Available online at www.sciencedirect.com

SciVerse ScienceDirect

Int. J. Human-Computer Studies 71 (2013) 206–210

International Journal of Human-Computer Studies

www.elsevier.com/locate/ijhcs

Knowledge acquisition and the web

Guus Schreiber

Computer Science,The Network Institute, VU University Amsterdam, The Netherlands

Accepted 25 September 2012
Available online 24 October 2012

Abstract

Knowledge-acquisition research started in the eighties as a small research community focusing on knowledge-intensive problems in relatively small domains. In this paper we look at the influence the Web has had on knowledge acquisition and *vice versa*. To this end we discuss in some depth four topics, namely the ontology language OWL, the vocabulary language SKOS, the notion of ontology alignment and the potential of semantic search. Even from this limited selection of research issues related to "Web knowledge" it is safe to conclude that the Web has had a large impact on knowledge acquisition, but also the other way around.

Keywords: Web-based knowledge acquisition; Future of knowledge acquisition; Web ontologies

1. Introduction

The knowledge acquisition (KA) community started meeting in 1986 at the first Knowledge Acquisition Workshop in Banff, Canada. Since then yearly meetings have taken place in North America, Europe and Asia. During the first 15 years the meetings were on purpose small-scale: 40–60 researchers. The community sense was strong but at the same time there were often debates about whether the KA community was having enough impact, i.e. "are we being heard outside?". The purpose of this contribution is to analyse the actual impact the community has had on the major technological development since 1986, namely the World-Wide Web.

This impact analysis does not have the pretense of being complete. Rather, this paper focuses on a selected set of Web technologies directly related to KA, in particular Semantic Web technologies. We discuss four topics: (i) the development of the Web Ontology language, OWL, (ii) publication of Web vocabularies, (iii) ontology alignment; and (iv) semantic search. The selection is biased by the background and personal experiences of the author. Still, looking back it appears fair to derive two conclusions: first that the Web has provided the KA community with a medium for knowledge acquisition at a scale we had not believed possible when the first KA workshops started, and second that KA research has significant impact on how the Web is shaping up.

2. Knowledge acquisition on the Web: selected topics

2.1. OWL

OWL (McGuinness and van Harmelen, 2004) was developed from an amalgamation of the DARPA-funded DAML language (Hendler and McGuinness, 2000) and the EU-funded OIL language (Fensel et al., 2000). Although OWL is often seen as a product of the knowledge-representation community, this is only partially true.

First and foremost, OWL is an *ontology* language. In essence, the use of the notion of ontologies in computer science is derived from the knowledge-level hypothesis of Newell (1982). This hypothesis states that knowledge should be represented at a level independent from particular implementation-level details. The KA community adopted this principle early on and was the key community in which ontology engineering was studied in its modern, pragmatic form: ontologies are not general theories of knowledge, but rather pragmatic and reusable specifications of concepts that represent a consensus view in a particular domain. The VT-Sisyphus experiment, initiated in the early 1990s by Gruber and others (Schreiber and

E-mail address: guus.schreiber@vu.nl

1071-5819/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ijhcs.2012.10.001

Birmingham, 1996), is, to the author's knowledge, the first comprehensive experiment in ontology reuse. It took until the end of the 1990s for ontologies to become fashionable within computer science at large. The most popular ontology editor to date is Protégé, developed and maintained by Musen, Noy et al. (2001) at Stanford University; the first version stems from their KA work in the early 1990s.

The results of the W3C Web Ontology Working Group,[1] which published the first OWL standard, were significantly influenced by KA researchers (van Harmelen, Fensel, Motta, Schreiber). This helped to prevent OWL from becoming a language only tailored to theoretical requirements from knowledge representation, in particular description logic. The documented working-group discussions and use cases provide evidence for this. OWL takes pragmatic knowledge-acquisition principles into account. For example, the hard requirement of allowing meta-classes and meta-properties was put forward as essential for real-world modeling in the diverse Web world with many different perspectives. This resulted in OWL having two types of possible semantics: the strict description-logic OWL DL semantics and the open-ended OWL-Full semantics.

