Semantic Web 1 (2010) 25–32
DOI 10.3233/SW-2010-0001
IOS Press

The role of space and time for knowledge organization on the Semantic Web

Editor(s): Pascal Hitzler, Wright State University, USA
Solicited review(s): Christoph Schlieder, University of Bamberg, Germany
Open review(s): Pascal Hitzler, Wright State University, USA

Krzysztof Janowicz
GeoVISTA Center, Department of Geography, The Pennsylvania State University, USA
E-mail: jano@psu.edu

Abstract. Space and time have not received much attention on the Semantic Web so far. While their importance has been recognized recently, existing work reduces them to simple latitude-longitude pairs and time stamps. In contrast, we argue that space and time are fundamental ordering relations for knowledge organization, representation, and reasoning. While most research on Semantic Web reasoning has focused on thematic aspects, this paper argues for a unified view combining a spatial, temporal, and thematic component. Besides their impact on the representation of and reasoning about individuals and classes, we outline the role of space and time for ontology modularization, evolution, and the handling of vague and contradictory knowledge. Instead of proposing yet another specific methodology, the presented work illustrates the relevance of space and time using various examples from the geo-sciences.

Keywords: Semantic heterogeneity, ontologies, context, space and time, sensors and observations, geospatial semantics

1. The beauty of semantic heterogeneity

Overcoming semantic heterogeneity is a core topic of many contributions to such diverse research fields as semantic interoperability, semantics-based information retrieval, service composition, the Sensor Web, and ontology engineering. But do we really want to *overcome* or *resolve* semantic heterogeneity and what would this mean for the Semantic Web?

In contrast to syntactic heterogeneities caused by differences in data types, signatures, and protocols, semantic heterogeneity refers to differences in the intended interpretation, i.e., meaning, of information. While homonyms or polysemes are classical linguistic examples, semantic heterogeneities in information science tend to be more subtle. Web service interfaces to weather stations offer an impressive example. Two services can return a string called *wind direction* as output and even specify that the results are numerical values ranging from 0–360° but still have a contradicting interpretation of *wind direction*. For instance, one service refers to *wind blows to* while the other adapts a *wind blows from* semantics. Combining the results to compute the dispersion of a toxic gas plume would lead to meaningless and potentially dangerous results [32]. It seems obvious that such incompatibilities need to be resolved by overcoming semantic heterogeneity. While this is true in many cases and at the core of classical data integration, it may not be the most appropriate solution for a Web following the AAA slogan[1].

Consider the following simple, yet entertaining example of potholes in the UK [23]. Due to a severe winter millions of potholes need to be repaired by the local councils that are legally responsible for the maintenance of roads within their administrative boundaries. While potholes are defined as cracks of more than 30mm depth in North East Somerset, they must be of the width of a 'large dinner plate' (300mm) and

[1] **A**nyone can say **A**nything about **A**ny topic [1].

the depth of a 'golf ball' (40mm) in Gloucestershire. Worcestershire, in contrast, defines potholes by the width of a smaller 'dinner plate' (200mm) with a minimum depth of a 'fist' (40mm). In Coventry, a pothole can be reported by citizens if its depth is 'a pound coin and a 1p coin side by side'. These, and many other, councils have different conceptualizations of the term *pothole* for good reasons – probably because of the budget they would need to invest in fixing them. Consequently, it is unlikely that they want to resolve semantic heterogeneity in the first place.

Assuming one knows all local definitions of pothole and all potential cracks in roads, answering the question how many potholes are there in Britain becomes a complex, yet feasible task. In contrast, the question of how many lakes are there in Minnesota, USA cannot be answered this way. While the Department of Natural Resources lists 11,842 lakes over 10 acres, *lake* is a vague concept by nature. Its intended interpretation is not restricted to a degree which would allow to decide whether a water body is a single lake or two lakes connected by a watercourse, or how to distinguish them from ponds [28]. In fact, many size-based definitions take 5 acres as criterion [41]. One could argue that the size of a lake is all it needs for its definitions, but a flooded grassland is not a lake while a temporarily dry basin may still count as lake.

What appears to be an academic exercise only is, in fact, a common problem in cross-border Spatial Data Infrastructures (SDI). To query and exchange data between administrative units or states requires to take local conceptualizations into account. Similarly, in most cases the reuse of sensor data fails due to different measuring practices and requirements and, hence, data is collected again and again. One approach to ensure that, e.g., a forest does not stop at a state's border and continues as (wood) pasture on the other side just because of varying definitions of *forest*[2] is top–down standardization. The Infrastructure for Spatial Information in the European Community INSPIRE is such a large scale standardization endeavor aiming at cross-scale, cross-language, and cross-border interoperability and access to geo-data.

However, creating top–down definitions of geographic feature types bears the danger of excluding local definitions [8]. To take yet another example from geography, the European Water Framework Directive defines *river* as '[a] body of inland water flowing for the most part on the surface of the land but which may flow underground for part of its course'. Simplifying, European member states have to encode their data according to such global schemata. Nevertheless, rivers in Southern Europe may lack any flowing water for long parts of the year. Therefore, the local definitions may contradict with the global schemata.

While the previous examples involved space as a criterion for their variety, the following example also involves time as driving force. One key concept in ecology is *succession*. It describes the ordered, sometimes cyclic sequence of changes resulting in transitions between ecological communities within the same geographic location. An often cited, cyclic example are beaver dams. By changing the water flow of streams they create ponds in forested areas. These ponds repress the trees, hence, change the composition of the habitat and, therefore, may not offer the right food sources for beavers anymore. Such ponds will then be abandoned by the beavers and dry out again. The resulting meadows form yet another habitat with optimal conditions for plants requiring more direct light. However, they will turn back into forested areas on the long term and serve as beaver habitats again. From an ontological point of view, this raises several questions about how to define identity criteria for such places and how to model them. We can define the state before the stream is dammed and after the pond is replaced by a meadow; the question from which point in time a stream *segment* becomes a pond and how much water is required to distinguish the pond from a meadow is more difficult. Finally, if the cyclic succession at the same location creates ponds again and again, are these ponds the same entity[3]?

Summing up, one reason for the success of Semantic Web technologies in life sciences such as medicine is based on canonical definitions. While we can define a human hand as having five distinguishable fingers in a specific order[4], the above examples illustrate that there is no context free definitions of lakes, forests, and many other geographic feature types. Context however, as will be discussed in the following, is largely determined by space and time. With respect to the question of overcoming semantic heterogene-

[2] and there are, for various reasons, several hundred local definitions of forest [25].

[3] This argument should also be kept in mind when arguing for *Web of Things* related approaches to grounding, e.g., by assigning URIs to real world entities.

[4] and we consider deviations such as caused by Polydaktylie, i.e., having supernumerary fingers, as deformities.

ity, the introduced examples illustrate the need for a change in perspective. Namely, shifting from resolving heterogeneity to accounting for it and acknowledging the importance of local conceptualizations by focusing on negotiation and semantic translation. In previous work, we have discussed how semantic similarity can be used to estimate how accurately an ontology captures the user's initial conceptualization [18]. Such work could also be used for the negotiation of semantics on the Web.

2. Contexts and concepts

Categorization is an essential prerequisite for interacting with and reasoning about our environment. Nevertheless, there is no a priori conceptualization of the world and the creation of entities and types is an act of cognition and social convention [24,26]. The decision of how to carve out fields of sensory input depends on context, i.e., factors such as cultural background, previous knowledge, language, personal goals, the current situation, and especially also on space and time. What is a *deep lake* for recreation may be a *shallow pond* for navigation purposes; see also [35]. In fact, conceptualization is the act of introducing distinctions for certain needs – making these decisions explicit in a formal way, i.e., constraining their interpretation, is what ontology engineering should be about. Concepts and relations between them are not fixed but emerge from the context [9].

The importance of contextual information has been widely recognized in information retrieval; which role does it play for the Semantic Web? Today, the Web is essentially still about documents and fixed links between them. These documents encapsulate information by providing structure and context for the inherent data and, hence, support their interpretation. The forthcoming Data Web, however, is about linking data, not documents. Data sets are not bound to a particular document but can be freely combined outside of their original creation context. In theory, users can query the Linked Data cloud to answer complex queries spanning multiple sources and establish new links between data on-the-fly. However, retrieving meaningful results is more difficult than one may expect. While uncoupling data from documents eases their accessibility it puts the burden on their interpretation.

Data is always created for a particular purpose, even if it may be as broad as the creation of a free and collaborative encyclopedia such as the Wikipedia. Consider the following gedankenexperiment as illustration: do all appearances of a particular term in the Wikipedia conform with its definition in the according Wikipedia main article? For instance, the article about *time* is based on modern physics, while the term is used in a colloquial way throughout hundreds of thousands of Wikipedia articles. If a future DBpedia version would capture more data from Wikipedia then it does so far, would it assign the same ontological concept *time* to all of them? To a certain degree DBpedia already faces such problems. For instance, searching for actors may be done using rdf:type 'Actor' or by the relation 'occupation' with the filler 'actor' – both SPARQL queries produce overlapping but different result sets.

Revisiting the examples in the previous section also illustrates that similar difficulties arise for the creation of meaningful URIs for Linked Data. Entities are often constructed based on social convention and differ between information communities or even individuals. *Downtown* or other vague regions may act as examples [29]. If we do not want to end up in assigning URIs to single pixels of raster data or the whole swath width of sensors, we need to make some choices about how to extract entities, e.g., points of interest, from datasets. These choices are arbitrary to a certain degree and should therefore be encoded in the URI. As recently discussed by Halpin and Hayes, using *owl:sameAs* for identity links is not sufficient and may be even misleading [15].

How do humans establish communication if the meaning of terms is influenced or even determined by local contexts? Leaving the physical layer, e.g., the cortex and the role of mirror neurons aside, substantial work from cognitive science argues for a situated nature of categorization [2]; see also [6]. Instead of rigid and pre-defined conceptualizations with clear boundaries, many concepts arise by simulating situations. A classical example are so-called ad-hoc categories such as *things-to-extinguish-a-fire* which include such diverse entities as bed sheets and water. The function of artifacts, for instance, may be best understood in terms of the HIPE theory, i.e., by their History, Intentional perspective, the Physical environment, and Event sequences [3].

Humans can interact not because they share the same conceptualizations but because they can make sense of each other's statements by putting them into context. Meaningful communication, i.e., semantic interoperability in terms of the Web, can be established as long as the consequences, e.g., actions, of our

counterpart are consistent and meet our expectations [13,33]. For instance, while *hill* and *mountain* may have clearly defined distinctions of social importance[5], they are irrelevant for many everyday situations such as agreeing to climb its peak. The reason why we are not confused when our counterpart uses the term *mountain* for what we would call a *hill* is that the potential interpretations of the terms are sufficiently restricted by the sentence or the context, e.g., the surrounding landscape. Where and when we use a term restricts its interpretation towards the intended model. Consequently when describing the nature of the Web, the AAA slogan may be more appropriately described as AAAAA slogan – **A**nyone can say **A**nything about **A**ny topic at **A**ny time and **A**nywhere.

Acknowledging the role of context for conceptualization and the importance of the resulting heterogeneities sheds new light on the vision of establishing ontologies for the Web. Learning from the success of user generated content on the social Web, a promising approach would be to support users in becoming active knowledge engineers instead of trying to develop de-contextualized ontologies top–down. A set of building blocks and tools could support information communities in specifying their local conceptualizations. Semantic annotations should connect local ontologies with Linked Data on-the-fly.

How to define such building blocks without falling into the symbol grounding trap, i.e., how to avoid an endless regress? While this topic is too complex to be discussed here, embodiment seems to be a crucial part of potential solutions [39]. Humans do not share the same conceptualization of the world but commonalities can be established by fundamental properties of our bodies and sensor systems. Experiences of surfaces, containment, paths, center-periphery, blockage, and many more are shared as they are directly observable based on our bodily interaction with the environment [20,21]. It is interesting to note how many of these so called *image schema* have spatial roots. A prominent linguistic examples illustrating the same argumentation are spatial and temporal metaphors [22]. Gibson's notion of Affordances [10], i.e., action possibilities arising from the combination of the actor's physical properties and those of the environment is another approach, and has been recently used to demonstrate how to ground geographic categories in observations [36] as well as for robot control. Strictly speaking, one may object that the argumentation provided above requires an inter-subjective *stimulus meaning* and a similar sensory reception. However, as argued by Quine, we can stay with private stimulus meanings as the inter-subjectivity is provided by the use of language [33].

The ontological question of *what is there* bears the danger of introducing entities and fixed types in an early stage instead of focusing on *observation categoricals* [33]; see also [37] for the construction of bodies from observations. Ontologies should act as bridges between the continuous fields of sensor-based observations, numerical models, and the rather entity-centric use of language. Highlighting the importance of observations does not exclude social aspect of semantics. With respect to the pothole example, all local definitions share an observable component – a depression in the road – while the required size is a matter of social convention and negotiation.

Space and time are two of the most fundamental ordering relations used in human cognition, language, and even on the physical level in the formation of patterns inside the human cortex. While we may not agree on the definition of *chair* by referring to shape, size, the number of legs, or the existence of a backrest – we can reach agreement in stating that their surfaces offer support and hence *sitability*. To demonstrate the impact of space on categorization[6], another approach to understand whether a visually perceived object is a chair is via its *position* relative to a table, bin, or other objects. Context and categorization influence each other mutually. While entering an unfamiliar building we constantly make predictions on what we expect to encounter [16]. Once we have identified a room as office by recognizing tables though stacks of paper *placed on* them, a partially visible gray box positioned *under* a table (that could not be categorized before) is likely to be a computer. In other words, unfamiliar objects can be categorized based on their *place* and at the same time give feedback about whether our assumptions about the current context were appropriate. If we cannot identify chairs in the room, the office hypothesis may need to be revised. This could also affect the interpretation of other objects categorized before. Personal information managers (PIMs) use this fact to split to-do lists based on contexts such things to do *at* the office, home, or during travel.

[5] Cineastes are referred to the movie *The Englishman Who Went Up a Hill But Came Down a Mountain* for an example.

[6] Marked *italic* in the following sentences.

Summing up, to ensure the meaningful usage of (linked) data requires to restrict their potential interpretations. Ontologies are one method to make the underlying distinctions explicit but depend on context themselves. The attempt to develop stable and global ontologies contradicts with the nature of the Web. While this section illustrates the role of context and situated concepts, various approaches have been proposed in the knowledge representation and reasoning literature within the last decades – recent examples include C-OWL [5] or Bennett's notion of standpoint semantics [4]. Embodiment, sensors, and observations are crucial elements for establishing common building blocks to align or translate between user-contributed ontologies. To combine two buzzwords: a promising approach for the future may be to ground the Semantic Web in the Sensor Web.

3. Giving order by space and time

While the previous section focused on knowledge representation, this section describes how to structure and organize concepts and ontologies. Since the impact of context is not random, reasoning about and building bridges between local ontologies requires a meta-theory explaining which kind of contextual information matters, which refines, and how context causes diversification. In contrast to Semantic Web research, understanding the user's context and trying to infer implicit information out of it is a central task in information retrieval and related areas.

The challenge of handling local conceptualizations at a global level is a prominent topic in artificial intelligence research since decades. One core idea is to be consistent at the local level but allow contradicting conceptualizations within the global knowledge base. One approach is to organize knowledge in domain specific *microtheories* (also called contexts) and has been used in OpenCyC. Each microtheory is developed as a coherent set of statements and can be thought of as a single ontology; see also work on ontology modularization [12]. Separate microtheories may hold information about the same concept but contain incompatible facts. Using the time example introduced above, one microtheory may be more precise and rigid with respect to physical properties and laws of nature, while another microtheory may be based on weaker constraints to support *naïve physics* [17].

Microtheories are organized in subsumption hierarchies, i.e., facts specified in the super-microtheory must also hold in each of its sub-theories. In contrast, sibling-theories can store contradicting facts. More formally, the hierarchy of microtheories is established through the generalization relationship $genlMt$ [27]. Given $ist(mt, p)$ is the *is true in* relation between a microtheory mt and a predicate p, then $genlMt$ is the anti-symmetric, reflexive, and transitive binary predicate by which the theory hierarchy is constructed by adding axioms of the form

$$mt_0 : \forall p\ ist(mt_g, p) \wedge genlMt(mt_g, mt_s) \longrightarrow ist(mt_s, p)$$

to the topmost theory mt_0; where mt_g is the more general and mt_s the more specific theory; see also [14] for details.

Surprisingly, alternative ordering principles based on space, time, or cultural background have not been discussed so far. While the previous sections illustrate the impact of climatic, geological, ecological, administrative, and further factors on the categorization of geographic feature into types, this impact does not occur randomly but follows gradually changing patterns[7]. In other terms, using Tobler's famous First Law of Geography: 'Everything is related to everything else, but near things are more related than distant things' [40]. For instance, the definition of *river* changes gradually from northern to southern European countries. Similarly, temporal examples can be found in the domain of cultural heritage research which has to deal with incomplete, biased, and contradicting information. For instance, beliefs about the solar system from the Middle Ages may be organized in a different branch of a knowledge base than microtheories describing beliefs from the age of industrialization. To structure microtheories by spatial (or administrative) containment $genlMt$ can be enriched.

$$mt_0 : \forall p\ ist(mt_g, p) \wedge genlMtC(mt_g, mt_s) \longrightarrow genlMt(mt_g, mt_s) \wedge \circledcirc(mt_g, mt_s)$$

Hence, $genlMtC(mt_g, mt_s)$ holds if mt_s is a subtheory of mt_g and all footprints of individuals of geographic feature types specified in mt_s are (spatially or administratively) contained in mt_g; see [8] for more details. This containment predicate ($\circledcirc$) requires a spatial footprint for the individuals as well as for the spatial scope of the theory; a formal semantics including

[7] Which does not exclude crisp borders between them as in case of some administrative factors.

the Region Connection Calculus (RCC) is left for further work.

The usefulness of this approach can be demonstrated by the INSPIRE example. Adding $genlMtC$ to the meta-theory structuring local ontologies ensures that geographic feature types defined by states that are administratively contained by the European Union must be sub-types of the EU wide definition. Based on this requirement, instead of developing common schemata for all European member states top–down, local conceptualizations and non-standard inference such as computing the Least Common Subsumer (LCS) [7] and similarity reasoning [19] can be employed to automatically infer an appropriate top-level which does not violate local definitions. Consequently, if Spanish rivers do not necessarily contain flowing water but rivers in Germany do, the computed top-level for the EU should not define rivers based on the feature of flowing water; see [8] for details.

Summing up, besides subsumption hierarchies ontologies can be organized using space and time. Understanding and modeling the relation and interaction between ontologies will support the development of and reasoning about user-centric ontologies for the Web.

4. Towards an ecology of concepts

The previous sections argue that conceptualization is influenced by spatial and temporal factors, and that these factors can be used on a higher abstraction level to establish structure between different conceptualizations. Consequently, if concepts are not static, how to study their evolution [30] and diversification in space and time? Since shifts in conceptualization are difficult to detect and quantify, one may search for an analogy to a well known process. An interesting approach would be to study how species evolve and what factors drive their diversification. One promising candidate may be the process of *adaptive radiation*. In short, it described the evolutionary diversification of a single ancestor into several species each adapted to a particular ecological niche; Darwin's finches are a classical example. Simplifying, the process is caused by some (sudden) change in the environment, e.g., the volcanic creation of an isolated island. To construct a meaningful analogy requires to establish partial mappings between the evolution of concepts and biological evolution. While concepts and emerging new sub-concepts can be mapped to the radiation of species, the changing environmental, e.g., spatial, aspects can be mapped to a semantic space [34], distance to semantic distance, i.e., similarity, and so forth. The sudden diversification of *pothole* definitions caused by climatic and economic conditions may serve as a first example. An alternative approach based on time-geography was recently presented by Raubal [34] arguing for a time-indexed representation of concepts in GIScience. Schlieder discusses the related notion of *semantic ageing* for the long-term preservation of digital data [38].

For an impressive example on how theories from ecology and evolutionary biology can explain human strategies in gathering information, see Pirolli's Information Foraging Theory [31].

5. Conclusions

In this work we discussed the role of space and time for knowledge engineering and organization from three distinct perspectives: (1) their role for the definition of individual concepts, (2) their role for the organization of these concepts, and (3) their role for understanding their mutual interaction. We illustrated the need for semantic heterogeneity and the situated nature of conceptualization using various examples from the geo-sciences, outlined how space and time can act as structuring principles between local ontologies, and sketched an approach to model how concepts evolve in space and time. To create and maintain such local ontologies, we have to raise the user from a content creator to an active knowledge engineer. Citizens as sensors [11] and the Sensor Web may serve as a foundation for Semantic Web ontologies based on observation categoricals.

Acknowledgments

The comments from the reviewers and discussions with Werner Kuhn, Martin Raubal, and Simon Scheider provided useful suggestions to improve the content and clarity of this paper.

References

[1] D. Allemang and J. Hendler. *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2008.

[2] L. Barsalou. Situated simulation in the human conceptual system. *Language and Cognitive Processes*, **5**(6):513–562, 2003.

