Laying the conceptual foundations for data integration in the humanities

Michele Pasin, Arianna Ciula
Centre for Computing in the Humanities
Kings College
London, UK
(michele.pasin, arianna.ciula) @kcl.ac.uk

The purpose of this paper is to promote the discussion on what are the key dimensions of humanities' scholarship, and how they can be best represented by means of formal languages in the context of the Semantic Web. Quite often, available formalizations of knowledge domains and practices in the humanities have been inspired by previous work on more rigorous scientific domains. As a result, we believe that the models thus created tend to oversimplify, if not totally misunderstand, the complexity and peculiarity of the work of humanities' scholars. In this paper, we want to highlight a number of characteristics that need to be taken into account when modeling humanities' data. We argue that only by keeping in mind such requirements we will be able to lay out solid foundations for facilitating non-trivial information integration in humanities domains. We are currently testing these ideas in our department by reflecting upon a number of preexisting digital humanities projects. The final paper will give a more extensive description of this evaluation.

1. Introduction

In recent years we have seen a proliferation of research and commercial projects aiming at the dissemination of a large number of structured or semi-structured data. On the academic side, for example, enterprises such as the Semantic Web (Berners-Lee et al., 2001) have long attempted to support the creation of a vast-scale layer of machine-processable data, which should work as an 'extension' of the traditional web. Less academic examples are instead Freebase (Freebase, 2007), a web application aiming at becoming an "open, shared database of the world's knowledge" which can be freely edited by registered users, and the DBpedia (Auer et al., 2007), a community effort to extract structured information from Wikipedia and make it available on the web by means of a public API[1].

In this paper, we associate these developments in web technologies with

[1] Application Programming Interface, that is, an access point by which such data can be retrieved or manipulated programmatically.

the term '*semantic* web' (SW), as they all share the intent to encode formally (with varying degrees of complexity) aspects of the *meaning* of the resources or artifacts they refer to.

It is worth asking then, why should we as digital humanists be adopting a semantic web approach? A primary advantage of having structured data exposed on the web is the possibility to integrate and reuse them in novel ways. For example, we can imagine a scenario where data coming from an archeological project about Tutankhamun are being accessed by other archeologists interested in pottery produced in Egypt in the same period. Pushing it a little further, we could also think of a research group in sociology of science examining the same data, looking for anomalous patterns in the archeologists' daily data-collection practices.

From an examination of the most recent literature, it is easy to conclude that semantic web technologies have already been tested in a variety of domains. These include both hard science domains, such as physics (Friedland and Allen, 2004), biology (Bechhofer et al., 2006), mathematics (Habel and Magnan, 2007), but also humanities' disciplines such as history of art (Hildebrand et al., 2006), literature (Nowviskie, 2005), music (Schraefel et al., 2005).

However, this spectrum of experimentations leads us to a further consideration. Since scientific domains are highly structured they can more easily be mapped into formal conceptual schemas, so as to be used in SW applications - e.g., a *gene* ontology, or an ontology of *hardware components*. This is not the case for all humanities domains, especially where scholars give high value to processes like the expression of *subjective interpretations* and the *debate* on the subject in question, rather than aiming to search for *objective schemas* or *universal taxonomies*. In other words, the task of modeling knowledge domains in the humanities through formal languages (so as to allow computability and data integration) presents various challenges which are still to be tackled by existing research on the Semantic Web front.[2]

For example, it is our view that systems such as /facet (Hildebrand et al., 2006) or CultureSampo (Eero Hyvönen et al., 2007), although providing advanced interfaces for exploring humanities' data, do not investigate enough the type of semantic 'services' humanities' scholars often engage with in their research practices. In fact, very often such systems make use of very 'shallow' semantic models (e.g., a 'person' who *created* a 'work' which *belongs-to* a 'style'), thus oversimplifying the actual discourse that makes a statement valuable within a humanities discipline. As a consequence, data thus structured can hardly be of use to the humanities scholar in her research and activities.

[2] It has to be noted that some of these challenges have been faced in previous efforts (preceding the advent of the web) of formalizations in the humanities: good examples are the creation of domain-specific thesauri and taxonomies, or the classification systems in library studies.