Looking back, the take-up of OWL shows that this decision was probably the best one. There are domains which lend itself well to OWL-DL style modeling, such as medicine (see e.g., the work of Rector et al., 2004), but other uses show the need for a more "scruffy" approach to knowledge modeling (e.g., our own work in virtual heritage collections, Schreiber et al., 2008). In many cases pieces of OWL have been picked up for Web usage, not the language as a whole. The best example is the wide-spread use of `owl:sameAs` in Linked Data. The use of DL-type reasoning appears to be mainly limited to ontology validation.

2.2. Web vocabularies

In many domains, experts have been writing down descriptions of the domain concepts, typically in a hierarchical taxonomy-like fashion. This work has been going on for decades, in some areas for centuries. Typical examples of such vocabularies are disease classifications, library subject headings, biological taxonomies and lexical terminologies. With the advent of the Web and of Linked Data these knowledge sources have become important assets. Domain vocabularies can be used for describing and indexing Web data, as these resources typically represent a consensus about classifying terms in a particular field of interest.

SKOS (Simple Knowledge Organization System), which was published as a Web standard in 2009,[2] was developed as a model for publishing such vocabularies on the Web. SKOS provides a set of meta-concepts to model a vocabulary in RDF/OWL (Miles and Bechhofer, 2009). One of the early adopters of SKOS was the US Library of Congress which published their complete set of Subject Headings, used world-wide to index books. Many organizations responsible for vocabularies have since then done the same. This has made a huge source of structured knowledge available.

KA research has had a significant influence on both the standardization of SKOS and on its subsequent deployment. The SKOS meta-model was designed with the "minimal ontological commitment" strategy, propagated by Gruber (1994), in mind: include only those distinctions in an ontology that are absolutely necessary. This has greatly helped the wide deployment of SKOS. To outsiders SKOS looks really easy to use; it fits their mental model of a vocabulary. For example, in the large-scale Europeana effort[3] to make all EU museum, archive and library data available online the process of converting the omnipresent heritage vocabularies to Web formats is now colloquially called "skossification".[4] KA researchers have provided methods and tools for supporting this conversion process.[5]

One type of vocabularies deserves special attention, namely lexical terminologies. Princeton's WordNet (Miller, 1995) is the prime example of such a vocabulary. This vocabulary contains a comprehensive set of terms and concepts in the English-American language in the form of one hierarchy with many different types of additional relations, such as part-of relations. WordNets have by now been developed for dozens of other languages. The importance of such resources on the Web is enormous. For example, concepts from Princeton WordNet are used to provide types for Wikipedia articles. Princeton WordNet was in fact one of the very first pieces of Linked Data published on the Web. This work was done by KA researchers (van Assem et al., 2006) and has served as a source of inspiration for many others. As an aside, it should be noted that the Web version of WordNet was used successfully by the IBM Jeopardy system.[6]

2.3. Ontology alignment

In the 1970s and 1980s ontology alignment was not a major research topic in the KA community. Knowledge acquisition focused on relatively small and closed areas. The Web has changed all this. Nowadays, we are confronted with many different knowledge sources, such as SKOS vocabularies that represent key concepts in different but sometimes related domains. For example, if we have Web data of two different libraries that have indexed their books with two different vocabularies, then we are likely to

[1] http://www.w3.org/2001/sw/WebOnt/
[2] http://www.w3.org/2004/02/skos/
[3] http://www.europeana.eu
[4] For one of may examples, see https://www.ocs.soton.ac.uk/index.php/CAA/2012/paper/view/678
[5] See, e.g. http://code.google.com/p/skosify/
[6] http://www.heatonresearch.com/content/free-and-open-software-behind-ibm%E2%80%99s-jeopardy-champion-watson

have a keen interest in semantic relations between the two vocabularies, as this will enable us to provide a combined search. Both OWL and SKOS provide built-in constructs for such alignments, the best known being the aforementioned `owl:sameAs` and also `skos:closeMatch`.

Ontology alignment[7] has therefore become a primary focus of attention in KA research. Since the middle of the previous decade a large number of papers has been published on this subject and initiatives such as the OAEI (Ontology Alignment Evaluation Initiative[8]) have seen the light of day. The proceedings of K-CAP and EKAW are evidence of this trend.