[3] L. Barsalou, S.A. Sloman, and S.E. Chaigneau. *Representing functional features for language and space: Insights from perception, categorization and development*, chapter The HIPE theory of function, pages 131–147. Oxford Univ. Press, 2005.

[4] B. Bennett, D. Mallenby, and A. Third. An ontology for grounding vague geographic terms. In *Formal Ontology in Information Systems – Proceedings of the Fifth International Conference (FOIS 2008)*, volume 183, pages 280–293. IOS Press, 2008.

[5] P. Bouquet, F. Giunchiglia, F. van Harmelen, L. Serafini, and H. Stuckenschmidt. C-owl: Contextualizing ontologies. In D. Fensel, K. Sycara, and J. Mylopoulos, editors, *The Semantic Web – ISWC 2003, Second International Semantic Web Conference*, pages 164–179, 2003.

[6] B. Brodaric and M. Gahegan. Experiments to Examine the Situated Nature of Geoscientific Concepts. *Spatial Cognition and Computation*, **7**(1):61–95.

[7] W. Cohen, A. Borgida, and H. Hirsh. Computing least common subsumers in description logics. In *10th National Conference on Artificial Intelligence*, pages 754–760. MIT Press, 1992.

[8] S. Duce and K. Janowicz. Microtheories for spatial data infrastructures - accounting for diversity of local conceptualizations at a global level. In *6th International Conference on Geographic Information Science (GIScience 2010)*, 2010; forthcoming.

[9] L. Gabora, E. Rosch, and D. Aerts. Toward an ecological theory of concepts. *Ecological Psychology*, **20**(1):84–116, 2008.

[10] J. Gibson. The theory of affordances. In R. Shaw and J. Bransford, editors, *Perceiving, Acting, and Knowing – Toward an Ecological Psychology*, pages 67–82. Lawrence Erlbaum Ass., Hillsdale, New Jersey, 1977.

[11] M. Goodchild. Citizens as Sensors: the World of Volunteered Geography. *GeoJournal*, **69**(4):211–221, 2007.

[12] B.C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler. Modular reuse of ontologies: theory and practice. *Journal of Artificial Intelligence Research*, **31**(1):273–318, 2008.

[13] H.P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and semantics*, volume 3. New York: Academic Press, 1975.

[14] R. Guha, R. Mccool, and R. Fikes. Contexts for the semantic web. In *International Semantic Web Conference (ISWC 2004)*, volume 3298 in *Lecture Notes in Computer Science*, pages 32–46. Springer, 2004.

[15] H. Halpin and P. Hayes. When owl:sameas isn't the same: An analysis of identity links on the semantic web. In *Linked Data on the Web (LDOW2010)*, 2010.

[16] J. Hawkins and S. Blakeslee. *On Intelligence*. Times Books, 2004.

[17] P.J. Hayes. The naive physics manifesto. In D. Michie, editor, *Expert Systems in the Micro Electronic Age*, pages 242–270. Edinburgh University Press, 1979.

[18] K. Janowicz, P. Maué, M. Wilkes, S. Schade, F. Scherer, M. Braun, S. Dupke, and W. Kuhn. Similarity as a quality indicator in ontology engineering. In C. Eschenbach and M. Grüninger, editors, *5th International Conference on Formal Ontology in Information Systems*, volume 183, pages 92–105. IOS Pres, October 2008.

[19] K. Janowicz and M. Wilkes. $SIM - DL_A$: A Novel Semantic Similarity Measure for Description Logics Reducing Inter-concept to Inter-instance Similarity. In L. Aroyo, P. Traverso, F. Ciravegna, P. Cimiano, T. Heath, E. Hyvoenen, R. Mizoguchi, E. Oren, M. Sabou, and E. Simperl, editors, *6th Annual European Semantic Web Conference (ESWC2009)*, volume 5554 of *LNCS*, pages 353–367. Springer, 2009.

[20] M. Johnson. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Univ. of Chicago Press, 1987.

[21] W. Kuhn. An image-schematic account of spatial categories. In S. Winter, M. Duckham, L. Kulik, and B. Kuipers, editors, *Spatial Information Theory, 8th International Conference (COSIT 2007)*, volume 4736 of *LNCS*, pages 152–168. Springer, 2007.

[22] G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

[23] R. Lefort. Telegraph.co.uk: Pothole britain: councils disagree about when hole becomes a pothole. http://www.telegraph.co.uk/motoring/news/7436237/Pothole-Britain-councils-disagree-about-when-hole-becomes-a-pothole.html, Published: 13.03.2010.

[24] S. Lehar. *The World in Your Head: A Gestalt View of the Mechanism of Conscious Experience*. Lawrence Erlbaum, 2003.

[25] G. Lund. Definitions of forest, deforestation, afforestation, and reforestation. Technical report, Forest Information Services, 03/22/2010.

[26] D.M. Mark. Toward a theoretical framework for geographic entity types. In *Spatial Information Theory: A Theoretical Basis for GIS, International Conference COSIT '93*, pages 270–283, 1993.

[27] J. McCarthy and S. Buvac. Formalizing context (expanded notes), 1996.

[28] D.R. Montello and P.C. Sutton. *An introduction to scientific research methods in geography.* Sage Publications, 2006.

[29] D.R. Montello, M.F. Goodchild, Jonathon Gottsegen, and Peter Fohl. Where's downtown?: Behavioral methods for determining referents of vague spatial queries. *Spatial Cognition & Computation*, **3**(2):185–204, 2003.

[30] N. Noy and M. Klein. Ontology evolution: Not the same as schema evolution. *Knowledge and Information Systems*, **6**(4):428–440, 2004.

[31] P. Pirolli. *Information Foraging Theory: Adaptive Interaction with Information*. Oxford University Press, Inc., New York, NY, USA, 2007.

[32] F. Probst and M. Lutz. Giving Meaning to GI Web Service Descriptions. In *International Workshop on Web Services: Modeling, Architecture and Infrastructure (WSMAI 2004)*, 2004.

[33] W.V.O. Quine. *Pursuit of Truth; revised edition*. Cambridge, MA: Harvard University Press, 1992.

[34] M. Raubal. Representing concepts in time. In C. Freksa, N. Newcombe, P. Gärdenfors, and S. Wölfl, editors, *International Conference on Spatial Cognition*, volume 5248 of *LNCS*, pages 328–343. Springer, 2008.

[35] M. Raubal and A. Adams. The semantic web needs more cognition. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):69–74, 2010.

[36] S. Scheider, K. Janowicz, and W. Kuhn. Grounding geographic categories in the meaningful environment. In K. Hornsby, C. Claramunt, M. Denis, and G. Ligozat, editors, *Conference on Spatial Information Theory (COSIT 2009)*, volume 5756 of *LNCS*, pages 69–87. Springer, 2009.

[37] S. Scheider, F. Probst, and K. Janowicz. Constructing bodies and their qualities from observations. In *Proceeding of the 2010 Conference on Formal Ontology in Information Systems (FOIS 2010)*, pages 131–144. IOS Press, 2010.

[38] C. Schlieder. Digital heritage: Semantic challenges of long-term preservation. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):143–147, 2010.

[39] L. Steels and T. Belpaeme. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, **28**:469–529, 2005.

[40] W. Tobler. A computer model simulating urban growth in the detroit region. *Economic Geography*, **46**(2):234–240, 1970.

[41] P. Williams, M. Whitfield, J. Biggs, S. Bray, G. Fox, P. Nicolet, and D.A. Sear. Comparative biodiversity of rivers, streams, ditches and ponds in an agricultural landscape in southern england. *Biological Conservation*, **115**(2):329–341, 2004.

Semantic Web 1 (2010) 69–74
DOI 10.3233/SW-2010-0002
IOS Press

The Semantic Web needs more cognition

Editors: Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited reviews: Philipp Cimiano, Universität Bielefeld, Germany; Aldo Gangemi, ISTC-CNR Rome, Italy

Martin Raubal[a,*] and Benjamin Adams[b]
[a]*Department of Geography, University of California, Santa Barbara, 5713 Ellison Hall, Santa Barbara, CA 93106-4060, USA*
[b]*Department of Computer Science, University of California, Santa Barbara, Engineering I, Room 2104, Santa Barbara, CA 93106-5110, USA*

Abstract. One of the key deficiencies of the Semantic Web is its lack of cognitive plausibility. We argue that by accounting for people's reasoning mechanisms and cognitive representations, the usefulness of information coming from the Semantic Web will be enhanced. More specifically, the utilization and integration of conceptual spaces is proposed as a knowledge representation that affords two important human cognitive mechanisms, i.e., semantic similarity and concept combination. Formal conceptual space algebra serves as the basis for the Conceptual Space Markup Language (CSML), which facilitates the engineering of ontologies using a geometric framework. We demonstrate the usefulness of the approach through a concrete example and suggest directions for future work, especially the need for combining geometric representations and reasoning mechanisms with existing Semantic Web structures.

Keywords: Conceptual space, Conceptual Space Markup Language, CSML, cognition, representation, markup language

1. Introduction

In 2004, Peter Gärdenfors argued that "the Semantic Web is not semantic" because it is good for syllogistic reasoning only and there is more to semantics than that [9]. In 2010, we claim here that the Semantic Web is still not semantic in the human sense because it does not sufficiently account for people's cognition, i.e., human conceptual representations and reasoning mechanisms. This must not to be confused with a search for *Strong Artificial Intelligence*, i.e., a Semantic Web whose intellectual ability cannot be distinguished from that of a human being [22]. But eventually, what comes out of the Semantic Web should be useful for people and it is our conviction that the better we integrate and account for people's reasoning mechanisms and cognitive representations the more useful such information will be.

Consider the example of looking for a warm climate vacation (Fig. 1). This search involves several questions that cannot be handled by the current Semantic Web, such as what is the meaning of 'warm' in a particular person's context of climate and how important is this dimension compared to other dimensions, such as distance and cost? This example makes it clear that the Semantic Web has to be based on a solid foundation of human concept processing, including limited knowledge and uncertainty in order to become truly semantic. In addition, representation and processing of context information, is key. Semantic models of context and contextualizing ontologies must account for human sensors and move into the direction of dynamic processes [5,17].

More specifically, we argue that knowledge representations underpinning the Semantic Web should afford two important human cognitive tasks: the *efficient calculation of semantic similarity* (in the context of the vacation example: how similar is the result to my ideal warm climate vacation?) and *combinations of concepts* ('warm' and 'climate'). However,

* Corresponding author. E-mail: raubal@geog.ucsb.edu.

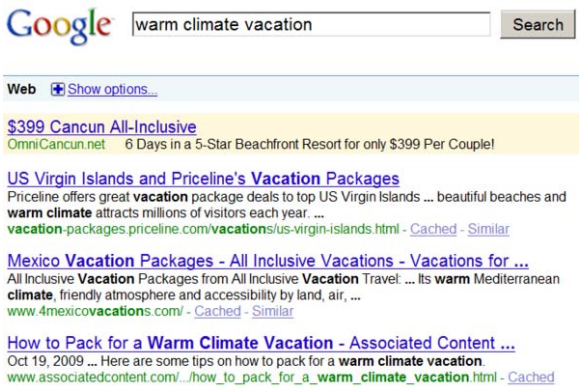

Fig. 1. Search for warm climate vacation.

the existing logical foundations of the Semantic Web—description logics and rules—presume a set-based classification scheme that does a poor job of facilitating these operations. By adopting a geometric and topological representational framework called *conceptual spaces* to describe semantics at the conceptual level, these operations can be defined in terms of an efficient vector algebra. This opens up the possibility to go beyond the classical concept combination possibilities of conjunction, disjunction, and negation. Conceptual spaces were conceived as a theory for how concepts are learned based on the paradigm of cognitive semantics [14], which emphasizes the role of similarity and prototype effects in categorization [20], and the importance of metaphorical and metonymic reasoning. Combined with natural language processing and existing methods of sentential representation, geometric conceptual representation has the potential to create a much richer and cognitively plausible Semantic Web [6].

2. Conceptual space algebra

Conceptual spaces were introduced to represent information at the conceptual level [8]. They can be utilized for knowledge representation and sharing, and account for the fact that concepts are dynamic and change over time [3,19]. A conceptual space is a set of quality dimensions with a geometric / topological structure for one or more domains. Domains are represented by sets of *integral* dimensions, which are separable from all other dimensions. Concepts cover multiple domains and are modeled as n-dimensional regions. Every instance of a category can be represented as a vector in the conceptual space [18]. This allows for expressing the similarity between two instances as a function of the spatial distance between their vectors. The utilization of conceptual space theory within the Semantic Web requires a solid mathematical foundation. Adams and Raubal [1] presented a metric conceptual space algebra, which consists of formal definitions of its components and operations that can be applied to them. Conceptual spaces are defined as multi-leveled structures and a distinction is made between the representation of the *geometric* elements (regions, points) and the *conceptual* elements (concepts, properties, instances). Furthermore, contrast classes—special types of properties, which have meanings that are dependent on the concepts they modify—are specified. Context is defined as a set of salience weights that can be applied to components of any type in the conceptual space, and is therefore a first-order element of a conceptual space. Different algebraic operations, such as metric operations on points and regions, and context-dependent similarity and concept combination query operations can then be applied to the elements of a conceptual space.

In order to facilitate the engineering of ontologies [12] using a geometric framework, languages must be developed to describe the geometric structures. The Conceptual Space Markup Language (CSML) is based on the described algebra and designed for this purpose.

3. Conceptual space markup language

CSML [2] is an XML-based language that allows one to create an ontology of concepts, properties, instances, contrast classes, and contexts as defined in the algebra above. The following shows the *climate* domain described in CSML with two dimensions temperature and precipitation.

```
<csml:Domain csml:ID="Climate">
  <csml:QualityDimension
csml:ID="Temperature">
    <csml:Scale> interval </csml:Scale>
  </csml:QualityDimension>
  <csml:QualityDimension
csml:ID="Precipitation">
    <csml:Scale> ratio </csml:Scale>
  </csml:QualityDimension>
</csml:Domain>
```

Different climate properties (e.g., wet, dry, hot, cold, Californian, temperate) are represented as regions within the climate domain. In CSML properties and contrast classes are described as systems of linear

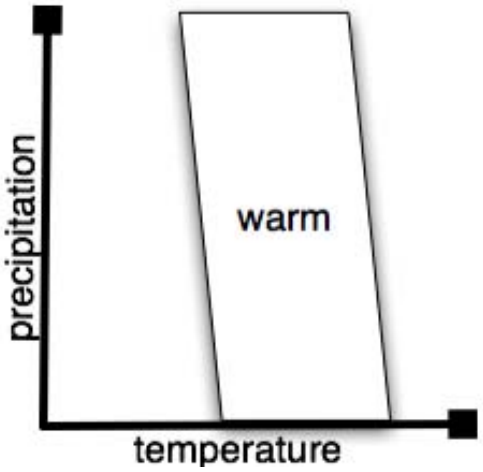

Fig. 2. Warm contrast class.

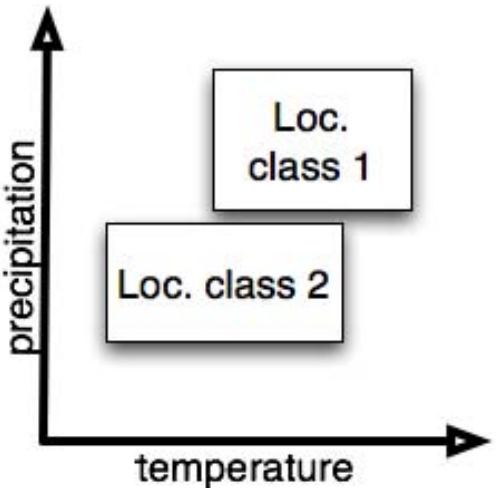

Fig. 3. Climate properties.

inequalities expressed using a variant of MathML. The following shows how one can represent *warm* as a contrast class in CSML.

```
<csml:ContrastClass csml:ID="Warm"
csml:DomainID="Climate">
     <csml:aVector>
       <cn> 5.4 </cn>
       <cn> 1.0 </cn>
     </csml:aVector>
     <csml:qVector>
       <ci> Temperature </ci>
       <ci> Precipitation </ci>
     </csml:qVector>
     <csml:ccMin> -3.0 </csml:ccMin>
     <csml:ccMax> 4.6 </csml:ccMax>
</csml:ContrastClass>
```

As well, different classes of locations have different climate properties (e.g., California climate) represented as regions in the climate domain bounded by, for example, minimum and maximum average temperatures and precipitation measures.

For the scenario where the user wants to search for a warm climate vacation, it is straightforward to represent the requisite elements in CSML in a manner that affords semantic search based on context. Specifically, one can frame the goal of this semantic search query by identifying the concepts that are most semantically similar to the user's idealized or prototypical warm vacation location depending on the user's location. Here there are really two different kinds of context at play. First, there is context represented as salience weights on the dimensions for the purpose of similarity measurement. In the example, precipitation might be weighted as highly as temperature because for two locations to have the same climate both precipitation and temperature matter

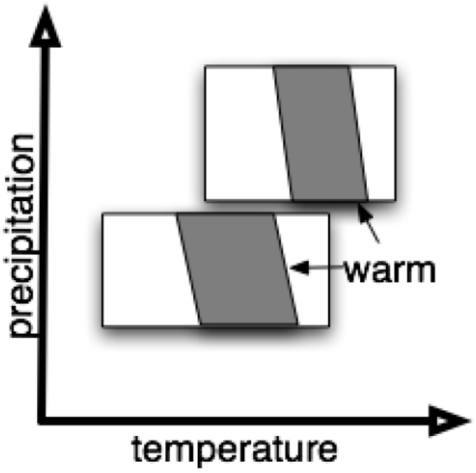

Fig. 4. Combination of warm with climate properties (dark areas).

equally. This first kind of context is described using the `csml:Context` tag in CSML. Second, there is context in terms of which climate property *warm* should modify.

The concept of a *warm German vacation* entails different semantics for the term *warm* than does *warm California vacation* (or for that matter *warm coffee*, which is actually cold!). As a contrast class, *warm* is represented in a conceptual space as a subregion of the entire climate domain (Fig. 2). The combination of *warm* with another class is not the intersection but rather a geometric affine projection of the warm region onto the other class' climate property (Figs 3 and 4). And unlike union and intersection this class constructor is asymmetric.

With this kind of concept combination operation one can very easily reason non-monotonically that what is warm in Sweden is not warm in Europe even if European country is modeled as a generalization or super-class of Sweden. Further, the geometric representation allows one to represent classes in terms of prototypical instances, i.e., as vectors or regions in a conceptual space [21]. This prototype representation is far more natural for representing classes without clear necessary and sufficient features (i.e., classes with degrees of membership determined by similarity to a prototype ideal, such as classes of shapes).

We should note that a semantic search query such as the one above does require a system that can identify which terms are modifying other terms. However, this problem is true for description logics as well and illustrates the need for a natural language processing [16] layer for the Semantic Web.

4. Where to go from here

Since it requires identifying the measurable dimensions of a property, the geometric representation might on the face of it seem overly restrictive. However, there is ample evidence that spatial metaphors are used in conceptualizations for many domains of knowledge [15], including any ordinal, interval, and ratio scaled measurements of observable phenomena. In addition, this representation does not necessarily require that the dimensions be identified in the cases when they are modeled as latent variables using techniques such as multidimensional scaling. Further, from the ontology engineer's perspective it makes little sense in many cases to translate the semantics of metric, spatially ordered data into a description logic representation, because 1) it adds unnecessary complexity, since transitivity, disjointness, and other logical characteristics emerge directly from the order topology of the space and 2) it affords the use of linear algebra and computational geometry algorithms as the foundations for many reasoning operations, which can be much more efficient. From a cognitive perspective the latter point aligns with the argument that much similarity measurement happens at the perceptual level without the need for higher-order cognitive representation [10]. Nevertheless, an important future development will be the formalization of mappings between conceptual space representations and OWL based ontologies. This includes the representation of vague information in Fuzzy OWL [16] and comparing the semantic expressiveness between conceptual spaces and Fuzzy OWL.

Description logics and conceptual spaces are two different knowledge representation frameworks with different degrees of semantic expressiveness and thus mappings between the two can result in a loss of information. Generally speaking, numeric datatype properties and object properties can be mapped to dimensions and regions in domains, respectively, but there are exceptions to this rule. In most cases the taxonomic relationships in a conceptual space representation can be 'frozen' into a description logic-based representation, but in doing so it loses expressiveness. For example, categories that are defined using prototypes will entail different memberships depending on context (i.e., dimension weights, which may be set by the user or automatically be assigned through learning from user behavior), so a conceptual space can generate a number of different OWL ontologies dependent on context. In addition, the notion of membership in a category existing on a continuum based on similarity is lost. The representation of regions as sets of linear inequalities might be achievable with the proposed OWL 2 Linear Equations data range extension, but arguably in a very cumbersome manner[1]. Since it is likely that ontology engineers will want to retain their ability to use all the existing features of OWL, a hybrid (or dual) representation will be in order. Such a hybrid representation would give ontology engineers the flexibility to define classes based on necessary and sufficient features or prototypes and use set based class constructors as well as more cognitively plausible methods based on contrast classes.