If data sharing and integration in the humanities is recognized to be worth pursuing, it is therefore necessary to build some solid foundations for a truly useful semantic web framework in the humanities. The first activity that will help us in this respect is a thorough consideration of the typical *entities* and *practices* emerging in humanities' research. Accordingly, in section 3 we outline a number of key requirements humanities' semantic models should support.
In the following section we spend some words on the approach that drives our usage of ontologies for data integration.

2. Ontology: a beauty or a beast?

A central notion in the semantic web and in the world of data integration is that one of *ontology*. The widely used definition by Gruber (Gruber, 1993) describes it as a "an explicit specification of a conceptualization". Being a *conceptualization* an ontology is therefore a *stylized representation* of the world; secondly, since it is expressed in a *formal* language, an ontology can be defined *unambiguously*. As a consequence, ontologies are well suited representation languages for describing data and sharing information; their employment is also endorsed by the W3C (W3C, 2004).

Besides this quite conventional view of what an ontology is, the debate is ongoing about the status of an ontology with respect to the world it represents. For example, some authors such as Smith (Smith, 2003) hold a *realist* position, while others such as the aforementioned Gruber (Gruber, 2003) support a more *pragmatic* view. Such positions affect inevitably the way ontologies are developed and used. For example, in the first case (realist) the implicit assumption is that the ontology should approximate to a 'true' reality; as a consequence, multiple ontologies about the same subject should ultimately converge in their modeling choices. On the contrary, the second class of ontology-design approaches (pragmatist) see an ontology essentially as an engineering artifact: thus, it does not hold any absolute value about the reality it depicts, but it provides a practical solution to the 'problems' it was designed to tackle (i.e. it is a *mean to an end*).

Although in the SW world both approaches have many followers, the context in which digital humanities practitioners and researchers operate, in our opinion, is much closer to the pragmatic approach. Indeed, the humanities are often perceived as the place where all the voices -provided they are respectful of certain argumentative conventions- can be heard, and where all the assumptions can be questioned. Therefore, ontologies for the humanities must support *diversity* and *variety* of *viewpoints*; thus they cannot adhere to an underlying model which neglects multiplicity in favor of a monolithic vision of the world.

Following Gruber (Gruber, 2003), we therefore intend to promote the concept of an ontology as the *agreement* reached by multiple *parties* (e.g., programmers, scientists, collaborators, librarians) with the aim of accomplishing some *objectives* (e.g., data exchange between applications, communication between people, integration of disparate representations). Using a metaphor, ontologies are *contracts*, they are the *currency* used to perform some valuable operations. Thus, their importance is ultimately related not to their truth or beauty, but to the ease they bring to the collaboration among people[3]. To use a less 'commercial' metaphor an ontology is a *compromise* or a point of contact between specific and possibly divergent models. The issue is therefore not only to identify commonalities between projects, for instance, but also to agree that the compromises so found won't diminish the value of the underlying idiosyncratic models, the specificity of any single project or interpretation. We believe that in the humanities this agreement is not necessarily reachable once for all or hoped for, because it may imply the negation of the interpretative efforts that make a work or a project unique and the negation of the evolutionary nature of scholarship. However, we also think that the possibility to make two incommensurable categorical systems communicate could be a challenge worth pursuing.

3. Defining humanities' research

As mentioned above, at a general level it is useful to characterize humanities scholarship by highlighting the points of contrast with the hard sciences. Humanities scholars are traditionally engaged with the expression of interpretative statements and the elaboration of debates on a disparate range of sources of knowledge, rather than with the seeking of firm objective schemas or universal taxonomies.
One of the authors analyzed more specifically the characteristics of a humanities domain (Pasin et al., 2007) -philosophy- and identified those key elements that define its scholarship and make it hard to model Some of these elements are outlined below:

1. **historical** events, that is, events which are inherently time-dependent (e.g., the publication of a book, or an author's subscription to a viewpoint);
2. generic **uncertainty**, that is the frequency of statements about facts which cannot be located exactly in the time and space dimensions (e.g., the birth of Heraclitus);
3. **information objects**, i.e. *texts* in a semiotic sense and especially language-based information objects (e.g. a book), as they are the traditionally preferred medium philosophical contents are expressed with;

[3] Note that we are focusing on the conceptual implications here rather than on the challenges of an ontology implementation by using specific computer languages.

