Web Science: the Digital-Heritage Case

Guus Schreiber

VU University Amsterdam, Computer Science
De Boelelaan 1085, 1081 HV Amsterdam, The Netherlands
E-mail: schreiber@cs.vu.nl; WWW: http://wiki.cs.vu.nl/web-media

Abstract. Web Science studies the interpay between web technology and the human behaviour it induces at the micro, meso and macro level. In this extended abstract we examine Web Science research issues by taking a closer look at the area of digital heritage. We discuss engineering, communication and socio-economic aspects.

1 What is Web Science?

Over the past 15 years the Web has had an increasing impact on the life of people. Web technology has changed the way people operate and communicate, both in their personal and working lives. In many ways the Web is a new phenomenon, for which the principles of the physical world do not always hold. Web science is a new scientific discipline that studies this phenomenon, in particular the interplay between Web technology and the effect it has on human behavior at the personal organizational and societal level.

Good introductions into Web Science can be found in the articles of Berners-Lee *et al.* [1] and Shneiderman [2]. The program and proceeddings of the 1st Web Science Conference in Athens[1] give a good impression of the field. In this extended abstract we illustrate research issues in Web Science by looking at one particular area, namely the digital heritage domain. Digital heritage comprises access to and interaction with large-scale virtual cultural heritage collections. We look at this domain from the experiences gained in a series of digital heritage projects, such as the ongoing work on the Europeana culture portal Europeana[2]. We limit the discussion here to engineering, communication and socio-econmic aspects of digital heritage.

2 Engineering of digital-heritage collections

Cultural heritage is an extremely knowledge-rich domain. Institutions in this field have been gathering knowledge for decades or even centuries. This knowledge is gathered in the form of a multitude of vocabularies, thesauri, classification schemes and other knowledge organization systems, which are used to describe

[1] http://www.websci09.org/
[2] http://www.europeana.eu

heritage objects, such as paintings, books and archival documents. These knowledge organization systems display a enormous richness, but are seldomly described in the formal way favored by computer scientists. The biggest challenge in digital-heritage ventures, such as Europeana, lies in interoperability. The description of heritage objects inherently shows a large variety of perspectives, caused by differences in type of object, time, place, culture and language.

When constructing a web portal for cultural heritage, such as Europeana or E-Culture[3], we are faced with a number of research questions. Firstly, we have to provide mechanisms for explicating heritage knowledge in a machine-readable web format. SKOS[3] is a recently released web standard for achieving this. The design of SKOS reflects some important principles, in particular the principle of "minimal ontological commitment": schema's for publishing knowledge on the Web need to be restricted to the minimum level of required constructs and semantic constraints for these to be usable across the field. For computer scientists with their formal background this is often counterintuitive. Secondly, to enable collection interoperability we need to provide techniques for *partial alignments* between knowledge organization systems. In other words: unification is infeasible in diverse domains such as cultural heritage; the best we can do is uncover the agreement and overlap that does exist. For this reason vocabulary-alignment techniques have become an active area in web research. Thirdly, due to the fact that cultural-heritage descriptions often partially consist of textual descriptions we also require a range of knowledge-extraction techniques, such as from natural-language processing and machine learning. Finally, we have to deal with large amounts of data, typically billions of statements ("the web of data"). This requires scalable search techniques. Given such amounts of data, the traditional notions of correctness and completeness make no sense. Therefore, alternative approaches to reasoning in a web of data are an active area of research[4].

Engineering web data in other domains, such as health care and biology leads to similar research issues. A pervasive common issue is also the notion of web identity. What should be the URI for Pablo Picasso or for the European union? Despite the many research efforts in this area, this is still an open research problem. It is likely that solutions will require some form of societal consensus.

3 Communication in digital heritage

An often-heard opinion in the cultural-heritage field is that the digital experience can never replace the "real thing". In this view digital access is is a surrogate and should ideally be a teaser for the web visitor to come to the museum. This is a lopsided view: the virtual world provides us with alternative and complementary forms of interaction with cultural heritage. An example is the generation of personalized museum tours [4]. Such tours can be first generated for a digital experience and then be downloaded on a mobile device for a physical tour in a museum. The physical tour has limitation ins time and space; the virtual tour

[3] http://www.w3.org/2004/02/skos/
[4] http://www.larkc.eu

has limitations in freedom of experience. It should also be noted that digital techniques help in accessing objects that would otherwise be inaccessible for conservation reasons (e.g. old manuscripts). Thus the virtual and physical visit both have their own pros and cons.