In the semantic web layer cake we conceive of the CSML layer as being a layer that sits on top of XML and beside rules and OWL. CSML can be an earlier stage in the pipeline for building an OWL representation, though it also has a role within the reasoning pipeline, e.g., when doing similarity measurement. Furthermore, reasoning can be done on conceptual spaces without mapping to OWL and this reasoning can exploit the characteristics of the geometric representation as a foundation for more complex sorts of class constructors. The following steps illustrate an example of how the CSML layer can be used to map a set of classes represented by prototypes in a feature space to an OWL ontology (see also Fig. 5).

1. A machine learning algorithm is used to learn points of central tendency for classes of observations. These points are interpreted as proto-

[1] http://www.w3.org/TR/2009/WD-owl2-dr-linear-20090421/

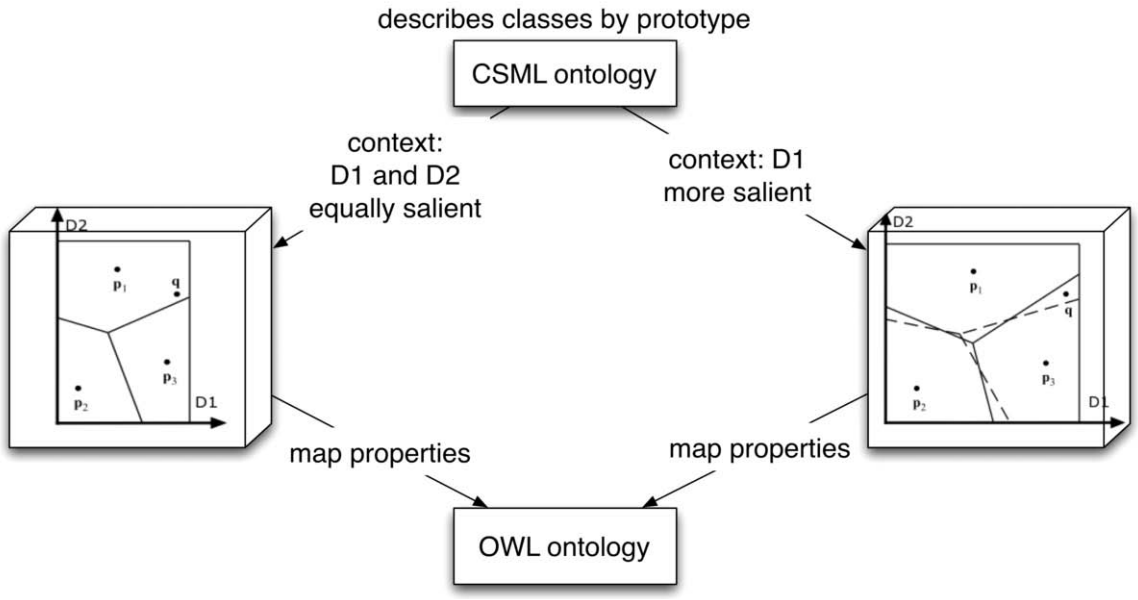

Fig. 5. Example of context changing classification.

typical instances of the classes. Note that most machine learning algorithms of this sort are flat, so in conceptual space parlance it is in fact properties that are being learned not classes (i.e., they are in one domain).

2. The dimensions and points are represented as quality dimensions and instances in CSML.
3. If the classes are disjoint then the Voronoi tessellation of the space based on these prototypes can be used to identify regions representing different properties. By placing different saliency weights on the dimensions the Voronoi tessellation may classify a particular observation differently [7].
4. A mapping is made from CSML to OWL for a given context.

A mapping from a CSML ontology to an OWL ontology is a morphism that reifies the quality dimensions, properties, instances, and concepts as OWL classes, properties, and individuals (see also [13]). In a metric conceptual space quality dimensions are mapped to transitive datatype properties, properties to object properties, instances to individuals, and concepts to classes. Alternately, domains can be mapped to object properties and conceptual space properties to individuals. A logical formalization of this mapping as a function that takes two input parameters, a conceptual space knowledge base and a set of context weights, and outputs a SROIQ(D) description logic knowledge base is a current research objective. Note in particular, that by mapping conceptual space properties to object properties information about the geometric structure is lost, therefore maintaining a link to the CSML representation that generated the OWL ontology can be used for finer-grained similarity reasoning. In addition, conceptual spaces are well-suited for non-monotonic changes based on new observations, e.g., adding new points to the original space can change the points of central tendency and the resulting property regions, which can be mapped to an updated OWL ontology.

This leads directly to the question of how we will be able to generate cognitively plausible ontologies in the future not only from measurement data but from the mass of user-generated data such as Volunteered Geographic Information (VGI) [11]. It will also be necessary to contextualize these ontologies on the fly (see the warm vacation example used here). With conceptual space representations this may be done by putting weights on the dimensions and modifying the classifications. If we know about people's prototypical concepts for different domains, how can we construct ontologies from there?

The future will show whether what is out there can be integrated with conceptual space theory and whether such combination and integration of ideas will eventually pave the way to a truly cognitively plausible Semantic Web, a Semantic Web that is useful for its users.

Acknowledgments

This work is supported by a UCSB faculty research grant. The two reviewers' suggestions helped improve the content of the paper.

References

[1] Adams, B. and M. Raubal, A Metric Conceptual Space Algebra, in *Spatial Information Theory – 9th International Conference, COSIT 2009, Aber Wrac'h, France, September 2009*, K. Stewart Hornsby, et al., Editors. 2009, Springer: Berlin. p. 51–68.

[2] Adams, B. and M. Raubal, The Conceptual Space Markup Language (CSML): Towards the Cognitive Semantic Web, in *Third IEEE International Conference on Semantic Computing (ICSC 2009), 14–16 September, Berkeley, California*. 2009, IEEE Computer Society. p. 253–260.

[3] Barsalou, L., W. Yeh, B. Luka, K. Olseth, K. Mix, and L. Wu, Concepts and meaning, in *Papers from the parasession on conceptual representations*, K. Beals, et al., Editors. 1993, University of Chicago: Chicago Linguistics Society. p. 23–61.

[4] Bobillo, F. and U. Straccia, An OWL Ontology for Fuzzy OWL 2, in *Foundations of Intelligent Systems, 18th International Symposium, ISMIS 2009, Prague, Czech Republic, September 14–17, 2009, Proceedings*, J. Rauch, et al., Editors. 2009, Springer: Berlin. p. 151–160.

[5] Corcho, O. and R. García-Castro, Five challenges for the Semantic Sensor Web. *Semantic Web – Interoperability, Usability, Applicability*, 2010. **1**(1,2): p. 121–125.

[6] Gangemi, A. and V. Presutti, Towards a pattern science for the Semantic Web. *Semantic Web – Interoperability, Usability, Applicability*, 2010. **1**(1,2): p. 61–68.

[7] Gärdenfors, P. and M. Williams. Reasoning about Categories in Conceptual Spaces, in *Fourteenth International Joint Conference of Artificial Intelligence*. 2001. p. 385–392.

[8] Gärdenfors, P., *Conceptual Spaces – The Geometry of Thought*. 2000, Cambridge, MA: Bradford Books, MIT Press.

[9] Gärdenfors, P., How to Make the Semantic Web More Semantic, in *Formal Ontology In Information Systems*, A. Varzi and L. Lieu, Editors. 2004, IOS Press: Amsterdam. p. 17–34.

[10] Goldstone, R. and L. Barsalou, Reuniting perception and conception. *Cognition*, 1998. **65**: p. 231–262.

[11] Goodchild, M., Citizens as sensors: the world of volunteered geography. *GeoJournal*, 2007. **69**: p. 211–221.

[12] Guizzardi, G., Theoretical foundations and engineering tools for building ontologies as reference conceptual models. *Semantic Web – Interoperability, Usability, Applicability*, 2010. **1**(1,2): p. 3–10.

[13] Janowicz, K., B. Adams, and M. Raubal, Semantic Referencing – Determining Context Weights for Similarity Measurement, in *Geographic Information Science – Sixth International Conference, GIScience 2010, Zürich, Switzerland, September 14–17, 2010, Proceedings*, S. Fabrikant, et al., Editors. 2010, Springer: Berlin.

[14] Janowicz, K., M. Raubal, A. Schwering, and W. Kuhn, Semantic Similarity Measurement and Geospatial Applications (Editorial). *Transactions in GIS*, 2008. **12**(6): p. 651–659.

[15] Johnson, M., *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. 1987, Chicago: The University of Chicago Press.

[16] Jurafsky, D. and J. Martin, *Speech and Language Processing – An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd ed. 2009, Upper Saddle River, New Jersey: Pearson – Prentice Hall.

[17] Keßler, C., M. Raubal, and C. Wosniok, Semantic Rules for Context-Aware Geographical Information Retrieval, in *Smart Sensing and Context – 4th European Conference, EuroSSC 2009, Guildford, UK, September 16–18, 2009*, P. Barnaghi, et al., Editors. 2009, Springer: Berlin, Heidelberg. p. 77–92.

[18] Raubal, M., Formalizing Conceptual Spaces, in *Formal Ontology in Information Systems, Proceedings of the Third International Conference (FOIS 2004)*, A. Varzi and L. Vieu, Editors. 2004, IOS Press: Amsterdam, NL. p. 153–164.

[19] Raubal, M., Representing concepts in time, in *Spatial Cognition VI – Learning, Reasoning, and Talking about Space. Proceedings of the International Conference Spatial Cognition 2008, Freiburg, Germany*, C. Freksa, et al., Editors. 2008, Springer: Berlin. p. 328–343.

[20] Rosch, E., Principles of Categorization, in *Cognition and Categorization*, E. Rosch and B. Lloyd, Editors. 1978, Lawrence Erlbaum Associates: Hillsdale, New Jersey. p. 27–48.

[21] Schwering, A. and M. Raubal, Measuring Semantic Similarity between Geospatial Conceptual Regions, in *GeoSpatial Semantics – First International Conference, GeoS 2005, Mexico City, Mexico, November 2005*, A. Rodriguez, et al., Editors. 2005, Springer: Berlin. p. 90–106.

[22] Searle, J., Minds, brains, and programs. *Behavioral and Brain Sciences*, 1980. **3**(3): p. 417–457.

Semantic Web 1 (2010) 75–81
DOI 10.3233/SW-2010-0003
IOS Press

A taskonomy for the Semantic Web

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Manfred Hauswirth, DERI, National University of Ireland, Galway, Ireland; Mark Gahegan, University of Auckland, New Zealand
Open review(s): Frank van Harmelen, Donovan Artz, Pascal Hitzler, Osma Suominen

Tom Heath
Talis Systems Ltd., Knights Court, Solihull Parkway, Birmingham Business Park, B37 7YB, United Kingdom
E-mail: tom.heath@talis.com

Abstract. The modalities of search and browse dominate current thinking about interaction with the Web. Given the Web's origins as a global hypertext system, it is understandable that these document-centric interaction patterns prevail. However, these modalities alone are inadequate as a conceptual model of interaction with the global Linked Data space that is the Semantic Web. Realising the full potential of the Semantic Web requires a fundamental reconsideration of Web interaction patterns in the light of Linked Data, and this renewed conceptualisation must drive the research agenda related to user interaction and the Semantic Web. This paper argues that a fundamental understanding of user goals and tasks is the appropriate perspective from which to approach the research and development of Semantic Web applications. However, the *Web* in Semantic Web should not detract from the potential for cross-platform data interoperability enabled by the Semantic Web technology stack. In this context we propose a *taskonomy* of data- and object-centric user tasks derived from an analytical abstraction of existing research, not simply in the fields of Web search and browse but also email and Instant Messaging, that can help shape the direction of research and application development in the Semantic Web field.

Keywords: Semantic Web, Linked Data, email, Instant Messaging, user interaction, taskonomy, activity-centred design

1. The ubiquitous document metaphor

"The correct approach to the support of behavior is activity-based classification." [15]

Look closely enough and it becomes apparent that document-centric metaphors are fundamental to the concept, design and realisation of our most widely used computing systems. Computers have desktops, files live in folders, we add pages to our Web sites. The terminology of email (*mailbox*, *postmaster*, *blind carbon copy*, *attachments*) reflects a communication platform conceived in the era of memos and postal systems. The Web emerged from a desire to share information between scientists [2], and owes much to the influence of the document-centric fields of hypertext and information retrieval:

"Computers give us two practical techniques for the man-knowledge interface. One is hypertext, in which links between pieces of text (or other media) mimic human association of ideas. The other is text retrieval, which allows associations to be deduced from the content of text. In the first case, the reader's operation is typically to click with a mouse (or type a reference number) – in the second case, it is to supply some words representing that which he desires. The W3 ideal world allows both operations, and provides access for any browsing platform." [1]

1.1. Classifying Web search

This original model of the Web has defined our view of how it is used ever since, with searching and browsing remaining the prevalent lens through which we view human interaction with the Web [16], even occurring in more data-centric analyses such as [19].

The dominance of the document-metaphor manifests itself not only in the computing applications we develop, but in the research conducted to try and un-

derstand how people use the Web in practice. While work such as [6] has attempted to understand the range of activities conducted on the Web, e.g. banking, job-hunting, or finding travel information, numerous others have attempted to identify and classify various forms of Web search:

- Guha, McCool and Miller [7] distinguish between *navigational* searches, where "the user is using the search engine as a navigation tool to navigate to a particular intended document" and *research* searches, where the user is "trying to locate a number of documents which together will give him/her the information s/he is trying to find" (pp. 702).
- Broder [4] identifies three types of Web search: *navigational* and *informational* searches, that map closely onto the *navigational* and *research* searches of Guha et al. [7], and *transactional* searches where the user intends "to reach a site where further interaction will happen" (pp. 6), such as a shopping site or a site where images or music can be downloaded.
- Related work by Rose and Levinson [17] yielded top-level categories with many similarities to those of Broder [4], in addition to a number of more specific sub-categories (e.g. *download*, *entertainment*, *interact*, and *obtain*).
- Morrison, Pirolli, and Card [13] describe a taxonomy of Web activities with three variables: the *purpose* of a search, the *method* used, and the *content* of the information being searched for.

1.2. Distortions in the search-centric lens

As comprehensive models of Web search, these classifications have a number of limitations. For example, in the work of Broder [4], the range of possible *transactions* a user may wish to perform, and the underlying reasons for wishing to perform them, are not explored. Similarly, consideration is not given to why a user may wish to *navigate* to a particular Web site or document. Presumably this destination does not represent an end in itself, but part of the strategy for performing another task, such as finding a phone number or arranging car rental.

In addition, while Rose and Levinson [17] give a number of examples to illustrate their sub-categories, the distinctions between them are often based on technical aspects of how the target object will be used, rather than the fundamental nature of the task the user is performing. For example, the target of the *download* goal is "a resource that must be on my computer or other device to be useful" (pp. 15), and the authors cite the example of a piece of software. However, the same definition could equally apply to the adult movie example used to illustrate the *entertainment* sub-category. In both cases the key feature is the attempt to *locate* something specific; drawing arbitrary category distinctions between these serves only to obscure the commonality in the underlying goal of the user.

At first glance the variables proposed by Morrison, Pirolli and Card [13] appear neatly defined. However, the classification of some activities suggests the variables may not be mutually exclusive in the form presented by the authors. For example, some methods are seen to be triggered by a particular goal (*find*, *collect*) whereas others (*explore*, *monitor*) are not. On the contrary, there is a strong argument that *explore* and *monitor* represent goals in their own right, and should be classed under *purpose*.

The focus of these studies on classifying search behaviors may be valuable in informing the ongoing development of Web search engines. However, by taking a search-centric perspective on Web usage these classifications may often obscure the true goal of the user in being online and perpetuate the ubiquity of the document metaphor in attempts to understand how the Web can support people in achieving their goals. The search-specific focus of these studies means none can account for more complex tasks performed on the Web. While the *resource-interact* goal of Rose and Levinson [17] and the *transactional* queries of Broder [4] suggest an intention to carry out further interaction beyond the search (perhaps indicating a greater overall goal), the search itself is still seen as the user's primary task. No mention is given of, for example, *arranging a holiday* as an overarching reason for being online, or even for carrying out a search. While analysis of search query logs is unlikely to show many queries such as "arrange holiday", this likely reflects an awareness among users of the narrow scope of search engines rather than a lack of desire the use the Web for this purpose.

2. That which is not search

A number of studies have investigated a broader range of tasks beyond simply Web search:

- Sellen, Murphy, and Shaw [18] describe six types of activity carried out on the Web, based on a study of Web use by twenty-four knowledge workers: *finding*, *information gathering*, *browsing*, *transacting*, *communicating*, *housekeeping*).
- Kellar, Watters, and Shepherd [11] report a study into how people use "web browser navigation mechanisms", in which participants were asked to classify their Web usage according to the following tasks: *fact finding*, *information gathering*, *just browsing*, *transactions*, and *other*. This classification was informed by previous studies, such as [18], but refined based on the findings of pilot studies with users.
- Kellar [10] refines the top level classification of Kellar et al. [11], grouping *fact finding*, *information gathering*, and *browsing* under an *information seeking goal*; *transactions* are joined by *communications* under *information exchange*; lastly a new top-level goal of *information maintenance* is added, containing a single *maintenance* task.

The classifications of [10,11,18] are not limited to describing variations of Web search and do attempt to capture the user's needs or goals in using the Web, with some success. For example, the concept of *transacting* is a first-class citizen of all three classifications, without the degree of indirection present in the notion of a *transactional search*. In contrast, it is not clear whether the notion of *browsing* represents a goal in its own right, or simply an activity in support of some further (unspecified) goal. As already noted, this confusion of purpose and method is a consistent theme in attempts to understand how people use the Web and how the Web supports individuals in meeting their goals.

2.1. Fragmented platforms for communication

Perhaps the most significant limitation in all the work reviewed to this point is the focus purely on Web-based tasks. Sellen, Murphy and Shaw [18] define their *communicating* task as "Using the Web in order to participate in chatrooms or discussion groups" (pp. 229), but exclude email activities from the data. Similarly, Kellar [10] introduces a *communications* task, but uses email as an illustrative example.

As intuition would suggest, and these findings corroborate, the Web is regularly used for two-way communication of the sort conducted through email, chatrooms and discussion groups. *Communicating* accounted for just four percent of observed activities by Sellen et al.[18], however it is likely that the inclusion of email in the analysis, in addition to increased use of Webmail services in recent years and the advent of Web-based *microblogging* services, would result in a significantly higher percentage if the study was repeated.

An examination of analogous research into how people use email reveals, unsurprisingly, significant usage of this *platform* to conduct asynchronous communication. For example, in a study of 20 office workers, of varying roles, Whittaker and Sidner [20] focus on three main email functions: *task management*, *personal archiving*, and *asynchronous communication*. Of particular note in this case is that Whittaker and Sidner found evidence of email being used for a significantly wider range of tasks than purely asynchronous communication, for which is was originally conceived. They refer to this process as *email overload*.

Additional evidence for the overloading of email as a platform comes from an investigation of members of a large research laboratory [12], in which email was identified as supporting the following work functions: *information management*, *time management*, and *task management*. It was noted that those participants for whom email served an information management function may have job roles that involve staying abreast of developments by tracking information, a task that bears a noteworthy similarity to the *monitoring* activities identified in the work of Kellar [10] and Morrison, Pirolli and Card [13].

Ducheneaut and Belotti [5] identify additional ways in which people use email to perform tasks. For example:

- All but one participant reported regularly using email to exchange files with others.
- Eighty percent of respondents reported using email to arrange meetings.
- Seventy-two percent of participants used email to make decisions.

In the case of the latter point, how this decision-making is achieved in practice is not discussed by Ducheneaut and Belotti. Presumably email is used as a medium for discussion from which a decision can be reached.

While mainstream use of Instant Messaging (IM) is a more recent phenomenon than adoption of email and the Web, research in this area reveals noteworthy overlap with the tasks identified in the literature discussed above related to email and the Web. While use of IM for *communicating* is to be expected, the litera-

ture demonstrates that this platform also plays a role in tasks such as *finding* and *arranging*.

Following a study of how IM is used in the workplace, Nardi, Whittaker and Bradner [14] describe a number of informal communication tasks supported by this technology:

- *quick questions and clarifications*
- *coordination and scheduling*
- *organising impromptu social meetings*
- *keeping in touch with friends and family*

Whilst the latter three tasks seem rather distinct, further examination shows that they share a common theme of the participants making arrangements. Such arranging may vary from purely work-related, such as scheduling a meeting, to workplace social arrangements such as meeting colleagues for lunch, to coordinating social activities with friends and family outside of work.

Isaacs, Walendowski, Whittaker, Schiano and Kamm [9] refine the work of Nardi et al. [14] by showing how the prevalence of particular tasks in IM usage varies across different types of users. Frequent IM partners, or those seen as heavy users, used the medium predominantly for working together. Lighter users or infrequent partners generally used IM to carry out scheduling tasks.