4. **interpretation** events, intended as the process of attributing an abstract content to an information object (e.g., when we say that "Aristotle's fourth book of the Metaphysics states an ontological principle");
5. coexistence of **contradictory** information, which is a direct consequence of 4 (e.g., when people claim different or opposing views on the same proposition);
6. **viewpoints**, and other non-material entities ("philosophical ideas"), for they are the objects philosophers are usually engaged with by studying and expressing them.

Although philosophy has often been defined as the queen of the sciences, these reflections on its nature as discipline may not stand true for the humanities as a whole. In order to highlight all the dimensions that make the modeling of humanities domains such a unique task, we surely need a thorough investigation of other humanities' domains too. Furthermore, for space reasons we have deliberately not mentioned other works in the digital humanities, such as (Jones, 2006) and (Eide, 2008), where the issues tackled are remarkably similar to ours, although the approach is not necessarily ontology-oriented. We intend to elaborate more on these topics in the final version of the paper.

4. Conclusions

In this extended abstract we addressed a number of problems emerging from the employment of semantic web technologies in humanities domains. In particular, we focused on the notion of *ontologies for data-integration,* highlighting the great challenges these technologies will bring *especially* to the digital humanities' practitioner. To this aim we also provided some examples from our previous research in the philosophical domain. In the final paper we will expand this research also by drawing from the results of a detailed analysis of the various projects ongoing in our department. It is our hope that this research will stimulate further discussions and the formulation of a preliminary but comprehensive research agenda.

References

- Auer, S. et al. Dbpedia: A Nucleus for a Web of Open Data. *6th International Semantic Web Conference (ISWC 2007)* (2007).
- Bechhofer, S., Stevens, R. D. & Lord, P. W. Gohse: Ontology Driven Linking of Biology Resources. *Web Semantics: Science, Services and Agents on the World Wide Web* **4**, (2006).

- Berners-Lee, T., Hendler, J. & Lassila, O. The Semantic Web. *Scientific American* (2001).
- Eero Hyvönen et al., CultureSampo-Finnish Culture on the Semantic Web. The vision and first results, in *Information Technology for the Virtual Museum* (ed. Robering, K.) (LIT Verlag, 2007).
- Eide, Ø. The Exhibition Problem. A Real-Life Example With a Suggested Solution. *Literary and Linguistic Computing* **23**, 27-37 (2008).
- Freebase. An open, shared database of the world's knowledge. (2007), Retrieved 20 Feb. 2009, http://www.freebase.com/
- Friedland, N. S. & Allen, P. G. Project Halo: Towards a Digital Aristotle, (2004). Retrieved 20 Feb. 2009, http://www.projecthalo.com
- Gruber, T. It Is What It Does: The Pragmatics of Ontology. *Invited presentation to the meeting of the CIDOC Conceptual Reference Model committee* (2003). http://tomgruber.org/writing/cidoc-ontology.htm
- Gruber, T. R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* **5**, 199-220 (1993).
- Habel, G. & Magnan, F. General Poncelet Meets the Semantic Web: A Concrete Example of the Usage of Ontologies to Support Creation and Dissemination of Elearning Contents. *World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2007* 908–915 (2007).
- Hildebrand, M., van Ossenbruggen, J. & Hardman, L. /Facet: A Browser for Heterogeneous Semantic Web Repositories. *International Semantic Web Conference (ISWC2006)* (2006).
- Jones, A. (ed) *Summit on Digital Tools for the Humanities: Report on Summit Accomplishments*. (2006). Retrieved 20 Feb. 2009, http://www.iath.virginia.edu/dtsummit/SummitText.pdf
- Nowviskie, B. COLLEX: semantic collections & exhibits for the remixable web. (2005). Retrieved 20 Feb. 2009, http://www.nines.org/about/Nowviskie-Collex.pdf
- Pasin, M., Motta, E. & Zdrahal, Z. Capturing Knowledge About Philosophy. *Knowledge Capture (KCAP)* (2007).
- Schraefel, m. c. et al. The Mspace Classical Music Explorer: Improving Access to Classical Music for Real People. *V MUSICNETWORK OPEN WORKSHOP: Integration of Music in Multimedia Applications* (2005).
- Smith, B. *Ontology*, in Blackwell Guide to the Philosophy of Computing and Information (ed Floridi, L.) (Blackwell, Oxford, 2003).
- W3C. OWL Web Ontology Language Overview. (2004). Retrieved 20 Feb. 2009, http://www.w3.org/TR/owl-features/