Social tagging is receiving considerable attention in the museum world (see for example the Steve Museum[5]). Tagging is a way to involve web visitors more actively in the collection. Also, given the large amounts of poorly-described objects the heritage institutions are keen on using web visitors for creating object metadata. This raises two issues. Firstly, institutions have to come up with incentive schemes for web visitors to help annotating the objects. Secondly, quality is seen as a key characteristic of curated metdata and is not yet clear which strategies should be followed in quality control of non-curated metadata. This is now an active area of research.

Central in web communication is the issue of user identity and user profiling. Current practice is that user identity is mostly handled at the level of individual web sites or applications. For web visitors this means they have to recreate their identity and their profile in many different places. The control of users over their own profile is limited, as it is usually stored in application-specific cookies. There are several proposals to "put the web visitor back in the driver seat", such as FOAF[6], OpenID[7] and OpenSocial[8]. We expect these mechanisms to change the scenery significantly over the next few years.

4 Social and economic issues in digital heritage

Although heritage institutions, at least in Europe, are almost all public institutions funded with public money, this does not mean that open access to cultural-heritage data should be taken for granted. Projects like Europeana face social barriers in it strive for open access. This is understandable: the heritage institutions have built up their knowledge and data over a long time with an eye on quality control, and are anxious to make this available with the risk of it being used or interpreted in the wrong way. For open access to be become the norm instead of the exception leaders in the field have to set the example. In the library world Library of Congress has done this by making their Subject Headings publicly available in a SKOS format[9]. Major European national libraries, musea and archives are now doing the same within Europeana.

Access to object data, such as images of paintings, still gives rise to rights and authority issues. New license schemes are required. Creative Commons[10] is frequently mentioned in this context, but is not really tailored to the type of heritage field. Similar schemes, but targeted specifically at data collections, have

[5] http://www.steve.museum/
[6] http://www.foaf-project.org/
[7] http://openid.net/
[8] http://code.google.com/apis/opensocial/
[9] http://id.loc.gov/authorities/
[10] http://creativecommons.org/

been proposed. Open Data Commons[11] appears to be a promising candidate for deployment in the area.

The business models for digital heritage depend to some extent on the rights issue. A typical web business model would assume that all primary access to the collections is free, including low and medium-resolution images. Secondary services can be profit-based, such as access to high-resolution images, use of objects within a virtual museum shop (e.g., posters, clothing) and integration with tourist services (e.g. combining heritage access with city walks).

5 Outlook

The web should be viewed as a new ecosystem with an ecology that is in various ways different from the systems we know. It deserves scientific attention from a multidisciplinary angle. Universities are already setting up their first Web Science curricula. The researcher in Web Science works in a new playing field, with new types of interplay between technology and society.

Acknowledgements The ideas in this paper are based on joint work and discussions with many people, in particular colleagues of the VU Semantic Web group, collaborators in research projects, and co-members of the research council of the Web Science Research Initiative[12].

References

1. Berners-Lee, T., Hall, W., Hendler, J.A., O'Hara, K., Shadbolt, N., Weitzner, D.J.: A framework for web science. Foundations and Trends in Web Science **1**(1) (2006) 1–130
2. Shneiderman, B.: Web science: a provocative invitation to computer science. Commun. ACM **50**(6) (2007) 25–27
3. Schreiber, G., Amin, A., Aroyo, L., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Omelayenko, B., van Ossenbruggen, J., Tordai, A., Wielemaker, J., Wielinga, B.: Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. J. Web Semantics **6**(4) (2008) 243–249
4. Wang, Y., Stash, N., Aroyo, L., Gorgels, P., Rutledge, L., Schreiber, G.: Recommendations based on semantically enriched museum collections. J. Web Semantics **6**(4) (2008) 283–290

[11] http://www.opendatacommons.org/
[12] http://webscience.org/