At a general level, the findings of Isaacs et al. [9] are supportive of Nardi et al. [14]. Evidence was found for a number of similar functions:

- *Simple questions and information* bears a strong likeness to Nardi et al.'s *quick questions and clarifications*, whilst placing more emphasis on the simplicity of the exchange rather than the duration over which it occurred.
- Directly equivalent *scheduling* and *coordination* tasks are present in both classifications.
- The *social talk* task of Isaacs et al. is comparable to aspects of the *keeping in touch with friends and family task.*

The evolution of email into a platform for a wider range of tasks than those for which it was originally intended is consistent with the development of the Web into a general purpose platform for a variety of tasks that go beyond its original role as an information distribution platform. While evident in literature and practice, it should be noted that the *function creep* affecting email and the Web does not imply that these platforms are well adapted to the tasks for which they are being used. On the contrary, they may represent the best of several poor options.

With the literature also indicating that Instant Messaging supports a broad range of heterogeneous tasks, any comprehensive attempt to understand the tasks and goals of Web users must not examine this platform in isolation, but instead take a holistic view of how all Internet platforms (e.g. Web, email, IM) are used. If the goal of understanding user tasks and goals is to drive improvements in available applications and services, this understanding must be shaped not by existing applications and services that embody potentially counterproductive assumptions and metaphors, but by the notion of *activity-centred design* [15] and a fundamental examination of the underlying goals of Internet users. To do otherwise would be to confuse purpose with the method employed.

3. A Web of actions

To summarise the arguments so far, it is apparent from the literature that Internet platforms such as email, IM and the Web are widely used to support tasks for which they were not originally intended. While this is acceptable (of course), and perhaps an inevitable indicator and consequence of their success, it is not necessarily optimal from a user perspective, as applications developed for one purpose may not be well adapted to others. Is your email client optimised for task and project management? In addition, many of the applications we use to interact with email and the Web, and their underlying conceptual models, reflect a document-centric perspective that is not adequate for a world of Linked Data [3].

Linked Data, and the Semantic Web that has arisen from the large-scale publication of Linked Data, is about *things* and the *connections* between things. Linked Data is about the ability to publish descriptions of any aspect of any thing. It's about giving *identifiers* to those things, and maybe even interacting directly with those things, rather than just with documents that describe them.

A Web that is document-centric only enables users to interact directly with documents; it does not allow users to interact with or perform actions on the things described by those documents. The Linked Data paradigm changes that, by encouraging data publishers to assign HTTP URIs to any object or concept they wish to refer to. The existence of these identifiers paves the way for applications that support direct interaction

with things identified by URIs, or at least interaction that is less incumbered by the layer of indirection inherent in document-centricity.

If Linked Data is about *things* rather than just *documents*, what happens to the old metaphors that underpin so many of our computing systems? Are desktops and filing systems appropriate metaphors for organising and accessing things that are not documents, or is a broader perspective required? Does consideration not of what can be done with documents, but what can be done with *things in general* provide a different perspective – a set of thing-centric and action-centric metaphors to shape and inspire the Linked Data applications we build? What Linked Data brings to the table are the means to identify the things we want to act upon, and also to describe the kinds of actions that it is possible to perform on things of certain types.

In a world where people are identified by URIs, should a person *A* who wants to share a photo with person *B* have to choose between multiple platforms (e.g. email, IM, photo-sharing Web site, social networking Web site) in order to share the photo? Does it make conceptual sense to create a new email message and then attach the photo to that *document*, or simply to post the photo to the URI of the recipient and allow her to decide how it is handled on arrival? In the latter case, the recipient benefits because she gets to *choose* how and where the photo is received and stored, while the sender benefits precisely because he does not have to.

Similarly, is the fragmented nature of current communication channels optimal, whereby a person *C* wanting to notify person *D* of something must choose between multiple channels through which to achieve this goal, many of which may be suboptimal for the recipient at a particular time? Should it not be up to the recipient to choose the notification method, with the sender simply posting notifications to a canonical identifier for that recipient, perhaps accompanied by some indicator of the perceived urgency of the notification?

4. A taskonomy for the Semantic Web

With these questions in mind, the following list presents a *taskonomy* of user activities and goals online. This taskonomy was developed by distilling the tasks and activities identified in previous literature, and removing those that represented *means* or *methods* rather than *ends* or *purposes*, or only reflected artefacts of existing Internet platforms.

- **Locating**: Looking for an object or chunk of information which is known or expected to exist.
  Example: Locating an article from a journal, an image for a school project, a colleague's phone number, or information about a book a friend recommended.
- **Exploring**: Gathering information about a specific concept or entity to gain understanding or background knowledge of that concept or entity.
  Example: Exploring a philosophical theory to understand its central tenets; getting background information about an organization before a job interview.
- **Grazing**: Moving speculatively between sources with no specific goal in mind, but an expectation that items of interest may be encountered.
  Example: Following links that spark your interest on someone's blog.
- **Monitoring**: Regularly or repetitively checking known sources that are expected to change, with the express intention of detecting the occurrence and nature of changes.
  Example: Monitoring news Web sites during an election; checking email accounts for new messages; watching discussion fora for new ideas or information.
- **Sharing**: Making an object or chunk of information available to others.
  Example: Sharing holiday photos with a colleague; uploading a journal article to your personal Web site.
- **Notifying**: Informing others of an event in time or a change of state.
  Example: Emailing a group of friends to tell them you will be going to a concert at the weekend.
- **Asserting**: Making statements of fact or opinion available, with no discursive expectation.
  Example: Writing a review of a film, or stating on your Web site that you own a certain book.
- **Discussing**: Exchanging knowledge and opinions with others, on a specific topic.
  Example: Posting a comment on a discussion forum stating that you disagree with a previous post, explaining why, and then receiving responses from others.
- **Evaluating**: Determining whether a particular piece of information is true, or assessing a number of alternative options in order to choose between them.
  Example: Choosing which film to see at the week-

end, based on what's showing, where, and at what time.
- **Arranging**: Coordinating with third parties to ensure that something will take place or will be possible at a certain time.
  Example: Arranging travel and accommodation for an international conference.
- **Transacting**: Transferring money or credit between two parties.
  Example: Paying a bill.

With this taskonomy as a reference point, what forms of applications should we develop to exploit the unique capabilities of the Semantic Web?

Starting with the classic modalities of *search* and *browse*, we should not be developing Semantic Web search engines and browsers that crudely apply existing (document-centric) interaction styles to Linked Data [8]. Instead we should build services that allow us to *locate* specific information as efficiently as possible by incrementally supplying as much concrete information as we can and as is needed in order to narrow the search space sufficiently. This does not mean simply tweaking *Query by Example* interfaces to work over RDF data, but enabling query terms to be combined with background contextual information about the user in order to refine the result set.

Interfaces for *exploring* should not just be document browsers, but applications that integrate and summarise information about a specific thing of interest, based on its type, and adapt the interface accordingly.

Monitoring applications need to be able to adapt their interfaces based on the rate of change of different information sources, and the relative significance of these changes. Some information sources (e.g. Twitter) will change frequently with little consequence, while others (e.g. natural disaster warning systems) will change rarely but carry great significance. Effective monitoring applications that do not fragment user attention across multiple channels will need to account for each combination of information significance and rate of change, and adjust their behaviour accordingly.

Sharing, *notifying*, *asserting*, and *discussing* are currently supported by applications that frequently tie data to the application or system in or through which it was created. For example, notification emails stay in email systems, discussion forum posts stay unconnect to related posts made in different fora, files are uploaded to specific systems and then shared with others, rather than vice versa. A shift is needed such that applications emphasise the posting of notifications, discussion points, assertions, files, etc. into the Web at large, from where they can be retrieved as required by authorised parties, rather than simply into an application-specific silo.

Applications wishing to support *evaluating* and *arranging* may stand to gain the most in the near term from Linked Data and the Semantic Web. It is not hard to envisage how price comparison Web sites could be enhanced through Linked Data, such that products can be evaluated not simply based on price, but on local availability, delivery times, product reliability, guarantee terms, and environmental impact. This capability is feasible at present, but very costly due to the complexity of integrating data from numerous sources, each with proprietary interfaces.

Similarly, many domain specific *arranging* applications exist, such as flight comparison and booking Web sites. Where these applications fall short is in their rigidity; integration of arbitrary data relevant to a trip, but not specifically flight related, comes at a significant cost and not all types of information will warrant the investment despite potential value to a long-tail of users.

How *transacting* applications will truly benefit from the Semantic Web is not immediately clear. In one respect *transacting*, *sharing*, and *notifying* have much in common: in all cases the Semantic Web infrastructure allows a recipient to be uniquely identified, independently of any specific application or service. The result of this could be a democratisation of online payment services for end users, supported by common protocols and standards for payment interoperability.

Grazing, as defined above, is an activity with no specific, explicit user goal. In contrast it is likely to serve as a *displacement activity* that allows the user to defer performance of another (likely more important) task. Key factors in *grazing* would appear to be novelty, serendipity, and human interest. With much of the promise of the Semantic Web centred on increased precision, it is unclear what form a *grazing* application for Linked Data may take. Further research may increase our understanding of *grazing* and reveal whether this is a meaningful match for Semantic Web technologies.

5. Conclusions

As the adoption of Linked Data and Semantic Web principles and technologies continues, informal questions are increasingly being asked about the kinds of applications that could and should be developed to

make best use of these technologies. Meaningful answers to these questions can only be achieved through principled analysis that attempts to understand the areas in which Linked Data and the Semantic Web can make a unique contribution relative to conventional technologies.

The fundamental shift from *document-centricity* to *thing-centricity* brought about by the Linked Data paradigm creates opportunities for new forms of *activity-centred applications* and also challenges the research and development community to reassess the established metaphors that underpin computing applications and services. The ubiquity of documents in human culture suggests that alternative metaphors may not be easily identified, however the taskonomy presented in this paper can form the basis for discussion and innovation in the research community that can begin to address these challenges.

References

[1] T. Berners-Lee, R. Cailliau, J.-F. Groff, B. Pollermann, World-Wide Web: The Information Universe, *Electronic Networking: Research, Applications, and Policy*, **2**(1) (1992), 52–58.

[2] T. Berners-Lee, R. Cailliau, A. Luotonen, H.F. Nielsen, A. Secret, The World-Wide Web, *Communications of the ACM* **37**(8) (1994), 76–82.

[3] C. Bizer, T. Heath and T. Berners-Lee, Linked Data – The Story So Far, *International Journal on Semantic Web and Information Systems* **5**(3) (2009), 1–22.

[4] A. Broder, A Taxonomy of Web Search. *ACM SIGIR Forum* **36**(2) (2002), 3–10.

[5] N. Ducheneaut, V. Bellotti, E-mail as Habitat: An Exploration of Embedded Personal Information Management, *Interactions* **8**(5) (2001), 30–38.

[6] C. Emmanouilides, K. Hammond, Internet Usage: Predictors of Active Users and Frequency of Use, *Journal of Interactive Marketing* **14**(2) (2000), 17–32.

[7] R. Guha, R. McCool, E. Miller, Semantic Search, In *Proc. 12th International Conference on World Wide Web* (2003).

[8] T. Heath, How Will We Interact with the Web of Data?, *IEEE Internet Computing* **12**(5) (2008), 88–91.

[9] E. Isaacs, A. Walendowski, S. Whittaker, D.J. Schiano, C. Kamm, The Character, Functions, and Styles of Instant Messaging in the Workplace, In *Proc. ACM Conference on Computer Supported Cooperative Work* (2002).

[10] M. Kellar, An Examination of User Behaviour During Web Information Tasks, In *Proc. CHI'06 Extended Abstracts on Human Factors in Computing Systems* (2006).

[11] M. Kellar, C. Watters, M. Shepherd, The Impact of Task on the Usage of Web Browser Navigation Mechanisms, In *Proc. Graphics Interface* (2006).

[12] W.E. Mackay, Diversity in the Use of Electronic Mail: A Preliminary Inquiry, *ACM Transactions on Office Information Systems* **6**(4) (1988), 380–397.

[13] J.B. Morrison, P. Pirolli, S.K. Card, A Taxonomic Analysis of What World Wide Web Activities Significantly Impact People's Decisions and Actions, In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (2001).

[14] B. Nardi, S. Whittaker, E. Bradner, Interaction and Outeraction: Instant Messaging in Action, In *Proc. ACM Conference on Computer Supported Cooperative Work* (2000).

[15] D.A. Norman, Logic versus usage: the case for activity-centered design, *Interactions* **13**(6) (2006), 45,63.

[16] C. Olston, E.H. Chi, ScentTrails: Integrating Browsing and Searching on the Web, *ACM Transactions on Computer-Human Interaction* **10**(3) (2003), 177–197.

[17] D.E. Rose, D. Levinson, Understanding User Goals in Web Search, In *Proc. 13th International Conference on World Wide Web* (2004).

[18] A.J. Sellen, R. Murphy, K.L. Shaw, How Knowledge Workers Use the Web, In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (2002).

[19] F. van Harmelen, A. ten Teije, H. Wache, Knowledge Engineering rediscovered: Towards Reasoning Patterns for the Semantic Web, In *Proc. 5th International Conference on Knowledge Capture* (2009).

[20] S. Whittaker, C. Sidner, Email overload: exploring personal information management of email, In *Proc. SIGCHI Conference on Human Factors in Computing Systems* (1996).

Semantic Web 1 (2010) 111–115
DOI 10.3233/SW-2010-0004
IOS Press

The knowledge reengineering bottleneck

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Bernardo Cuenca-Grau, University of Oxford, UK; Sören Auer, Universität Leipzig, Germany

Rinke Hoekstra [a,b]
[a] *Department of Computer Science, VU University Amsterdam, The Netherlands*
E-mail: hoekstra@few.vu.nl
[b] *Leibniz Center for Law, Universiteit van Amsterdam, The Netherlands*
E-mail: hoekstra@uva.nl

Abstract. Knowledge engineering upholds a longstanding tradition that emphasises methodological issues associated with the acquisition and representation of knowledge in some (formal) language. This focus on methodology implies an ex ante approach: "think before you act". The rapid increase of linked data poses new challenges for knowledge engineering, and the Semantic Web project as a whole. Although the dream of unhindered "knowledge reuse" is a technical reality, it has come at the cost of control. Semantic web content can no longer be assumed to have been produced in a controlled task-independent environment. When reused, Semantic Web content needs to be remoulded, refiltered and recurated for a new task. Traditional ex ante methodologies do not provide any guidelines for this ex post knowledge reengineering; forcing developers to resort to ad hoc measures and manual labour: the knowledge reengineering bottleneck.

Keywords: Knowledge engineering, ontology reuse, design patterns, linked data, dirty data, data reuse, provenance

1. Introduction

The field of knowledge engineering upholds a longstanding tradition that emphasises methodological issues associated with the acquisition and representation of knowledge in some (formal) language. Examples are the development of task-independent ontologies and the recent interest in design patterns. However, the focus on methodology implies an ex ante approach: "think before you act". And in fact, the same attitude is prevalent in traditional web-based publication of information. Information is moulded, filtered and curated in a way that befits the purpose of the information provider. In this position paper, I argue that the field of knowledge engineering is facing a new challenge in the linked data age as information providers become increasingly dependent on external data and schemas.

1.1. Ex ante knowledge engineering

The ex ante approach of knowledge engineering originates in the problems identified in the development of large scale expert systems in the eighties and early nineties. Well known examples are Clancey's identification of types of knowledge in a knowledge base [4], the KADS and CommonKADS methodologies of [2,16] that separate a conceptual domain model from problem solving methods in the specifications of a knowledge based system, and Gruber's now famous characterisation of 'ontology' [10] and their physical reuse in the Ontolingua server [7] that culminated in the now commonplace use of the term to refer to a set of axioms that can be exchanged as a file. Ontologies soon became the center of attention for the field of knowledge acquisition – leaving problem solving methods largely ignored until only recently in e.g. [18]. It is the type of knowledge represented as an ontology – terminological knowledge – that was the main inspiration for the data model and semantics of the main Semantic Web languages.

The main focus was now directed towards the specification of design criteria and corresponding methodologies that ensured the development of ontologies suited for their main purpose: reuse in multiple sys-

tems [9]. For, it was thought, if ontologies are well-designed, they can be reused as task-independent knowledge components, enabling and facilitating more rapid construction of knowledge based systems by circumventing the knowledge acquisition bottleneck [8]. In the late nineties, and early 2000s, with the expected increase in the number of ontologies, a similar bootstrap seemed attainable by developing methods for reusing (parts of) ontologies in developing new ontologies, thus spawning research on ontology types, ontology merging, ontology alignment [15], ontology mapping, and – more recently – ontology modularisation.

In [12] I criticised the underlying assumptions of the alignment and merging of ontologies as these inevitably alter the ontological commitments of an ontology, rendering the claim of more reusable and compatible knowledge system components an empty one.[1] This criticism is moderated by the fact that many (if not most) ontologies are never used as a component of an expressive knowledge based system, but rather as facilitator for knowledge *management*; i.e. as 'semantic' annotations of information resources (documents, users). Knowledge management has indeed turned out to be the key use case for ontologies (and vocabularies) on the Semantic Web [6,12,18]. This is partly given by limitations of web-scale reasoning on expressive ontologies, although these limitations are of decreasing severity [17].

2. The bottleneck

The methodologies and technical solutions we briefly discussed in the preceding section have been motivated and developed in a world *without actual data*: ontology engineering is an activity that takes place at design time. In a knowledge management setting, ontologies are often used for the annotation of fresh data. But the world has changed; the linked data cloud is growing at an exponential pace, and more and more applications become dependent on it. This has a significant effect on the way in which knowledge is being reused on the web.

Feigenbaum's knowledge acquisition bottleneck refers to the difficulty of correctly extracting expert knowledge into a knowledge base:

> "The problem of knowledge acquisition is the critical bottleneck problem in artificial intelligence." [8, p. 93][2]

In contrast, the *knowledge reengineering bottleneck* refers to the general difficulty of the correct and continuous reuse of *preexisting* knowledge for a new task. The first difference between the two bottlenecks is that knowledge acquisition concerns the extraction of *generic* knowledge from a domain expert, while knowledge reengineering involves both generic and *assertional* knowledge. Indeed, knowledge engineering has contributed a lot to *enabling* schema level reuse, but traditional ex ante methodologies do not provide any guidelines for this ex post knowledge reengineering. Semantic web developers therefore resort to ad hoc measures and manual labour. The second difference is that on the linked data web, reuse is not a copy-and-paste operation, but rather a continuous relation of trust between a knowledge provider and its 'clients'.

Simply replace 'applied artificial intelligence' with 'the semantic web' in the following quote from Feigenbaum:

> "If applied artificial intelligence is to be important in the decades to come, we must have more automatic means for replacing what is currently a very tedious, time-consuming and expensive procedure." [8, p.93]

The tedious procedure alluded to by Feigenbaum is the procedure by which we integrate (existing) knowledge into a new system. The web of data may be more accessible than expert knowledge in a human brain, it is often expressed in a very convoluted manner, making it hard to reuse [11].

3. Challenges

The rapid increase of both quantity and importance of linked data poses new challenges for knowledge engineering and the Semantic Web project as a whole:

Challenge 1: Data Dependency Knowledge engineering is not yet fully accustomed to the ubiquity of instance data. An example is current work on ontology and vocabulary alignment. The Ontology Alignment Evaluation Initiative (OAEI) annually specifies

[1] In fact, this extends to the reusability of ontologies and ontology design patterns.

[2] The knowledge acquisition bottleneck is often misunderstood as the high threshold in effort before knowledge representation starts to pays off, and practical reasoning problems can be solved.

a set of ontologies for benchmarking alignment systems. These systems are evaluated against a reference alignment, or checked for coherence, but not against a set of instance data.[3] At the moment, this does not seem to be a very pressing issue. The most prominent use case for ontology alignment is information retrieval, and formal characteristics of the aligned ontologies and datasets play only a limited role. In a retrieval setting, alignment quality can be assessed by comparing precision and recall with or without using the alignments. A limited loss of retrieval quality can be outweighed by the added advantage of search using two vocabularies. In a more knowledge intensive setting, however, loss of quality has a more significant effect: instance data can be classified under the wrong type. How current ontology alignment techniques will scale to use cases for tasks that require higher expressiveness is at the present time still an open question.

Challenge 2: Limited Control Although the dream of unhindered knowledge reuse is a technical reality, it has come at the cost of control. Similar to the Web 2.0 revolution, where information consumers transformed into information producers; semantic web content can no longer be assumed to have been produced in a controlled environment. First of all, this means that data is 'dirty'; it may not be the latest version, it may be inconsistent, it may use multiple identifiers for the same resource, it may have gaps in coverage, or be redundant. The prototypical example of the dangers of this type of issues is the excessive use of owl:sameAs assertions between resources in different data sources. Furthermore, there is no guarantee that the ontologies that define the classes and properties used in the data are used in the specified way: the relata of a property may not be of the correct type, the data may be expressed in terms of an older version of a schema, or the data may cause the schema to become inconsistent.