Despite the many research efforts, progress in this area is still hard. By its nature ontology alignment is a difficult problem. For example, Halpin et al. (2011) have shown that `owl:sameAs` is often used in a way that is inconsistent with its formal definition in OWL. One central methodological problem concerns the way to set up an evaluation study. For example, say you want to align WordNets from two different languages. Each language is based on a underlying cultural, societal and historical frame of reference. Even if those cultures happen to be closely related, concept alignment is still non-trivial. For example, the French have many different terms for food, both abstract and concrete. The British, on the other hand, have a much more limited vocabulary for the same domain.

Ontology alignment brings us to the edges of the prevailing paradigms of modern knowledge representation, in particular the view of a class or concept as a set of instances with clear boundary. When you ask in an evaluation experiment domain experts about the nature of the alignment relation between two concepts, it is often impossible for them to give an unequivocal answer. This is because we ask them to think in categories with precisely defined borders, while in practice people use concepts in an approximate way. Studies in cognitive psychology have shown that people think of categories as prototypes: instances are classified based on their relative distance to categories (Lakoff, 1987). This observation poses significant methodological problems for the evaluation of ontology-alignment techniques (Tordai et al., 2011). As it stands, ontology alignment is likely to continue to be an important research area for KA researchers in the near future.

2.4. Semantic search

Web search is still dominated by statistical information-retrieval methods. An open question for the KA community is still whether there is a role for explicit knowledge in Web search. The assumption that explicit knowledge about Web resources facilitates Web search lies at the heart of the notion of a Semantic Web. Some search engines are already deploying semantics in *presenting* search results. But can search itself be improved with semantics? The answer to this question still remains unclear. A number of applications submitted to the Semantic Web Challenge[9] have shown the usefulness of semantic search to some extent. Wolfram Alpha[10] and Siri[11] are examples of applications that use semantic search. The major search engines have also made significant investments to deploy semantics, such as the "schema.org" efforts.[12] An extensive survey of existing research efforts in this area is outside the scope of this paper, but it is safe to say that this is still an area in its early days. In this paper we limit ourselves to proposing two lines of approach that could help moving semantic search forward:

- Semantic-search should target knowledge-rich domains.
- Semantic search should focus on other types of search than the prevailing keyword-based paradigm.

Semantic search in knowledge-rich domains: In semantic search we try to deploy domain knowledge, attached via RDF annotations to Web resources. Adding metadata leads to a graph of interlinked Web data and associated concepts. It is our contention that it is difficult to put such graph search to use on a general Web scale, due to diverseness of perspectives and the heterogeneity of Web data. However, this does not mean that such types of search cannot be of use in (large) subdomains such as medicine and cultural heritage. In the medical area domain expertise plays a key role: you do not want "approximate search" on the best treatment for a disease; search results need to be precise and up-to-date. With the growth of the Web the need for domain-specific search engines is likely to increase.

Another example is a domain-specific engine for art works. With so many musea making images of their works plus associated metadata available online (see, e.g., the aforementioned Europeana effort) it becomes relatively easy to provide services that can answer questions such as "Give me art works depicting Moulin de la Gallette in Paris" (see the sample search results in Fig. 1).

Different search paradigms: We are now used to the keyword-style Web search. However, we should not be blinded by the success of this search paradigm. We should also try out other types of queries that may be better suited for the use of explicit background knowledge. One example that has been mentioned is what one could call *relation search*: "given two objects O1 and O2, find relations between these objects". An example of relation search that we have worked on in the art domain is shown in Fig. 2.

[7] We leave the debate whether vocabularies can be called ontologies here aside.

