenance information and avoiding contradictory statements will affect data consistency.

GO and PRoles, which are mainly devoted to describe above explained specific issues, will be used, where possible, to enrich description of relations in some way formalized but not still exploited for interrogation: e.g. where, in a latin text, a person's role is attested or could be deduced as strictly related to an ancient place, both models can be used to accomplish the descriptive task and we will expect to benefit of inferred information.

The expected result of data reorganization and enhancement will finally lead, through the faceted search, to discover and generalize which requirements are needed when creating datasets in a broad conceptualized perspective, allowing us to formalize a shareable workflow for dealing with data related to cultural objects.

6 Acknowledgments

This research is a result of a stimulating and fruitful collaboration on the single described projects. We would like then to thank all people that worked with us on both ontologies and digital collections: Ciro Gognano, Francesca Mambelli, Silvio Peroni, and Fabio Vitali.

References

1. M. Lana, F. Ciotti, D. Magro, S. Peroni, F. Tomasi, F. Vitali, *Annotating texts with ontologies, from geography to persons and events*, Digital Humanities 2014.
2. F. Ciotti, M. Lana, F. Tomasi, *TEI, ontologies, linked open data: geolat and beyond*, «JOURNAL OF THE TEXT ENCODING INITIATIVE», Issue 8, 2015 [in print].
3. M. Daquino, F. Tomasi, *Ontological approaches to information description and extraction in the cultural heritage domain*, in: *Humanities and Their Methods in the Digital Ecosystem*, F. Tomasi, R. Rosselli Del Turco, F. Rossi, A.M. Tammaro (eds.), New York, ACM 2015.

4. M. Daquino, S. Peroni S., F. Tomasi, F. Vitali, *Political Roles Ontology (PRoles): enhancing archival authority records through Semantic Web technologies*, «PROCEDIA COMPUTER SCIENCE» 38, Elsevier, 2014, pp. 60-67.
5. C.M. Gonano, F. Mambelli, S. Peroni, F. Tomasi, F. Vitali, *Zeri e LODE. Extracting the Zeri photo archive to Linked Open Data: formalizing the conceptual model*, in: Digital Libraries (JCDL), IEEE, London, 2014.
6. M. Lana, *Geolat: Geography for Latin Literature*, in: *ISAW* papers 7, Current Practice in Linked Open Data for the Ancient World Editors: Thomas Elliott, Sebastian (forthcoming) Heath, John Muccigrosso, http://sfsheath.github.io/lawdi-publication/isaw-papers-7.xhtml.
7. M. Daquino, S. Peroni, F. Tomasi, *HiCo, Historical Context Ontology Documentation* (2014), http://purl.org/emmedi/hico.
8. F. Tomasi. *L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere*, Ecdotica 2012, 9 (2013).
9. F. Mambelli, *Una risorsa online per la storia dell'arte: il database della fototeca Zeri*, in: *Digital Humanities: progetti italiani ed esperienze di convergenza multidisciplinare*, a cura di F. Ciotti. Quaderni Digilab, Università di Roma La Sapienza, 2014.
10. S. Peroni, F. Tomasi, F. Vitali, J. Zingoni, *Semantic lenses as exploration method for scholarly article*, in: *Bridging between Cultural Heritage Institutions*, Berlin, Springer Verlag, «COMMUNICATIONS IN COMPUTER AND INFORMATION SCIENCE», 2014, 385, pp. 118-129.
11. M. Pasin, *Browsing Highly Interconnected Humanities Databases Through Multi-Result Faceted Browsers*, Digital Humanities 2011.