Recently, the SIOC Project has made a change to its schema – an increasingly popular vocabulary for expressing social networking knowledge.[4] sioc:User was changed to sioc:UserAccount to avoid conflation of the class with foaf:Person. The change was announced on the SIOC website, and the schema owner advised users to change their data accordingly. Arguably, a change to such a widely used schema can have enormous consequences, certainly as we cannot assume that all occurrences of sioc:User will be replaced, nor that tool developers will provide the necessary update. But, what *are* these consequences, and how do we prevent or amend them?

The pragmatic, ad hoc approach to dirty data is to "just fix it". An example is the recently started "Pedantic Web" group; a group of concerned experts that functions as a communications channel between data owners, schema owners, and users, allowing them to file bug reports, and suggest fixes.[5] Indeed, repairing dirty data and schemas is a noble effort, but it is doubtful whether this initiative can scale and remain effective over the coming years.

In the end, data and schema quality have to be assured in some automatic way. Description logics reasoners will tell you whether a knowledge base is consistent, but there is a tradeoff in optimisation between expressive TBox reasoning, or reasoning on a large ABox (see e.g. [5]). Approaches that allow reasoning on very large amounts of (dirty) data, such as [17], are based on forward chaining algorithms that do not detect inconsistencies or other problems. An additional issue is that the results of tableaux algorithms are very hard to explain [13] and problems can only be fixed one at a time. Techniques for reasoning with inconsistent ontologies, such as e.g. [14], show promising results but their value depends on task context. Knowledge engineering can certainly play a role in investigating reasoning strategies for tasks on the web of data.

Another question is, what should a knowledge reuser do when encountering a problem? If it is not your own data, who should fix it? The model chosen by the BBC music website is to fix the original information source (e.g. Musicbrainz).[6] Clearly this model only works when dealing with community-developed *open* data; in a more restricted setting, other models will be more suited (including not fixing it). Different users may adopt conflicting models for the same data: a knowledge provider has to make clear how its data and/or schema should be used, what its versioning regime is, and has to provide provenance information for quality assurance.[7]

[3] See for evaluation methodology the OAEI and Ontology Matching workshop pages at http://oaei.ontologymatching.org/.

[4] SIOC: Semantically-Interlinked Online Communities. See http://sioc-project.org.

[5] "We want you to fix your data", see http://pedantic-web.org/

[6] See http://www.bbc.co.uk/music and http://www.musicbrainz.org, respectively.

[7] See other contributions in this volume, and the W3C Incubator Group on Provenance, http://www.w3.org/2005/Incubator/prov/.

Challenge 3: Increased Complexity The issues raised by the two preceding challenges are not new to many of us working with Linked Data. However, in context of the decennia-old debate between *neats* and *scruffies*,[8] these challenges are currently addressed only through the pragmatics of the latter perspective. Most of the experience gained there precipitates in blog posts, or best practices documents, rather than traditional scientific discourse.[9]

With scruffy linked data on the rise, it is likely that new Semantic Web applications [6] will capitalise on this data and move beyond the simple lookup and mashup services listed by [18]. These applications may not all live on the web or produce linked open data, but they will depend on it and require more expressiveness. As a consequence, the complexity and task-dependence of content on the web of data will increase, emphasising the need for a knowledge reengineering perspective. What does task-dependence of data mean on the web? Is there a role for knowledge engineering insights from the nineties, such as the problem solving methods of CommonKADS [3]? Understanding patterns in data reuse (as opposed to ontology design pattens) is currently uncharted territory.

Challenge 4: Increased Importance As the scale of the web of data increases, the number of applications that depend on it will increase as well. One of the major successes of the linked data initiative is the take-up by non-academic parties, such as the BBC, the UK and US governments, and more recently Google and Facebook. These parties are new stakeholders on the web of data, and it is not likely that this take-up is going to stop anytime soon. At the moment it is unclear how these non-academic parties will behave in the future, but linked data has already left the toy worlds of AI researchers and is increasingly mission critical to stakeholders. Facing the challenges iterated above becomes more important as coverage grows in influential domains such as commerce and legal and government information.

4. Discussion

In this short paper I call for a new role for knowledge engineering that takes the ubiquity of instance data into account. The challenges discussed in Section 3 are not new, but have to be faced in order to make the Semantic Web – and not just a web of data – a success. Indeed, that these challenges arise is a sign of a maturing domain. The dependency on data means that the web of data has become an object of study in its own right. It has grown beyond the control of the (academic) community that gave rise to it – similar to the Web itself [1].

Insights from knowledge engineering have played an important role in the initial design of Semantic Web technology, but the field seems to be sticking to its own turf rather than reaching out to help overcome the reengineering bottleneck.

[8] See [12, Ch.2] and http://en.wikipedia.org/w/index.php?title=Neats_vs._scruffies&oldid=323249466 for an overview.

[9] An example is Jeni Tennison's blog on her experiences with translating UK government data to RDF, http://www.jenitennison.com/blog/.

Acknowledgements

I would like to extend my thanks to the reviewers for their comments on the first version of this paper.

References

[1] T. Berners-Lee, W. Hall, J. Hendler, N. Shadbolt, and D.J. Weitzner. Creating a science of the web. *Science*, **313**, August 2006.

[2] J.A. Breuker and B.J. Wielinga. Knowledge acquisition as modelling of expertise: the KADS-methodology. In T. Addis, J. Boose, and B. Gaines, editors, *Proceedings of the European Knowledge Acquisition Workshop*, pages 102–110, Reading GB, 1987. Reading Press.

[3] J. Breuker and W. Van De Velde, editors. *CommonKADS Library for Expertise Modeling: reusable problem solving components*. IOS-Press/Ohmsha, Amsterdam/Tokyo, 1994.

[4] W.J. Clancey. The epistemology of a rule-based expert system – a framework for explanation. *Artificial Intelligence*, **20**(3):215–251, 1983. First published as Stanford Technical Report, November 1981.

[5] B. Cuenca Grau, B. Motik, Z. Wu, A. Fokoue, and C. Lutz. OWL 2 Web Ontology Language: Profiles. Technical report, W3C, 2009.

[6] M. d'Aquin, E. Motta, M. Sabou, S. Angeletou, L. Gridinoc, V. Lopez, and D. Guidi. Towards a new generation of semantic web applications. *IEEE Intelligent Systems*, **24**:20–28, 2008.

[7] A. Farquhar, R. Fikes, and J. Rice. The ontolingua server: a tool for collaborative ontology construction. *International Journal of Human-Computer Studies*, **46**(6):707–727, 1997.

[8] E.A. Feigenbaum. Knowledge engineering: the applied side of artificial intelligence. *Annals of the New York Academy of Sciences*, 426:91–107, 1984. Original publication in 1980 as report of the Stanford department of Computer Science.

[9] M. Fernández-López and A. Gómez-Pérez. Overview and analysis of methodologies for building ontologies. *The Knowledge Engineering Review*, **17**(2):129–156, 2002.

[10] T.R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Deventer, The Netherlands, 1994. Kluwer Academic Publishers.

[11] P. Hitzler and F. van Harmelen. A reasonable Semantic Web. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):39–44, 2010.

[12] R. Hoekstra. *Ontology Representation – Design Patterns and Ontologies that Make Sense*, volume 197 of *Frontiers of Artificial Intelligence and Applications*. IOS Press, Amsterdam, June 2009.

[13] M. Horridge, B. Parsia, and U. Sattler. Lemmas for justifications in OWL. In B. Cuenca Grau, I. Horrocks, B. Motik, and U. Sattler, editors, *Description Logics*, volume 477 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[14] Z. Huang, F. van Harmelen, and A. ten Teije. Reasoning with inconsistent ontologies. In L.P. Kaelbling and A. Saffiotti, editors, *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI'05)*, pages 454–459, Edinburgh, Scotland, Aug 2005.

[15] M. Klein. Combining and relating ontologies: An analysis of problems and solutions. In *Proceedings of the Workshop on Ontologies and Information Sharing (at IJCAI 2001)*, pages 53–62, Seattle, WA, 2001.

[16] G. Schreiber, H. Akkermans, A. Anjewierden, R. de Hoog, N. Shadbolt, W. Van den Velde, and B. Wielinga. *Knowledge Engineering and Managament: The CommonKADS Methodology*. MIT Press, 2000.

[17] J. Urbani, S. Kotoulas, J. Maassen, F. van Harmelen, and H. Bal. OWL reasoning with WebPIE: calculating the closure of 100 billion triples. In *Proceedings of the Seventh European Semantic Web Conference*, LNCS. Springer, 2010.

[18] F. van Harmelen, A. ten Teije, and H. Wache. Knowledge engineering rediscovered: Towards reasoning patterns for the semantic web. In N. Noy, editor, *The Fifth International Conference on Knowledge Capture*, pages 81–88. ACM, 2009.

Semantic Web 1 (2010) 121–125
DOI 10.3233/SW-2010-0005
IOS Press

Five challenges for the Semantic Sensor Web

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Christoph Schlieder, Bamberg University, Germany; Werner Kuhn, University of Münster, Germany
Open review(s): Roberto García, Universitat de Lleida, Spain

Oscar Corcho* and Raúl García-Castro
Ontology Engineering Group, Departamento de Inteligencia Artificial, Facultad de Informática, Universidad Politécnica de Madrid, Campus de Montegancedo, Boadilla del Monte, 28660, Madrid, Spain

Abstract. The combination of sensor networks with the Web, web services and database technologies, was named some years ago as the Sensor Web or the Sensor Internet. Most efforts in this area focused on the provision of platforms that could be used to build sensor-based applications more efficiently, considering some of the most important challenges in sensor-based data management and sensor network configuration. The introduction of semantics into these platforms provides the opportunity of going a step forward into the understanding, management and use of sensor-based data sources, and this is a topic being explored by ongoing initiatives. In this paper we go through some of the most relevant challenges of the current Sensor Web, and describe some ongoing work and open opportunities for the introduction of semantics in this context.

Keywords: Sensor, ontology, query language

1. Introduction

The combination of sensor networks with the Web, web services and database technologies, was named some years ago as the Sensor Web or the Sensor Internet [1,6,7,11,15]. Most of the work done on this topic, performed in some cases under the umbrella of the OGC Sensor Web Enablement Working Group[1], focused on the creation of specifications for different functionalities related to the management of sensor-based data (observations, measurements, sensor network descriptions, transducers, data streaming, etc.), and for the different types of services that may handle these data sources (planning, alert, observation and measurement collection and management, etc.).

Some additional work has focused on the provision of platforms that provide the services needed to develop sensor-based applications. These platforms include libraries for common domain-independent data management tasks, such as data cleaning, storage, aggregation, query processing, etc., and they are used to provide domain-specific aggregated services (e.g., coastal imaging [6], patient care [15]).

Finally, centralized registries for sensor-based data have appeared (e.g., Pachube[2], SensorMap[3]), focused on the registration of sensor-based data sources, and on the provision of access to them in multiple ways, by means of REST-based interfaces, web services, or ad-hoc query languages, to name a few.

Figure 1 presents a general architecture of Sensor Web applications; which can be characterised by:

- variability and heterogeneity of data, devices and networks (including unreliable nodes and links, noise, uncertainty, etc.);
- use of rich data sources (sensors, images, GIS, etc.) in different settings (live, streaming, historical, and processed);
- existence of multiple administrative domains; and
- need for managing multiple, concurrent, and uncoordinated queries to sensors.

* Corresponding author. E-mail: ocorcho@fi.upm.es.

[1] http://www.opengeospatial.org/projects/groups/sensorweb

[2] http://www.pachube.com/

[3] http://atom.research.microsoft.com/sensewebv3/sensormap/

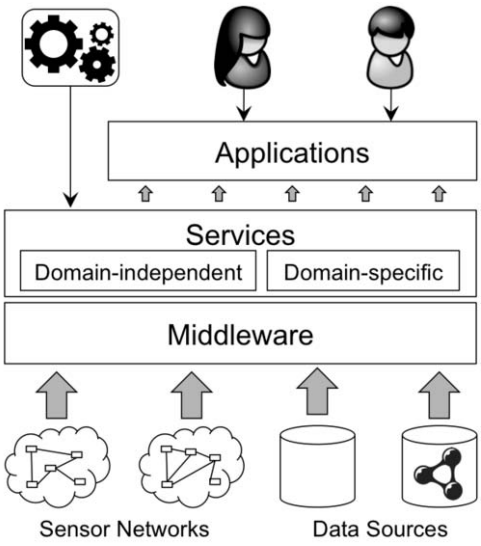

Fig. 1. Architecture of a Sensor Web application.

We will now review some of the most relevant challenges in this area, for which we will later propose descriptions of how semantic-based approaches could be applied.

2. Five challenges for Sensor Web applications

This section starts presenting those challenges in the area of the Sensor Web that have to do with the characteristics of the data sources that are handled in typical Sensor Web applications, and then moves into those challenges that have to do with the creation of applications based on these data sources. We do not aim at being exhaustive on the identification of challenges, but we hope that this categorization is useful to understand some of the open problems in this area.

One of the first challenges is related to the **abstraction level** in which sensor data can be obtained, processed and managed in general. Sensor data can be managed at a very low level, at the device- and network-centric levels, generally by means of using low-level programming languages and operating systems. But it can be also managed through higher-level formalisms (e.g., via declarative continuous queries over streams), thereby insulating clients and users from the infrastructural and syntactic heterogeneities of autonomously-deployed sensor networks.

Another challenge is related to the adequate characterisation and management of the **quality** (and **quality of service**) of sensor data. Issues like the unavailability of a piece of data over a period of time may have different meanings when seen from an application perspective: the sensor was not available, there was no event to trigger the data generation during that time, the communication with the sensor was broken, etc. Other issues like the accuracy of the sensed data may depend on a number of internal and external conditions to the sensor network. In summary, there are a number of quality characteristics that are relevant to the quality of service and that may affect the results obtained from a data observation process, normally with important trade-offs among each other (e.g., longevity vs. latency or completeness vs. throughput).

Another challenge has to do with the **integration and fusion of data** coming from autonomously-deployed sensor networks, with varying qualities of service and different throughput rates, geographical scales, etc. This is related not only with the integration of data coming from different sensor networks, but also with the combination of such data with data persisted in other sources, such as static data or archived sensor data.

Another challenge of utmost importance, related to the previous one, is the **identification and location of relevant sensor-based data sources** with which data integration and fusion tasks can be performed. The number of sensor networks being deployed in the real world is growing continuously, given the fact that the prices of hardware are decreasing. As a result, more experiments and initiatives deploy sensor networks in different (sometimes overlapping) areas, and finding the right information to be used in integration and fusion tasks is highly relevant.

Finally, another important challenge has to do with the need to enable the **rapid development of applications** that are able to handle sensor data, taking into account the aforementioned characteristics and challenges. This includes dealing with data integrity and validation issues as well as the need for common interfaces and formats between applications, databases, sensor networks, etc. This challenge requires enabling the development of applications with different resource models and qualities of service (e.g., energy, bandwidth, processing, storage) and facilitating the interaction with sensor data from the developer and user points of view.

3. Applying semantic-based approaches to Sensor Web challenges

In this section we provide brief descriptions of how the aforementioned challenges are being addressed in existing initiatives and projects, by means of semantic-based methods, techniques and technologies.

We start with the characterization of the abstraction level at which sensor data can be obtained, processed and managed. A number of **sensor network ontologies** have been defined in the literature [5], which aim at describing different aspects of sensor-based data, from the device point of view (focusing on the hardware that is being used in order to generate the data) to the domain point of view (focusing on the types of data that can be generated from sensors and sensor networks in the context of specific domains such as Health or Environment). Several aspects are relevant in the development of most of these ontologies, such as the distinction between raw observed data and derived data, the representation of aspects like accuracy, or the consideration of observations and measurements according to the relevant OGC models; the ontological representation of this last aspect has received attention on its own [8,9]. The development of an ontology in this area is one of the main tasks being performed in the W3C Incubator Group on Semantic Sensor Networks[4].

The aforementioned work on sensor network ontologies also takes into account the **quality of the data sources**, although it is not central to the work being performed in the context of the Incubator Group. Data quality is a large research area that is not only applicable to sensor-based data, but to any type of data that can be managed in an application. It is common to talk about data quality in relational databases, in semi-structured data sources, in user generated content, etc. Therefore, it is a property of data sources in general, and not of sensor-based data in particular. However, sensor-based data depends largely on the context of the sensor network, such as the network physical infrastructure, deployment strategy, or surrounding environmental conditions. This context may influence the quality of data (e.g., the accuracy of measurements) and has to be taken into account to correctly interpret them (e.g., to interpret the meaning of data gaps). Early work is being done on the definition of data quality models for this type of data, by categorising existing approaches for other types of sources and selecting and adapting them to the context of sensor networks. The same applies to the quality of service of sensor network sources, in terms of parameters that are also applied to other types of sources (e.g., reliability) and are specialized for sensor networks (e.g., reading rate, battery levels).

With respect to the integration and fusion of data, work has been done in the context of integrating and fusing heterogeneous data streams. Some of this work uses semantic techniques, and some does not. A recent research trend is focused on the generation of **Linked Data from sensor network data streams** [13,14] by means of transforming sensor-based data into RDF and making it available using HTTP by means of sensor-related URIs. This will allow the seamless navigation across sensor-based (and other types of) data. Other work is being done on the provision of **semantic queries** that are adapted to sensor-based data. They leverage declarative querying infrastructure to define logical views over sensor network data and open the way for view- and ontology-based techniques to be used. These approaches extend query languages like SPARQL with constructors normally applied to stream-based sources (e.g., time and tuple-based windows). Examples of such extensions are the C-SPARQL [2] or the Streaming SPARQL [3] languages, and an example of approaches that provide transformations between sensor data sources and these languages is the work described in [4].

In the context of identifying and locating relevant sensor-based data in the real world, work is being done on the definition of **sensor data registry interfaces**, and in the development of the appropriate infrastructure that can cope with the types of queries that are usually handled in sensor-based applications. These registries should provide support for spatio-temporal queries (e.g., "get sensor data sources that contain information about the temperature in this region for the last two days") and for metadata queries related to existing sensor network ontologies. Some work in this context can be found at [10].

Finally, another identified challenge is related to the development of **high-level application programming interfaces** (APIs) that ease **the rapid development of thin applications** (e.g., mashups) that use data from sensor networks and legacy databases. These programming interfaces should cope in a homogeneous way with the different types of data (persisted and streamed), support the use of the semantic extensions already identified (e.g., semantic-based descriptions of data, linked sensor

[4] http://www.w3.org/2005/Incubator/ssn/

data, semantic-based registries), and help users interact with and make sense of the potentially enormous and heterogeneous amounts of data coming from the Sensor Web. Examples of these interfaces are already available, although without much semantic support (e.g., SensorMap [12]) and some early work is also done to develop decision support systems for environmental management.

4. Conclusions and future work

In this paper we have described some challenges in the area of the Sensor Web and how these challenges are being addressed using semantic-based approaches.

We have covered issues that arise from the need to interpret, manage and integrate in a meaningful way data that is coming from heterogeneous sensor networks, with different levels of abstraction, different application areas, and different quality conditions. We have also described how applications that rely heavily on sensor-based data can be more flexibly created, and how they can make use of services to locate data sources that may not have been originally deployed for the specific purpose of the application.

Much work still remains to be done in all these areas, and also in others that have not been covered exhaustively in this position paper, such as event identification and management with sensor data or improved sensor network management using semantic techniques, to name a few.

Furthermore, the achievement of a Semantic Sensor Web is not a task to be made in isolation. We have shown how introducing semantics into the Sensor Web scenario presents new requirements over the Semantic Web specifications and technologies. Even if such requirements are currently being satisfied by extending these specifications and technologies, they can be a valuable input for advancement in the Semantic Web area that will, in turn, benefit the Semantic Sensor Web.

Acknowledgements

Some of these challenges are being addressed in the SemsorGrid4Env project[5], funded by the European Commission under grant FP7-223913. Other ongoing projects related to these challenges are: SENSEI[6], CONET[7], PECES[8], and ASPIRE[9].

A research agenda for the Semantic Sensor Web is also being discussed by the Future Internet Assembly working group on Real World Internet[10] and in the W3C incubator group on Semantic Sensor Networks.

Finally, we would like to thank all the SSG4Env project partners for their contributions to the identification and work on these challenges.

References

[1] K. Aberer, M. Hauswirth, and A. Salehi, A middleware for fast and flexible sensor network deployment. In *VLDB*, pages 1199–1202, 2006.

[2] D.F. Barbieri, D. Braga, S. Ceri, E. Della Valle, M. Grossniklaus, C-SPARQL: SPARQL for Continuous Querying. In *Proceedings of the 18th Int. World Wide Web Conference*, pages 1061–1062, 2009.

[3] A. Bolles, M. Grawunder and J. Jacobi, Streaming SPARQL – Extending SPARQL to process data streams. *The Semantic Web: Research and Applications*, pages 448–462, 2008.

[4] J.P. Calbimonte, O. Corcho, A.J.G. Gray, Enabling Ontology-based Access to Streaming Data Sources. *9th International Semantic Web Conference (ISWC2010). Shanghai, China,* November 2010.

[5] M. Compton, H. Neuhaus, K. Taylor, K.-N. Tran, A Survey of the Semantic Specification of Sensors. *Proceedings of the 2nd Int. Workshop on Semantic Sensor Networks (SSN09), Washington DC, USA,* October 26, 2009.