[8] http://oaei.ontologymatching.org/

[9] http://challenge.semanticweb.org/

[10] http://www.wolframalpha.com

[11] http://tomgruber.org/writing/semtech09.htm

[12] http://schema.org/

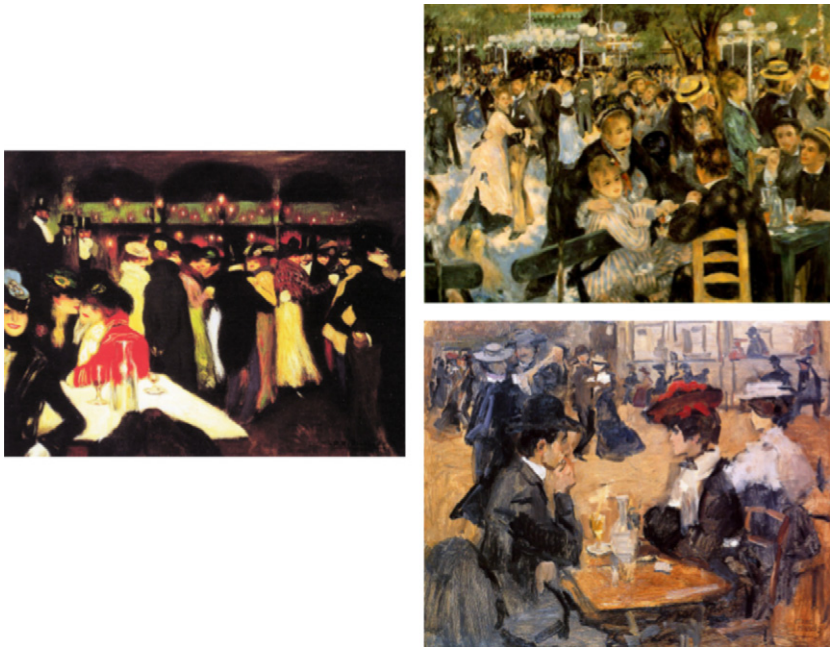

Fig. 1. Three paintings resulting from a semantic search for an art object depicting the same place. The place here is Moulin de la Galette in Montmartre, Paris. The painting at the left is an early work by Picasso, showing the inside of the place. The upper right shows the famous painting by Renoir of the terrace. The lower-right image is another painting of the inside by the Dutch artist Isaac Israels. Both the images and the metadata are nowadays available on the Web.

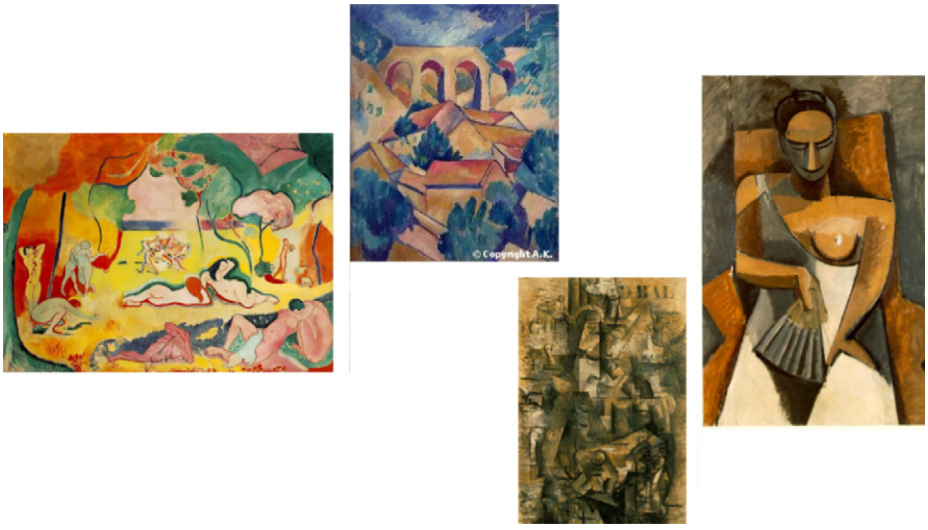

Fig. 2. Result of a semantic search to provide one possible answer to the query "How is Matisse related to Picasso?". Here, this question is answered by showing that George Braque painted in two art styles strongly connected to the artists mentioned in the query: the Fauve ("wild beast") style of Matisse on the left and the cubist style of Picasso on the right.

When answering a query like "What are the relations between Matisse and Picasso" is it relatively easy to come up with the result shown in the figure, as online sources such as the Union List of Artists Names (ULAN)[13] provide all the background knowledge needed (Georges Braques, another famous painter, was influenced by both painters; check the relations in ULAN).

Relation search requires algorithms for graph traversal that use knowledge patterns to come up with potential results (Nuzzolese et al., 2011). The notion of knowledge patterns was in the 1980s and 1990s a strong feature of KA research (e.g., the patterns of knowledge-intensive tasks). We see a new role for such pattern-focused research.