[6] M. Compton, C. Henson, H. Neuhaus, L. Lefort, A. Sheth, IrisNet: An Architecture for a World-Wide Sensor Web. *IEEE Pervasive Computing*, Volume 2, Number 4, October–December 2003.

[7] D. Havlik, G. Schimak, R. Denzer, B. Stevenot, Introduction to SANY (Sensors Anywhere) Integrated Project. *ENVIROINFO 2006*, Shaker, Graz, Austria, pages 541–546, 2006.

[8] C. Henson, J. Pschorr, A. Sheth, K. Thirunarayan, SemSOS: Semantic Sensor Observation Service. *Proceedings of the 2009 Int. Symposium on Collaborative Technologies and Systems (CTS 2009), Baltimore, MD, USA,* May 18–22, 2009.

[9] W. Kuhn, A Functional Ontology of Observation and Measurement. *3rd Workshop on Geosemantics (GeoS 2009), Mexico City, Mexico,* pages 26–43, December 3–4, 2009.

[10] K. Kyzirakos, M. Koubarakis, and Z. Kaoudi, Data models and languages for registries in SemsorGrid4Env. Deliverable D3.1 Version 1.0, SemSorGrid4Env, August 2009.

[11] J. Ledlie, J. Shneidman, M. Welsh, M. Roussopoulos, M. Seltzer, Open Problems in Data Collection Networks. *SIGOPS European Workshop, Leuven, Belgium,* September 2004.

[5] http://www.semsorgrid4env.eu/

[6] http://www.ict-sensei.org/

[7] http://www.cooperating-objects.eu/

[8] http://www.ict-peces.eu/

[9] http://www.fp7-aspire.eu/

[10] http://rwi.future-internet.eu/

[12] S. Nath, J. Liu, and F. Zhao, SensorMap for Wide-Area Sensor Webs. *IEEE Computer*, Volume 40, Number 7, pages 106–109, July 2007.

[13] K. Page, D. de Roure, K. Martinez, J. Sadler, O. Kit, Linked Sensor Data: RESTfully serving RDF and GML. *Proceedings of the 2nd Int. Workshop on Semantic Sensor Networks (SSN09), Washington DC, USA,* October 26, 2009.

[14] J. Sequeda, O. Corcho, Linked Stream Data: A Position Paper. *Proceedings of the 2nd Int. Workshop on Semantic Sensor Networks (SSN09), Washington DC, USA,* October 26, 2009.

[15] J. Shneidman, P. Pietzuch, J. Ledlie, M. Roussopoulos, M. Seltzer, M. Welsh, Hourglass: An Infrastructure for Connecting Sensor Networks and Applications. Harvard Technical Report TR-21-04.

Semantic Web 1 (2010) 105–110
DOI 10.3233/SW-2010-0006
IOS Press

User modeling and adaptive Semantic Web

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Michel Dumontier, Carleton University, Canada; Jie Tang, Tsinghua University Beijing, China

Lora Aroyo[a,*] and Geert-Jan Houben[b]
[a] *Computer Science, VU University Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands*
[b] *Web Information Systems, TU Delft, P.O. Box 5031, 2600 GA Delft, The Netherlands*

Abstract. Historically, personalization and adaptation have been important factors for the success of the Web and therefore they have been important topics in Web research. Many research efforts in the field of adaptive hypermedia and adaptive Web-based systems have resulted in solutions for user-adapted access to Web content, often in terms of systems that provide an adaptive hypermedia structure of content pages and hyperlinks. With pages and links that depend on the user, it is feasible to offer a high degree of personalization. Next to research into engineering and realizing adaptation, research into user modeling has been crucial for the success of adaptation. To apply the right adaptation it is necessary to know the user and her relevant properties and the research field of user modeling has focused on theories and techniques for eliciting knowledge about the user. Naturally, the research fields of adaptive Web-based systems and user modeling have always lived in close harmony. In order to create a similar success with personalization and adaptation in relation to the Semantic Web, adaptation and user modeling have to be redefined, with consequences for the research into these topics. In particular, the nature of user modeling changes significantly with the extended distribution and openness that we encounter on the Web of Data, with implications from problems studied in Web science. Promising research shows how Semantic Web-based solutions can aid in the representation of user properties for sharing and linking of user models. In this vision paper we outline the evolution of user modeling and adaptation in connection to the Semantic Web and list research questions and challenges for the relevant research fields.

Keywords: User modeling, adaptation, personalization, linked open data, Web science

1. Introduction

In this paper we outline the evolution we see and anticipate for the research fields of *adaptation* and *user modeling* in connection with the Semantic Web. Adaptation of content access to the user is by definition a process in which a description of the content and a description of the user are combined to decide whether and how to present the content to the user. On the Web this adaptation has typically been designed and realized in closed and scoped applications. The user modeling to represent relevant user descriptions has been characterized by the same assumptions. When we relate this to the Web of Data, new conditions and assumptions come into play. At the Semantic Web we have witnessed already the use of semantics in the integration of content, for the purpose of integrating and linking content between applications, but similarly, and perhaps more importantly, user modeling can and should also be aligned with the conditions and requirements of the new Web. Where the distributed and open nature of data has brought major advances for linking content, the same characteristics of distribution and openness find their way into user modeling.

In this vision we reflect on what is happening and sketch challenges and questions for the relevant research fields. In Section 2 we consider the concepts of adaptation and user modeling in their original context of the classical Web, before we turn to user modeling in relation to the Semantic Web in Section 3 and the corresponding challenges in Section 4.

* Corresponding author. E-mail: l.m.aroyo@cs.vu.nl.

2. Adaptation and user modeling

From the conception of the Web its hypertext-based nature triggered researchers to find ways to improve the nodes-and-links structure that gave the Web its success. With the observation that a single fixed hyperlink structure would not fit all users, the research field of adaptive hypermedia [4] investigated how the hyperlink structure could be made adaptive to the user, i.e. how the pages and hyperlinks could be made fitting for each single user.

This adaptation aims to present the best possible hyperlink structure to a user depending on relevant properties of the user, e.g., background, context, or goal. Educational applications have always been a good example for researchers to showcase adaptation solutions [2]. In educational applications often the users are students or learners for whom the acquired knowledge is a basis for adaptation. Imagine a teacher wants to present the student with relevant material to study a certain subject, then the teacher will create a structure of pages and hyperlinks that will make the student go through the material in a manner that satisfies the teacher's intentions and pedagogic principles. That structure will present the student with pages and links depending on the knowledge the student acquired before and during her study of the material. For this purpose, the hyperlink structure will include with pages and links preconditions that reflect the teacher's assumptions about the student's knowledge at that moment in the browsing.

In the research field of adaptive hypermedia [4] this approach has been studied in several other domains as well, e.g. e-commerce and tourism, and this has led to the development of systems that support the design and execution of adaptation in hypermedia-based information delivery. Due to the nature of the first trials and the technological hurdles that had to be taken, most research concentrated on systems with a well-defined and limited scope, to make it feasible for the system to "know" the user and how to respond to that. Obviously, the design of adaptation asks for a detailed understanding of the user's knowledge at the time and of the influence that the knowledge should have on the content to be presented. In "closed" applications and systems the design and execution of the adaptation proved to be already challenging enough for researchers to extensively investigate design and usage [20]. Later, the scoping was relaxed when the same approaches were being used at the Web [6,7].

Approaches for adaptation cannot be meaningfully applied without a thorough understanding of the user. That is why the field of user modeling [10] has concentrated on theory and techniques for the elicitation of user knowledge into user models that could serve as the basis for effective user-adaptation. Applying intelligent techniques to calculate relevant properties of the user for adaptation, for example in recommendation or teaching scenarios, researchers established theories and tooling for an accurate and relevant description of the user on the basis of the user's actions in the application.

In linking these two research fields [17], the use of an explicit user model is the classical approach for adaptive systems. Following the reference model from [12], a general view on adaptive systems is that a system contains a description of the domain content, i.e. a *domain model*, a description of the user, i.e. a *user model*, and a way to combine those two to adapt the content for a presentation fitting the user, i.e. an *adaptation model and engine*.

In many cases the user model overlays the domain model, meaning that the user knowledge is expressed as an overlay over the domain content. A good example from the educational scenario would be that the student's knowledge is expressed as a value attached to each domain subject reflecting the degree to which the student has learned the topic. The specific elements in a user model depend of course on the application, but aspects that we often see are history, background, preferences, knowledge level, goals and tasks, context of work, meta-cognitive skills, personality traits, affective states, and attitudes.

3. User modeling in a Web of Linked Data

With the content moving towards the Semantic Web, it is now interesting to see how the advances in the Semantic Web and the cloud of Linked Open Data [19] impact adaptation and in particular user modeling.

Where in the traditional adaptation approaches the adaptation was often confined to a single closed application, it is natural to try to share and integrate content data to profit from the investment in content made in multiple applications. In the evolution from adaptive hypermedia to adaptive Web-based systems this trend was already visible. Also, open hypermedia systems [3] show a similar approach where the linking is separated from the content data.

Similarly, the open corpus adaptive hypermedia systems [5] show how metadata-based approaches concentrate on content reuse and integration, with obvious connections to semantic techniques and languages. So, when it comes to integrating and linking data, adaptive applications do not differ from other applications and can equally well benefit from results obtained in Semantic Web research. Therefore, we concentrate in this vision paper further on the user modeling aspect.

3.1. Linking user knowledge

Semantic integration can of course also create benefits for the user knowledge. When adaptive applications have the opportunity to share user knowledge, for example in an educational setting the student's knowledge in a particular subject domain, then with richer and more relevant user knowledge from across application boundaries the applications can provide better adaptation.

This trend in semantic integration of user knowledge aligns with the trend in Web 2.0 and social networking where people share personal information.

Both trends show two aspects that are relevant for linking user model knowledge: the identification of the users and the representation of their properties.

3.2. User identification

When applications want to share information for and about users, they require mechanisms for the identification of users.

Identity-based protocols as OpenID[1] can be used for users to link their different identities on the Web. Systems can use authentication mechanisms, from basic http authentication to open protocols for secure API authorization like OAuth[2]. The Google Friend Connect[3] API exemplifies the use of OpenID (e.g. Yahoo) and OAuth to integrate existing login systems, registered users, and existing data with new social data and activities. It is based on open standards and allows users to control and share their data with different sites. The integration of social flows and data is realized via the OpenSocial[4] standard specification. The Facebook Platform uses the OAuth 2.0[5] protocol for authentication and authorization in Web applications (both desktop and mobile). The Facebook Connect extension makes it possible for users to "connect" their Facebook identity to any site by using trusted authentication and to also reuse, among others, their basic profile information and friends list around the Web.

Research like [9] presents an approach to enable interoperability of user-adaptive systems in a ubiquitous environment. It is centered around a semantics-based dialogue for exchanging user model and context data with focus on the user data clarification and negotiation tasks. Further, [8] looks at a framework for user identification for cross-system personalization. It exploits a set of identification properties that are combined using an identification algorithm.

3.3. User property representation

Contrary to the identification of users, for the representation of user properties there are hardly any standardized and generic solutions available. This is not a surprise of course, given the traditionally closed environment in which user model knowledge is created and used.

This representation issue can best be explained with an example. If for example in an educational setting a student's knowledge needs to be represented, then one often sees something like a value such as "well-learned" that is associated as the "degree of learning" with a subject like "Programming" or a value of "80%" for the "knowledge level" of a subject "Java". From these examples it is easy to see that for interoperability we need to align (a) the knowledge about the domain, e.g. about the domain concepts such as "Programming" or "Java", and (b) the knowledge *about* the knowledge about the domain, e.g. the "degree of learning" or "knowledge level" and their corresponding values.

Important aspects of the representation of user properties are

(1) to represent *uniquely the object* of the interest or preferences, e.g. "interested in *Java*" or "likes *Brad Pitt*",

(2) to provide a *shared vocabulary* to express different *user activities* which translate into user properties, e.g. *"like", "read", "view", "favor"*,

(3) to have a *shared scale(s)* to interpret the user property, e.g. "rate this video with *5* stars" and "rate

[1] http://openid.net

[2] http://oauth.net

[3] http://code.google.com/apis/friendconnect/

[4] http://www.opensocial.org/

[5] http://tools.ietf.org/html/draft-ietf-oauth-v2-10

this book with 2 stars" – it is handy to know that the first value was in a 10 point scale and the second in a 5 point scale, which makes them almost identical in terms of their value for the aggregated user interest,

(4) to represent the *notion of certainty and accuracy* of the collected and aggregated user properties, e.g. indicating the source or the context of the collected information and specifying how trustworthy or reliable the source is – if the user had bought a book on Amazon about Java programming or watched a video on TED about it, these could be pretty reliable sources for her interest in this topic, while if she just browsed through several web pages it might be questionable whether she actually read anything.

A good example for most of those aspects can be found in current extensions of FOAF[6], e.g. the Weighted Interest Vocabulary[7] for identifying context and source of the collected information, or e-FOAF[8] for defining temporal properties for the interest value. Additionally, a format like Activity Streams[9] is used for syndicating social activities on the Web and providing a shared vocabulary to express user activities across applications. This format has already been adopted by Facebook, MySpace, Windows Live, Google Buzz, BBC, Opera, Gowalla, among others. The base schema defines a set of Verbs, e.g. "mark as Favorite", "post", "tag", a set of Object Types, e.g. "article", "bookmark", "comment", a set of Activity Context Elements, e.g. for location and mood, and Event Verbs, e.g. "positive RSVP".

e-FOAF allows for temporal reasoning in the user interest value calculation over time. Knowing when a specific piece of evidence for the user interest has occurred (`e-foaf:interest_appear_time`) or when the interest value was updated (e.g. `e-foaf:interest_value_updatetime`) helps to increase the accuracy in the calculation of the user interest as an aggregation (`foaf:cumulative_interest_value`) of multiple pieces of evidence around the Web. Additionally, properties as `e-foaf:retained_interest_value`, help express decay or other time-related aspects of the interest values. With properties like `e-foaf:interest_longest_duration` and `e-foaf:interest_cumulative_duration` the strength of the interest can be varied in order to reflect the intensity of the evidence in terms of calculating the cumulative user interest value.

Semantic Web-based user model standards that are used widely in educational settings are IMS LIP[10] and IEEE PAPI[11].

3.4. Sharing adaptation functionality

For the sake of completeness we mention that following [12]'s reference model, after *domain knowledge* and *user knowledge* linking and integration, the integration of *adaptation functionality* can also be improved but this is an extremely challenging problem given the proprietary nature of many of the currently available solutions and systems. Projects like [15] show first steps in the integration of adaptation functionality, where also semantic technologies are used albeit mainly for domain and user knowledge.

With the advances in the Web of Data for linking domain and content knowledge and for linking user model knowledge as we have just described in this section, we see however the emergence of a new paradigm for adaptation, where adaptive applications are connected to a cloud of Linked (content) Data as well as a cloud of Linked User Data. While this Linked User Data is technically part of the new Web of Data as well, the *distributed and open* nature puts a whole new perspective on user modeling, and opens a whole new array of innovations.

4. Distributed and Open User Modeling

For sharing user model knowledge, experience from the Semantic Web can provide concrete solutions, as for example the research from [1,2,9,18] shows. The main benefit is that each application does not have to build up its user model knowledge alone, which specially in the beginning can be a problem when little knowledge is available: the so-called *cold start*. Also, with more knowledge available to construct a model the chance that the model *accurately* describes the user increases naturally.

This sharing of user knowledge is not just a matter for applications that like to adapt (as we discussed in the previous section), but practically the

[6] http://www.foaf-project.org/

[7] http://xmlns.notu.be/wi/#spec/20091224.html

[8] http://wiki.larkc.eu/e-foaf:interest

[9] http://activitystrea.ms/

[10] http://www.imsglobal.org/profiles/

[11] http://www.ieee.org

same ambition and techniques show in the Social Web when a user wants to share her own personal profiles between social networking systems, as for example [12] shows. In [14] it is shown how FOAF can be used to link social networks. Where existing social networking services are highly centralized, as are existing personalized services, the trend towards distribution is clearly visible and helps also to increase the control users have over their own data.

Another example is provided by the NoTube project [12] where a strategy for aggregating user data from various Social Web applications is provided. This work is based on a concrete use case of reusing activity streams to determine a user's interests, and then generating television programme recommendations from these interests. A key component to realize this is the NoTube BeanCounter [21]. The main design rationale is to provide a flexible and extensible architecture that exposes robust, scalable and reliable services to handle different kinds of responses of different social application platforms. A set of APIs allows for modeling the targeted responses in order to gather them, represent them with a set of suitable RDF vocabularies, and integrate them with other pulled information in a fully transparent way – with the help of service-specific adaptor *tubelets* (e.g. a twitter activity stream tubelet) and application server *modelets* (e.g. for movies or songs) which allow for the selection of data source and RDF vocabulary and the generation of RDF-ized user data, integrated with data coming from other adaptors.

Observing the promising results in user modeling for adaptive and social applications with the aid of Semantic Web-based solutions, we now outline research questions and challenges for the new paradigm of *Distributed and Open User Modeling* as a main ingredient for the Adaptive Semantic Web.

4.1. User identification

The major question in user identification is: *How do we identify a person (or a person's appearance)?*

In the conventional adaptive solutions, closed and with restricted scope, identification mechanisms are often proprietary or pragmatic, and usually these are also not fit for application at Web-scale. The new assumptions and requirements imply a number of research questions:

- How can a person identify herself to an application?
- How can a person manage her identities (across multiple applications)?
- How can applications find a user (identity) in other applications?
- How are trust and privacy provided in mechanisms for user identification?
- How do users behave in systems with shared user identification and what are the social and legal consequences?

The above questions do include technical challenges, but also constitute interesting problems in Web Science. Considering the specific Semantic Web angle we see that standard identification mechanisms and efficient corresponding indexing mechanisms need to be proposed.

4.2. User knowledge alignment

After users being identified, the main question with respect to user knowledge is: *How can user model knowledge be shared?*

Answering his question on the Web of Data brings up several research questions related to the representation of user properties:

- How can the objects of user properties uniquely be represented?
- How can a shared vocabulary be constructed for expressing user properties?
- How can shared scales be constructed to interpret values for user properties?
- How can notions of certainty and accuracy be attached to user properties?

The main challenge here for the user modeling research field is to derive from the vast amount of user modeling theories and experience [17], those properties that are typically and effectively used to model user properties and then turn to the area of the Semantic Web to create a standard vocabulary for those properties. Such a vocabulary could borrow from SKOS, RDF/OWL etc., as we see in some of the examples we mentioned before.

Besides its role for sharing user model knowledge, such a vocabulary-based approach would also be the ideal stepping-stone for Web Scientists to study user modeling on the Web of Data, and thus to analyze how user modeling performs under the new conditions of distribution and linking. As part of this

[12] http://notube.tv

study, openness and scrutability need to be investigated as well:

- How is an open approach to user knowledge perceived by the users?
- How can users be given the opportunity to inspect and correct the user knowledge an application maintains about them?

The role of the user in the elicitation and verification of user knowledge can also be extended, with the aid of semantic techniques, following [11].

Thus, we see the first examples of investigations into the new paradigm for user model knowledge.

5. Conclusion

In this paper we have considered how user modeling evolves from its original environment in connection to a closed and scoped adaptive system to a distributed and open existence in the Semantic Web. We have identified some promising research approaches that show how the Semantic Web can contribute with solutions for user identification and user knowledge representation. On the basis of that experience, we have formulated for the new paradigm a number of relevant engineering and scientific questions for inclusion in the research agenda.

References

[1] F. Abel, D. Heckmann, E. Herder, J. Hidders, G.J. Houben, D. Krause, E. Leonardi and K. van der Sluijs, A Framework for Flexible User Profile Mashups. In *Proceedings of APWEB 2.0 2009, workshop in conjunction with UMAP 2009* (2009).

[2] L. Aroyo, P. Dolog, G.J. Houben, M. Kravcik, A. Naeve, M. Nilsson, and F. Wild, Interoperability in personalized adaptive learning. *J. Educational Technology & Society*, **9**(2) (2006) 4–18.

[3] C. Bailey, W. Hall, D. Millar and M. Weal, Towards open adaptive hypermedia. In *Proceedings of AH 2002, Adaptive Hypermedia and Adaptive Web-Based Systems* (2002) 36–46.

[4] P. Brusilovsky, Adaptive hypermedia. *User Modeling and User Adapted Interaction*, **11** (1/2) (2001) 87–110.

[5] P. Brusilovsky and N. Henze, Open Corpus Adaptive Educational Hypermedia, in: *The Adaptive Web*. *Springer Lecture Notes in Computer Science* **4321**, 2007, 671–696.

[6] P. Brusilovsky, A. Kobsa and W. Nejdl (Eds.), The Adaptive Web. *Springer Lecture Notes in Computer Science* **4321**, 2007.

[7] P. Brusilovsky and M.T. Maybury, From adaptive hypermedia to the adaptive web. *Commun. ACM* **45**(5) (2002) 30–33.

[8] F. Carmagnola and F. Cena, User identification for cross-system personalisation. *Information Sciences* **179**(1–2) (2009) 16–32.

[9] F. Cena and L. Aroyo, A Semantics-Based Dialogue for Interoperability of User-Adaptive Systems in a Ubiquitous Environment. In *Proceedings of UM 2007* (2007) 309–313.

[10] C. Conati, K.F. McCoy and G. Paliouras (Eds.), User Modeling (UM 2007). *Springer Lecture Notes in Computer Science* **4511**, 2007.

[11] R. Denaux, V. Dimitrova and L. Aroyo, Integrating open user modeling and learning content management for the semantic web. In *Proceedings of UM 2005, User Modeling* (2005) 9–18.

[12] P. De Bra, G.J. Houben and H. Wu, AHAM: A Dexter-Based Reference Model for Adaptive Hypermedia. In *Proceedings of ACM Hypertext 1999* (1999) 147–156.

[13] R. Ghosh and M. Dekhil, Mashups for semantic user profiles. In *Proceedings of WWW 2008* (2008) 1229–1230.

[14] J. Golbeck and M. Rothstein, Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. In *Proceedings of AAAI 2008* (2008) 1138–1143.

[15] Grapple project: www.grapple-project.org.

[16] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz and M. von Wilamowitz-Moellendorff, GUMO – The General User Model Ontology. In *Proceedings of UM 2005, User Modeling* (2005) 428–432.

[17] G.J. Houben, G.I. McCalla, F. Pianesi and M. Zancanaro (Eds.), User Modeling, Adaptation, and Personalization (UMAP 2009). *Springer Lecture Notes in Computer Science* **5535**, 2009.

[18] T. Kuflik, Semantically-Enhanced User Models Mediation: Research Agenda. In *Proceedings of UbiqUM'2008, workshop in conjunction with IUI 2008* (2008).

[19] Linked Data: www.linkeddata.org.

[20] W. Nejdl, J. Kay, P. Pu and E. Herder (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2008). *Springer Lecture Notes in Computer Science* **5149**, 2008.

[21] C. van Aart, L. Aroyo, D. Brickley, V. Buser, L. Miller, M. Minno, M. Mostarda, D. Palmisano, Y. Raimond, G. Schreiber, and R. Siebes, The NoTube Beancounter: Aggregating User Data for Television Programme Recommendation. In *Proceedings of SDoW 2009, workshop in conjunction with ISWC 2009* (2009).

Semantic Web 1 (2010) 53–59
DOI 10.3233/SW-2010-0007
IOS Press

Inductive learning for the Semantic Web: What does it buy?

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA
Solicited review(s): Philipp Cimiano, CITEC, Universität Bielefeld, Germany; Bernardo Cuenca Grau, Oxford University, UK
Open review(s): Pascal Hitzler, Wright State University, USA

Claudia d'Amato *, Nicola Fanizzi and Floriana Esposito
Department of Computer Science, University of Bari, Italy
E-mail: claudia.damato@di.uniba.it, fanizzi@di.uniba.it, esposito@di.uniba.it

Abstract. Nowadays, building ontologies is a time consuming task since they are mainly manually built. This makes hard the full realization of the Semantic Web view. In order to overcome this issue, machine learning techniques, and specifically inductive learning methods, could be fruitfully exploited for learning models from existing Web data. In this paper we survey methods for (semi-)automatically building and enriching ontologies from existing sources of information such as Linked Data, tagged data, social networks, ontologies. In this way, a large amount of ontologies could be quickly available and possibly only refined by the knowledge engineers. Furthermore, inductive *incremental learning* techniques could be adopted to perform reasoning at large scale, for which the deductive approach has showed its limitations. Indeed, incremental methods allow to learn models from samples of data and then to refine/enrich the model when new (samples of) data are available. If on one hand this means to abandon sound and complete reasoning procedures for the advantage of uncertain conclusions, on the other hand this could allow to reason on the entire Web. Besides, the adoption of inductive learning methods could make also possible to dial with the intrinsic uncertainty characterizing the Web, that, for its nature, could have incomplete and/or contradictory information.

Keywords: Ontology mining, inductive learning, uncertainty

1. Introduction

The Semantic Web (SW) [3] view is grounded on the availability of domain ontologies to be used for semantically annotating resources. Most of the time ontologies are manually built thus resulting in a highly time consuming task that could undermine the full realization of the SW. For this reason several Machine Learning (ML) methods have been exploited to automatize the ontology construction task [23,28,33]. The main focus is on (semi-)automatically building the *terminology* of an ontology while less attention has been dedicated to the enrichment/construction of the *assertional* part, namely the *ontology population problem*, which results in an even more time consuming task.

In last few years, this problem has been tackled by customizing ML methods such as instance based learning [37] and Support Vector Machine (SVM) [40] to Description Logics (DLs) [1] representation that is the theoretical foundation of OWL[1] language which is the standard representation language in the SW. Specifically, the problem is solved by casting the ontology population problem to a classification problem where, for each individual in the ontology, the concepts (classes) to which the individual belongs to have to be determined [5,8,14].

Both methods for building terminology and assertions only marginally dial with another important problem that emerged in the last few years: "how

*Corresponding author.

[1] http://www.w3.org/TR/owl-features/

to manage the inherent uncertainty[2] of the Web"[3]. To face this problem, some proposals have been formulated. They mainly concern with: how to represent uncertain knowledge [27,30,32] and how to reason in presence of uncertain knowledge [10,31,42]. However, they generally assume that a probabilistic and/or fuzzy knowledge base already exists. Inductive learning methods could be used to build probabilistic knowledge bases by learning the probability that: an inclusion axiom, a relationship between two individuals, a concept assertion hold. Indeed, differently from deductive reasoning (generally adopted in the SW context) where, given a set of general axioms, correct and certain conclusions are drawn by the use of a formal proof, inductive reasoning has as input specific examples/data from which a possible/plausible generalization is computed. This generalization is also able to predict the behavior (i.e. the classification) of new and not previously observed examples.

Reasoning on ontological knowledge plays an important role in the SW since this allows to make explicit some implicit information (e.g. concept and role assertions, subsuption relationships). However, in presence of noisy/inconsistent knowledge bases, that could be highly probable in a shared and distributed environment such as the Web, deductive reasoning is no more applicable since it requires correct premises. On the other hand, inductive reasoning is grounded on the generalization of specific examples (assertions in the SW context) rather than correct premises, thus allowing the formulation of conclusions even when inconsistent/noisy knowledge bases are considered.

In this paper, we survey some inductive learning methods specifically focussing on their applicability for solving various *ontology mining* problems. For *ontology mining* we mean all those activities that allow to discover hidden knowledge from ontological knowledge bases (most of the time concept and role assertions are considered), by possibly using only a *sample* of data. The discovered knowledge could be exploited for building/enriching ontologies. Specifically, we envision the applicability of inductive methods for:

- learning new relationships among individuals
- learning probabilistic ontologies
- (semi-)automatizing the ontology population task
- learning probabilistic mapping for the ontology matching task
- refining ontologies
- reasoning on inconsistent/noisy knowledge bases

In the following some these aspects are analyzed. Particularly, in the Section 2 an overview of existing ML methods that have been exploited for solving some ontology mining problems is presented. Proposals on how existing inductive learning techniques can be exploited for facilitating the realization of the SW view are presented in Section 3. Conclusions are drawn in Section 4.

2. The present of ontology mining

One of the first proposals for automatically building terminologies is the *ontology learning* task [33]. It focuses on learning ontologies (mainly terminologies) from text documents by the use of clustering methods (drawn from Formal Concept Analysis (FCA) [18]) and association rules [22]. Concepts are extracted from documents by the use of Natural Language Processing techniques [34]. Hence, they are clustered to obtain an initial terminology which is further enriched with new relationships (not necessarily taxonomical) by means of association rules. The main limitations of this approach are: 1) the semantic relations among the terms are not fully clear; 2) the expressiveness of the adopted language is less than OWL.

In order to obtain more expressive knowledge bases, different approaches have been set up [23,26,28]. They assume the availability of an initial sketch of ontology that is automatically enriched by adding and/or refining concepts. The problem is solved as an unsupervised learning problem where given a set of positive and negative examples for the concept to learn, namely a set of individuals that are known to be respectively instances of the concept to learn and instances of the negation of the concept to learn, the goal is building a concept description such that all positive examples are instances of it while all negative examples are not instances.

As regards (semi-)automatizing the ontology population task, the problem has been focused by casting it to a classification problem. Given the concepts of an ontology, all individuals are classified with respect to each concept. In [8,16], the *Nearest Neighbor (NN)* approach [37] is adopted. A new instance (individual) is classified by selecting its most similar training ex-

[2] With the term "uncertainty", a variety of aspects are meant such as incompleteness, vagueness, ambiguity.

[3] http://www.w3.org/2005/Incubator/urw3/

amples (existing individuals in the knowledge base) and by assigning it the class (concept) that is majority voted among the training examples. This required to cope with: 1) the *Open World Assumption (OWA)* rather than the usual *Closed World Assumption (CWA)* generally adopted in ML; 2) the non-disjointness of the classes (since an individual can be instance of more that one concept at the same time) while, in the usual ML setting, classes are generally assumed to be disjoint; 3) the availability of new similarity measures to exploit the expressiveness of DLs.

In [5,12,14], a similar approach is adopted. The main difference is given by the use of SVM [40] rather than *NN* to perform the classification. SVM efficiently classifies instances by implicitly mapping, by the use of a kernel function, the training data and the input values in a higher dimensional feature space where instances can be classified by means of a linear classifier. The application of SVM to DLs representation required the definition of suitable kernel functions to cope with the language expressiveness.

A similar underlying idea has been exploited in [2] where FCA [18] has been used for completing both the terminological and the assertional part of an ontology.

Most of these approaches have also been adopted for performing inductive concept retrieval and query answering, namely for determining the set of individuals that are instance of an existing concept or of a concept generated on the fly from the existing concepts and relationships in the ontology. This is done by classifying all individuals in the ontology with respect to the considered concept. The interesting results of using inductive methods have been: 1) a very low error rate; 2) the ability to induce new knowledge, namely new assertions that are not logically derivable. They can be suggested to the knowledge engineer that has only to validate them. Moreover, most of the inductive methods that have been applied to ontological representation (e.g. NN or SVM) have polynomial complexity which would allow to scale on the whole Web.

3. Inductive learning for the future of Semantic Web

The adoption of inductive approaches for ontology mining is mainly motivated by the necessity of: a) semi-automatize the mining of the assertional part of an ontology (i.e. the ontology population task); b) overcoming the limitations showed by deductive reasoning in the SW context [44], namely its inability to: 1) scale on large ontologies; 2) reason on uncertain knowledge; 3) exploit data regularities. On the contrary, induction can be defined as the process of learning from data. In the following, an overview of how some existing inductive learning methods can be exploited for performing several ontology mining tasks is presented.

3.1. Inductive learning for building ontologies from folksonomies and Linked Data

A first fruitful usage of inductive approaches is to automatically build ontologies from source of information such as folksonomies and Linked Data [39]. Indeed, besides of the plethora of text documents and Web pages that are used as input for the *ontology leaning* process [6,19], folksonomies and Linked Data are becoming so popular to constitute a non-negligible source of knowledge. We envision the process of learning ontologies from folksonomies and Linked Data as structured in the following the three steps.

1. *Annotated data are clustered* to create meaningful groups. Well known clustering algorithms such as *K-Means*, *DB-SCAN*, *Simulated Annealing* [24] could be used. Clustering methods are generally grounded on the notion of similarity. Given a set of data, the goal of clustering methods is to find clusters that have high intra-cluster similarity and low inter-cluster similarity [37]. Different approaches could be used: hierarchical, partitional or fuzzy. Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called *dendrogram*[4]. The root of the tree consists of a single cluster containing all data, and the leaves correspond to individual data. Partitional clustering determines all clusters at once, generating a flat set of clusters. Both hierarchical and partitional methods usually assume that clusters are disjoint. On the contrary fuzzy clustering methods allow non-disjoint clusters: an instance can belong, with a certain degree of membership, to more than one cluster at the same time. Applying hierarchical (fuzzy) clustering methods (such as *K-Means* algorithm) to Linked Data, a taxonomy is obtained. It could represent a sketch of an ontology that

[4] A dendrogram is a nested grouping of patterns and similarity levels at which grouping changes. The dendrogram could be broken at different levels to yield different clustering of the data.

is populated with the resources to which Linked Data refer to. However, similarity measures that are able to cope with Linked Data representation need to be exploited. Moreover, at the current stage, no intentional concept definitions are available in the sketch of the ontology. In order to avoid this issue, the second step of the proposed process has to be taken into account.

2. *Concept descriptions for the taxonomy can be learnt* by the use of *conceptual clustering* methods [17] whose goal is to give intensional descriptions of the discovered clusters. Most of the conceptual clustering algorithms such as INC, COBWEB, CLUSTER/2 [17,21,36] often exploit generalization operators applied to propositional representations to set up intentional descriptions of the discovered clusters. The application of conceptual clustering methods to Linked Data will necessarily require the definition of new generalization operators that are able to cope with the considered representation. However, at this stage of our learning process, mainly a taxonomy is available. In order to enrich it with new and potentially more expressive concepts and relationships, the third step of the process has to be considered.
3. Some *data mining techniques such as association rules [22] can be used to further discover frequent patterns* both in a single cluster or in the entire data set. These pattern can be seen as positive examples for a concept (or a relation) to learn via a supervised learning process. However, a supervised learning process usually needs also negative example for the concept to learn. The availability of negative examples could be problematic because of the *OWA*. Indeed, differently from the *CWA* (usually adopted in ML) where negative examples are intended as those examples that are not instance of the concept to learn, in the *OWA* generally adopted in the SW context, negative examples should be instance of the negation of the concept to learn[5]. In this situation, where negative examples could be hardly determined, methods for learning from positive (and unlabeled) examples [7,48] only can be exploited.

[5]The problem does not exist if the examples are labelled by an expert as positive and negative examples of a concept (or relationship) to learn. However, this is not really realistic in an open and wide environment such as the Web.

3.2. Class-imbalance learning for concept retrieval and ontology population

As discussed in [2,5,8], inductive learning can be exploited for (semi-)automatizing the ontology population task by casting this problem to a classification problem and by classifying each individual in the ontology with respect to each concept in the ontology itself. The same approach could be adopted for performing inductive concept retrieval and query answering, namely for assessing all individuals that are instances of an existing concept or of a query concept that is built on the fly by composing (for instance via conjunction and/or disjunction) existing concepts. Induced assertions, namely assertions that cannot be logically derived, could be used for enriching the assertional part of an ontology.

However, as it has been experimentally shown [8, 11], this approach could be less reliable when individuals are not homogeneously spread in the ontology, namely when they are mainly instances of a subset of the concepts in the ontology while the remaining concepts have very few instances. In a setting like this, methods such as NN, that performs classification on the ground of the majority voted class among the most similar training examples, would fail. For instance, considering a case in which 97% of training examples belong to a class A and only 3% of them belong to another class, it will be highly probable that most of the time the classification result will be the class A. *Class-imbalance learning methods* [20,29,47] can be exploited to avoid this problem. They are generally used for performing classification in presence of *imbalanced data sets* [20,29,47], namely data sets where the number of examples of one class is much higher than the others. By the use of sampling techniques, class-imbalance learning methods first create a balanced dataset, namely a data set where instances are homogeneously spread among all categories, and then perform the inductive classification task.

3.3. Inductive learning for ontology refinement

Another important task is ontology refinement. Manually performing ontology refinement could turn out to be a very complex task, particularly for large ontologies. In order to (semi-)automatize this task, *learning Decision Trees* methods [37] could be interestingly used for the purpose. Given a set of positive and negative example for a concept to learn, they return a tree from which a concept description is induced.

The application of these methods in the SW context requires: the specification of positive, negative and unlabeled[6] examples (to cope with the OWA) and the exploitation of refinement operators for DL representations [23,28] giving as a output a *Terminological Decision Tree*[7] from which a new concept definition is derived [15]. Hence the ontology can be refined/enriched by adding the new concept or the whole tree, thus introducing a fine granularity level in the concept descriptions (some tentatives in this direction have been presented in [45,46]). Moreover, *Terminological Decision Trees* can be also exploited for classifying individuals with respect to the learnt concept thus having an alternative way for performing inductive concept retrieval and query answering [5,8].

3.4. Inductive learning for ontology evolution

Another interesting problem that can be tackled via inductive reasoning is *ontology evolution*. Indeed ontologies are not static, they evolve over the time, because new concepts are added (TBox evolution) or most of the time because new assertions are added (ABox evolution). Particularly, the ABox evolution could introduce new concepts that are only extensionally defined while their intentional definitions are missing. *Conceptual clustering* algorithms [17] can be crucial for discovering such kind of evolution [13]. Specifically, they can be employed for discovering *concept drift* or *the formation of new emerging concepts* in an ontology. In order to do this, all instances of the ontology are clustered and an overall evaluation of the clusters (called *global decision boundary*) is computed by the use of well known metrics such as Dunn's Index, Silhouette index, generalized medoid [4,25,38]. A new set of instances is considered as a candidate cluster. To determine its nature, namely if it represents a new concept, a concept drift or an already existing concept, the evaluation of the candidate cluster is performed and it is compared with the *global decision boundary*. If this evaluation is lower than the *global decision boundary* than the candidate cluster is assessed as being an existing concept otherwise it is assessed to represent a new/evolving concept. In the latter case, the intentional cluster description (that is a concept description) can be learned and then merged (by the use of the subsuption relationship) in the ontology. Furthermore, methods for tracking cluster transitions could be also exploited [41].

3.5. Incremental inductive learning for scaling on large ontologies

The interest in inductive reasoning and inductive learning methods is not only motivated by the fact that they allow to discover concepts and relationships that cannot be deductively derived. The other main reason is given by the limitation that the deductive approach has showed on reasoning at large scale. To cope with this problem *incremental* inductive learning methods [35,43] are particularly suitable. Indeed, these methods do not need the whole set of data. They are able to learn a first model from a sample of the available training examples and then to update the model when new examples are available. This could allow to learn ontologies, for example, by sampling the Web. Specifically, given an initial sample of the Web, a first (set of) ontology (ontologies) is learnt and then continuously updated when new instances are available. Moreover, differently from the deductive approach that cannot be applied to inconsistent knowledge bases, inductive reasoning is able to process data even in presence of inconsistent or noisy knowledge bases [8,9], situation that could be quite common in an open and heterogeneous environment such as the Web.

3.6. Inductive learning for building probabilistic ontologies

As showed in [8,11], inductive classification can be effectively exploited for performing inductive concept retrieval and query answering. Since the conclusions drawn from inductive reasoning are typically uncertain, this can be explicitly treated, that is the probability of an inductive result (for instance an individual belonging to a certain concept) could be computed. The explicit treatment of the uncertain results gives several advantages: 1) users or applications can have a measure of the reliability of the inductive results; 2) computed probabilities can be exploited for ranking the answers of a query; 3) a new way of formulating queries which include the chance of requiring likely information/event can be considered [44], i.e. a query of kind *finds all persons that live in Italy that are employees and are likely to own a Ferrari* could be treated; 4) probabilistic ontologies can be automatically built.

[6] Because of the *OWA*, for some instances could be not possible to assess if they belong to a certain concept or its negation so the case of unlabeled example has to be considered.

[7] A terminological decision tree is a decision tree from which DL concept description can be learnt.

Particularly, the last point refers to an another interesting open problem in the SW context: how to manage uncertainty. Even if some existing works have tackled the problem [10,27,32,42], mainly they focus on: (a) how to represent uncertain knowledge; (b) how to reason with uncertain knowledge. Almost all of them assume the availability of uncertain/probabilistic knowledge bases. Building probabilistic ontologies could be a task even more hard than building ontologies. The inherent uncertainty of inductive results could be effectively exploited for the purpose. For instance, the classification results for performing inductive concept retrieval can be accompanied by the probability values for which a certain result is true. Such probabilities can be exploited for building probabilistic ontologies by adopting a framework such as the one proposed in [32].

4. Conclusions

The role of inductive reasoning for ontology mining has been analyzed. A summary of the inductive methods currently adopted in ontology mining has been presented, hence a set (of potential new) ontology mining problems have been addressed and proposals for suitable inductive methods, jointly with a brief analysis of the issues to solve, have been done. The applications of inductive methods for learning probabilistic ontologies is considered one of the most challenging and interesting problems. Moreover, methods for learning event probabilities can be also exploited for assessing probabilistic mapping in the ontology matching task.

References

[1] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, editors. *Description Logic Handbook*. Cambridge University Press, 2003.

[2] F. Baader, B. Ganter, B. Sertkaya, and U. Sattler. Completing description logic knowledge bases using formal concept analysis. In M. Veloso, editor, *IJCAI 2007, Proc. of the Int. Joint Conference on Artificial Intelligence*, pages 230–235, 2007.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American Magazine*, 2001.

[4] J. C. Bezdek and N. R. Nikhil. Cluster validation with generalized dunn's indices. In *Proc. of the Int. Conference on Artificial Neural Networks and Expert Systems, ANNES '95*, page 190. IEEE Computer Society, 1995.