3. Discussion

The fear of the KA community for "not being heard outside" has proven to be unfounded. The Web has given the community a space in which knowledge acquisition has become an activity at an unprecedented scale. This is in all

[13] http://www.getty.edu/research/tools/vocabularies/ulan/

likelihood also the reason why so many of the researchers from the community have become influential in the Web field.

Wikipedia is a good example of the relevance of knowledge acquisition on the Web. Over the years Wikipedia pages have become more structured using categories of pages with predefined attributes, thus forming a class structure of information resources. Linked Open Data are another example of a fast growing knowledge base. Such bodies of knowledge and information provide immense opportunities for developing novel knowledge-acquisition theories, methods and tools.

Acknowledgments

This work is supported by the COMMIT Project funded by the Dutch Government.

References

Fensel, D., Horrocks, I., van Harmelen, F., Decker, S., Erdmann, M., Klein, M. 2000. OIL in a nutshell. In: Knowledge Engineering and Knowledge Management: 12th International Conference EKAW2000, Juan-les-Pins. Lecture Notes in Artificial Intelligence, vol. 1937. Springer-Verlag, Berlin, Heidelberg, pp. 1–16.

Gruber, T.R., 1994. Towards principles for the design of ontologies used for knowledge sharing. In: Guarino, N., Poli, R., (Eds.), Formal Ontology in Conceptual Analysis and Knowledge Representation. Kluwer, Boston.

Halpin, H., Hayes, P., Thompson, H., 2011. When owl: sameAs isn't the same redux: a preliminary theory of identity and inference on the semantic web. In: Workshop on Discovering Meaning on the Go in Large Heterogeneous Data 2011 (LHD-11), Barcelona, Spain, 16 July 2011, pp. 25–30.

Hendler, J., McGuinness, D., 2000. The DARPA agent markup language. IEEE Intelligent Systems 15 (6), 67–73.

Lakoff, G. (Ed), 1987. Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. University of Chicago Press.

McGuinness, D., van Harmelen, F., 2004. OWL Web Ontology Language Overview. W3c Recommendation, World-Wide Web Consortium.

Miles, A., Bechhofer, S., 2009. SKOS Simple Knowledge Organization System Reference. W3c Recommendation, World-Wide Web Consortium.

Miller, G., 1995. WordNet: a lexical database for English. Communication of ACM, 38(11), November.

Newell, A., 1982. The knowledge level. Artificial Intelligence 18, 87–127.

Noy, N., Sintek, M., Decker, S., Crubézy, M., Fergerson, R., Musen, M., 2001. Creating semantic web contents with protégé-2000. IEEE Intelligent Systems 16 (2), 60–71.

Nuzzolese, A., Gangemi, A., Presutti, V., Ciancarini, P., 2011. Encyclopedic knowledge patterns from Wikipedia links. In: The Semantic Web—ISWC 2011. Tenth International Semantic Web Conference, Bonn, Germany, 23–27 October 2011, Proceedings, pp. 520–536.

Rector, A., Drummond, N., Horridge, M., Rogers, J., Knublauch, H., Stevens, R., Wang, H., Wroe, C., 2004. Owl pizzas: practical experience of teaching owl-dl: common errors & common patterns. In: Engineering Knowledge in the Age of the Semantic Web, 14th International Conference, EKAW 2004, Whittlebury Hall, UK, 5–8 October, Proceedings, pp. 63–81.

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., 2008. Semantic annotation and search of cultural-heritage collections: the MultimediaN E-Culture demonstrator. Journal of Web Semantics 6 (4), 243–249.

Schreiber, G., Birmingham, W.P., 1996. The Sisyphus-VT initiative. International Journal of Human–Computer Studies 43 (3/4), 275–280 (Editorial special issue).

Tordai, A., van Ossenbruggen, J., Schreiber, G., Wielinga, B., 2011. Let's agree to disagree: on the evaluation of vocabulary alignment. In: Proceedings of the Sixth International Conference on Knowledge Capture (K-CAP 2011), 26–29 June 2011, ACM, Banff, Alberta, Canada, pp. 65–72.

van Assem, M., Gangemi, A., Schreiber, G., 2006. RDF/OWL Representation of WordNet. Technical Report, World-Wide Web Consortium W3C, 19 June.