[5] S. Bloehdorn and Y. Sure. Kernel methods for mining instance data in ontologies. In K. Aberer et al., editors, *Proc. of the International Semantic Web Conference*, volume 4825 of *LNCS*, pages 58–71. Springer, 2007.

[6] P. Buitelaar and P. Cimiano, editors. *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, volume 167. Frontiers in Artificial Intelligence and Applications, 2008.

[7] F. De Comité, F. Denis, R. Gilleron, and F. Letouzey. Positive and unlabeled examples help learning. In *Algorithmic Learning Theory: 10th Int. Conf., ALT'99*, volume 1720 of *LNCS*. Springer, 1999.

[8] C. d'Amato, N. Fanizzi, and F. Esposito. Query answering and ontology population: An inductive approach. In S. Bechhofer et al., editors, *Proc. of the 5th European Semantic Web Conference*, volume 5021 of *LNCS*, pages 288–302. Springer, 2008.

[9] C. d'Amato, N. Fanizzi, B. Fazzinga, G. Gottlob, and T. Lukasiewicz. Combining semantic web search with the power of inductive reasoning. In F. Bobillo et al., editors, *Proc. of the International Workshop on Uncertainty Reasoning for the Semantic Web at ISWC 2009*, volume 527 of *CEUR Workshop Proceedings*, pages 15–26. CEUR-WS.org, 2009.

[10] C. d'Amato, N. Fanizzi, and T. Lukasiewicz. Tractable reasoning with bayesian description logics. In S. Greco and T. Lukasiewicz, editors, *Scalable Uncertainty Management, Second International Conference, SUM 2008*, volume 5291 of *LNCS*, pages 146–159. Springer, 2008.

[11] C. d'Amato, F. Nicola, and F. Esposito. Analogical reasoning in description logics. In *Uncertainty Reasoning for the Semantic Web I: Revised Selected and Invited Papers*, pages 330–347. Springer-Verlag, 2008.

[12] N. Fanizzi and C. d'Amato. A declarative kernel for $\mathcal{ALC}$ concept descriptions. In F. Esposito et al., editors, *Foundations of Intelligent Systems, 16th International Symposium, ISMIS 2006, Bari, Italy, September 27–29, 2006, Proceedings*, volume 4203 of *LNCS*, pages 322–331. Springer, 2006.

[13] N. Fanizzi, C. d'Amato, and F. Esposito. Conceptual clustering and its application to concept drift and novelty detection. In *Proc. of the Europ. Semantic Web Conference*, volume 5021 of *LNCS*, pages 318–332. Springer, 2008.

[14] N. Fanizzi, C. d'Amato, and F. Esposito. Statistical learning for inductive query answering on owl ontologies. In A. P. Sheth et al., editors, *International Semantic Web Conference*, volume 5318 of *LNCS*, pages 195–212. Springer, 2008.

[15] N. Fanizzi, C. d'Amato, and F. Esposito. Towards the induction of terminological decision trees. In *Proc. of the Symposium on Applied Computing*, volume 2, pages 1424–1428. ACM, 2010.

[16] Nicola Fanizzi, C. d'Amato, and F. Esposito. Instance-based query answering with semantic knowledge bases. In R. Basili and M. T. Pazienza, editors, *AI*IA 2007: Artificial Intelligence and Human-Oriented Computing, 10th Congress of the Italian Association for Artificial Intelligence, Rome, Italy, September 10–13, 2007, Proceedings*, volume 4733 of *LNCS*, pages 254–265. Springer, 2007.

[17] D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, **2**:139–172, 1987.

[18] B. Ganter and R. Wille, editors. *Formal Concept Analysis: Mathematical Foundations*. Springer-Verlag, 1998.

[19] M. Grobelnik and D. Mladenić. Knowledge discovery for ontology construction. In *Semantic web technologies: trends and research in ontology-based systems*, pages 9–27. John Wiley and Sons, 2006.

[20] H. Guo and V. L. Herna. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *SIGKDD Explor. Newsl.*, **6**(1):30–39, 2004.

[21] M. Hadzikadic and D. Y. Y. Yun. Concept formation by incremental conceptual clustering. In *Proc. of the International Joint Conference on Artificial Intelligence (IJCAI'89)*, pages 831–836, 1989.

[22] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations*, **2**(2):1–58, 2000.

[23] L. Iannone, I. Palmisano, and N. Fanizzi. An algorithm based on counterfactuals for concept learning in the semantic web. *Appl. Intell.*, **26**(2):139–159, 2007.

[24] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, **31**(3):264–323, 1999.

[25] L. Kaufman and P. J. Rousseeuw, editors. *Finding Groups in Data: an Introduction to Cluster Analysis*. John Wiley and Sons, 1990.

[26] J. Kietz and K. Morik. A polynomial approach to the constructive induction of structural knowledge. *Machine Learning*, **14**(1):193–217, 1994.

[27] M. C. A. Klein, P. Mika, and S. Schlobach. Rough description logics for modeling uncertainty in instance unification. In F. Bobillo et al., editors, *Proc. of the 3rd ISWC Workshop on Uncertainty Reasoning for the Semantic Web*, volume 327 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[28] J. Lehmann and P. Hitzler. Concept learning in description logics using refinement operators. *Machine Learning*, **78**:203–250, 2010.

[29] X. Liu, J. Wu, and Z. Zhou. Exploratory under-sampling for class-imbalance learning. In *Proc. of the Int. Conf. on Data Mining*, pages 965–969. IEEE Computer Society, 2006.

[30] T. Lukasiewicz. Expressive probabilistic description logics. *Artif. Intell.*, **172**(6–7):852–883, 2008.

[31] T. Lukasiewicz. Uncertainty reasoning for the semantic web. In *Web Reasoning and Rule Systems, Int. Conf.*, volume 5837 of *LNCS*, pages 26–39. Springer, 2009.

[32] T. Lukasiewicz and U. Straccia. Managing uncertainty and vagueness in description logics for the semantic web. *J. Web Sem.*, **6**(4):291–308, 2008.

[33] A. Maedche and S. Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, **16**(2):72–79, 2001.

[34] C. D. Manning and H. Schütze, editors. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[35] R. S. Michalski, I. Mozetic, J. Hong, and N. Lavrac. The multipurpose incremental learning system aq15 and its testing application to three medical domains. In *AAAI*, pages 1041–1047, 1986.

[36] R. S. Michalski and R. E. Stepp. Automated construction of classifications: Conceptual clustering versus numerical taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **5**:396–410, 1983.

[37] T. Mitchell, editor. *Machine Learning*. McGraw Hill, 1997.

[38] P. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, **20**(1):53–65, 1987.

[39] N. Shadbolt, T. Berners-Lee, and H. Hall. The semantic web revisited. *IEEE Intelligent Systems*, **21**(3):96–101, 2006.

[40] J. Shawe-Taylor and N. Cristianini, editors. *Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[41] M. Spiliopoulou, I. Ntoutsi, Y. Theodoridis, and R. Schult. Monic: modeling and monitoring cluster transitions. In *Proc. of the Int. Conf. on Knowledge Discovery and Data Mining*, pages 706–711. ACM, 2006.

[42] U. Straccia. Reasoning within fuzzy description logics. *J. Artif. Intell. Res. (JAIR)*, **14**:137–166, 2001.

[43] N. A. Syed, H. Liu, S. Huan, L. K., and K. Sung. Handling concept drifts in incremental learning with support vector machines. In *Proc. of the Int. Conf. on Knowledge Discovery and Data Mining*, pages 317–321. ACM, 1999.

[44] V. Tresp, M. Bundschus, A. Rettinger, and Y. Huang. Towards machine learning on the semantic web. In P. C. G. da Costa et al., editors, *Uncertainty Reasoning for the Semantic Web I: Revised Selected and Invited Papers*, volume 5327 of *LNCS*, pages 282–314. Springer, 2008.

[45] J. Völker, P. Hitzler, and P. Cimiano. Acquisition of owl dl axioms from lexical resources. In *Proc. of the European Semantic Web Conference (ESWC 2007)*, LNCS, pages 670–685. Springer, 2007.

[46] J. Völker, D. Vrandecic, Y. Sure, and A. Hotho. Learning disjointness. In *Proc. of the European Semantic Web Conference (ESWC 2007)*, LNCS, pages 175–189. Springer, 2007.

[47] P. Yao. Comparative study on class imbalance learning for credit scoring. *Hybrid Intelligent Systems, International Conference on*, **2**:105–107, 2009.

[48] B. Zhang and W. Zuo. Learning from positive and unlabeled examples: A survey. In *International Symposium on Information Processing*, pages 650–654, 2008.

Semantic Web 1 (2010) 83–88
DOI 10.3233/SW-2010-0008
IOS Press

Accessing the Web of Data through embodied virtual characters

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Kunal Verma, Accenture, USA; Tom Heath, Talis, UK

Philipp Cimiano [a,*] and Stefan Kopp [b]
[a] *Semantic Computing Group, Cognitive Interaction Technology Excellence Center (CITEC), Bielefeld University, Germany*
[b] *Sociable Agents Group, Cognitive Interaction Technology Excellence Center (CITEC), Bielefeld University, Germany*
E-mail: kopp@cit-ec.uni-bielefeld.de

Abstract. The amount of data published on the Semantic Web has witnessed a tremendous growth in the last years to which the Linked Open Data (LOD) project has contributed significantly. While the Semantic Web was originally conceived of as an extension to the Web by addition of machine-readable data allowing automatic processing by machines, the question how humans can benefit from all the data published on the Web is becoming an important one. In the light of this question it seems crucial to make accessing the data on the Web as easy and intuitive as possible by adapting to the cognitive and information processing capabilities of humans. In this short position paper, we argue that one interesting and promising approach in this direction is to allow people to access semantic data on the Web through multimodal interaction with embodied virtual characters.

Keywords: Semantic web, interaction, multimodal access, virtual characters

1. Introduction

The amount of data published on the Web of Data (a.k.a. The Semantic Web) has witnessed a tremendous growth in recent years. The Linked Open Data (LOD) project[1] has contributed significantly to this growth. People are in fact massively following Tim Berners-Lee's advice to publish data on the Web following the "linked data" principles [3,5]. Linked Data is a term referring to the recommended best practices for exposing, sharing, and connecting RDF data via dereferenceable URIs on the Semantic Web. While the Semantic Web was originally conceived as an extension to the Web by the addition of machine-readable data allowing automatic processing by machines, the question how humans can benefit from all the data published on the Web is certainly becoming a more and more important one.

In the light of this question it seems crucial to make accessing the data on the Web as easy and intuitive as possible. One central concern is to adapt to the cognitive and information processing capabilities of humans by making interaction on the Semantic Web *"meaningful"* for the user. Clearly, this is not only a user interface question. While suitable user interfaces are definitely yet to be seen, appropriate interaction paradigms for the Web of Data need to provide answers to the following questions:

- How much of the RDF graph should remain visible to end users? Should we fully abstract from the RDF data model/graph? It has been argued that the graph is actually not particularly useful as a way to present semantic data [28]. While some tools for accessing the Web of Data stick quite closely to the (linked) data graph (see Tabulator

*Corresponding author. E-mail: cimiano@cit-ec.uni-bielefeld.de.
[1] http://linkeddata.org/

[4]), it seems important to abstract from the data graph when interacting with end users. After all, why should users care about the data model if all they want is relevant information?
- How should information be packaged? That means, which are the information units that users can handle optimally? At one end of the spectrum, we have a single triple (arguably the smallest information unit on the Web of Data), at the opposite end we can come up with complete (multimodal) presentations generated by integrating various resources, aggregating data, computing diagrams, etc.
- What is the ideal interaction paradigm to access the Web of Data? Keyword querying [33], browsing [4], query-by-example, natural language [26] or even by way of reciprocal conversation between the user and the interface [14]?
- How can users understand non-trivial concepts such as trust, provenance, confidence etc? What are appropriate metaphors to convey such meta-information?

Providing answers to the above questions does clearly not only pertain to research on mere user interface design, but rather constitutes a non-trivial and long-term endeavor for the Semantic Web and related fields of research. Developing new and effective paradigms for interacting with the Semantic Web has indeed been recognized as a key challenge in the field (see [16–18]).

2. Motivation

We argue that one interesting and promising approach is to allow people to access semantic data on the Web through multimodal communication with embodied virtual characters. Consider that you would like to get information about the relation between the two French painters *Claude Monet* and *Edouard Manet*. Suppose also that you would like to get an overview about all painters considered as impressionists or that you would like to receive information about US presidents in chronological order. A natural way to pose such queries to a system is by way of natural language. And, as the Web of Data is structured – in contrast to the traditional Web – we can indeed provide answers to such information needs by fetching and re-composing different pieces of data available. Such a composition of available information into new informational structures that meet a current information need would be much more difficult on the traditional Web as it requires to understand the textual content first.

The even more important question for our purposes here is: which kind of structure would we like to get back as answer to such a request? An unordered and unstructured set of triples crawled from the Web of Data? Certainly not. Rather it seems crucial to find approaches that allow to assemble the relevant triples into a logical and coherent structure that can be conveyed to users. Imagine that you have a virtual character as assistant that you have posed the above query to compare Monet and Manet. The character would provide the following spoken answer along with different non-verbal modalities (we highlight the ouput in non-textual modalities in bold font):

> **[Agent displays three photos of Claude Monet, Edouard Monet and one of Paris in the 19th century, respectively]** *"Both Claude Monet* **[points to the photo of Claude Monet]** *and Edouard Manet* **[points to photo of Manet]** *were French painters born in Paris in the 19th century* **[points to photo of Paris]**. *Monet was born on the 14th of November 1840, whereas Manet was born earlier on the 23rd of January of 1832. While both are associated with the Impressionism movement, Monet is also considered to belong to the realism movement. The most important works of Monet include Impression Sunrise, Rouen cathedral, London Parliament, Water Lilies and Poplar Series* [**agent sequentially blends in pictures of all these works, synchronized with its speech**]. *The most important works of Manet include "The Lunch on the grass".*

The strengths of such an approach to accessing the Web of Data by virtual embodied characters can be clearly appreciated: by packaging information into different modalities and units (e.g. sentences in speech) that people are used to from everyday conversation, we can generate a structured, yet compact, concise and amenable presentation. There are a number of further benefits, which we discuss with respect to the issues raised in the introduction:

- **Abstracting from the RDF data model:** It has been argued that the RDF graph is actually not particularly useful as a way to present semantic data [28]. Virtual characters are a promising way to realize a human-tailored access that abstracts from the RDF data model in order to transform

the information into units that can be presented via different natural modalities (speech, text, gestures etc.).

- **Multimodal communication as interaction paradigm:** Conveying multimodal output (using speech, intonation, gestures or facial expressions) allows to package information more effectively and compactly as different types of information can be conveyed across different suitable channels in parallel. As a corollary, this will lead to information packages that are closer to the information units that people are used to process and assimilate in daily interaction with human partners.
- **Tangible notion of a mediator:** Virtual characters are known to be entertaining and to increase the motivation to interact with a system [24]. In the difficult situation of wanting (or having) to access the abstract body of knowledge contained in the Semantic Web, we hypothesize that the presence of a virtual character can be beneficial because it makes tangible the notion of an assistant who is there to help users in finding the relevant information. Also, being able to formulate an information need in natural language is a custom and natural way for humans (and is increasingly supported by search engines, e.g. WolframAlpha[2], or by natural language interfaces to the Semantic Web, see [20,27]). Virtual characters are known to be social actors in the sense that they elicit the willingness to apply natural interaction patterns [10,23,24].
- **Expressing meta-information:** An embodied virtual character can use gestures, appropriate facial expressions, prosody and intonation, or linguistic modifiers such as *possibly* or *probably* to make clear that trust in a certain bit of information is low. As trust is an important building block of the Semantic Web (see [1,9,12,13,30,32]), conveying trust levels to human users becomes a crucial issue. Using gestures along with appropriate linguistic modifiers enables natural communication of such qualifiers together with the actual content, in a way that is more intuitive than presenting lists of items ranked by confidence or other symbolic or numerical representations of confidence values or trust levels.

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

3. Related work

Developing new and effective paradigms for user interaction with the Semantic Web has been recognized as a key challenge in the field. Heath et al. [18] have identified the following challenges that *"must be addressed if Semantic Web technologies are to enter into widespread usage"*: i) increasing awareness, ii) providing clear benefits and iii) delivering appropriate functionality, iv) giving guidance for users, v) improving usability, vi) ensuring coherence of Semantic Web applications and vii) creating a critical mass of participation. In fact, there have been several workshops on this topic since 2004, e.g. IDWS[3], EUSW[4], SWUI'07[5], SWUI'08[6], SWUI'09[7]. Some authors of papers at these workshops have already proposed that conversational interaction with the Web of Data is an important interaction paradigm to explore (see [14] and [11]). However, there have only been quite preliminary approaches [7,21] which allow such an access and even extend it to the use of embodied virtual characters. For example, Kimura and Kitamura [21] directly embed RDF queries into utterance rules specified in the chatterbot markup language AIML. This simple approach allows for responding to a certain input phrase with a fixed utterance in which predetermined parts are replaced with retrieved fragments, but it represents by no means a flexible and comprehensive method to collect semantic data and to turn this into coherent multimodal presentations to satisfy the user's information need.

It is important to emphasize that we are not stating that an approach to access the Semantic Web/Web of Data by way of embodied virtual characters will solve all of the challenges raised by Heath et al. [18]. However, as argued above, an approach based on embodied virtual characters has the potential to provide access to the Web of Data in an intuitive and natural manner and thus to *improve usability*. Given the massive amount of data available, techniques that gather this data and generate intuitive and appealing summaries are addressing a clearly defined user need and deliver a *clear benefit* and *appropriate functionality*. Heath et al. claim that in order for semantic technologies to *increase in awareness* and receive widespread adoption we would need

[3]http://interaction.ecs.soton.ac.uk/idsw04/
[4]http://www.ifi.uzh.ch/ddis/iswc2005ws.html
[5]http://swui.semanticweb.org/swui2007/
[6]http://swui.webscience.org/SWUI2008CHI/
[7]http://swui.webscience.org/SWUI2009/

to hide the label "Semantic Web" and convey the fact to users that technology is doing useful things. The interaction with embodied virtual characters as we propose here would indeed contribute to making the technology and data models used behind the scenes transparent to the user while focusing on the system's presence as a helpful and useful assistant. This capitalizes on the fact that an agent can naturally provide assistance and guidance to the user in case he/she is experiencing problems, and can provide an enjoyable interaction which has the potential to increase the participation toward a *critical mass*.

As argued for by Dickinson [11], many problems for which the Semantic Web or Linked Open Data is useful are inherently exploratory in nature, where the users start out with a vague idea of what to look for and develop further insights into the nature of the enquiry during the interaction. This is exactly the type of incremental interaction that we aim to support and elicit with an embodied conversational character. It has been further argued for by Dickinson that *"turn-taking dialogues are a natural fit for the iterative exploration moves in exploratory search"* [11]. We agree with Dickinson that access though conversation is the most natural interaction type possible for humans engaging in exploration tasks. The fact that conversation-based access to the Semantic Web is an alternative with high potential is also argued for by Golbeck and Mutton, who allow users to access services from Internet Relay Chat (IRC) [14].

4. Challenges and roadmap

We have highlighted the benefits of accessing the Web of Data by way of embodied virtual characters. However, this is a daunting endeavor requiring remarkable progress in a number of areas. We can, however, point out some of the challenges that need to be addressed:

- **Selecting and packaging information intro narratives:** An important challenge is to develop approaches that i) select the right information from the Web of Data to satisfy a user's information need, ii) construct plans how to convey this information in different modalities and iii) generate coherent narrative structures as output. The latter requires the generation of discourses beyond single sentences, which has been partially addressed in the language generation community (see [19,29,34]). The biggest challenge is to accomplish this robustly without requiring a fixed data schema to allow for scaling up to the size and heterogeneity of the Web of Data.
- **Verbalization of Information:** Conversational access to the Semantic Web requires that we are able to verbalize RDF data, possibly in different languages. We hence need generation algorithms that can exploit linguistic knowledge captured in models such as LexInfo [8] about how data elements are to be realized linguistically in order generate language output. First approaches to verbalize semantic data have been presented [6] but are restricted to very rigid schemas and require manual effort by the user to adapt the system.
- **Generation of appropriate non-verbal behavior:** Flexibly producing gestures, facial expressions, or head movements that can accompany other modalities (speech, audio, video, text) is a non-trivial problem and subject of ongoing research [2]. In our context, one key challenge is to extract the information from a semantic resource that allows for generating behaviors that communicate differently from speech, e.g. gestures modulating or complementing it with imagistic or indexical information.
- **Language-based interfaces:** We need to support language-based interaction between the user and the conversational agent that goes beyond mere question-answering functionality (see [26]). One key issue is to be able to interpret language input for the user's information need, which is most often only possible by embedding it in the context of previous requests and presentations.
- **Synchronizing different modalities:** While expressing information through different channels (modalities) allows to compactly and efficiently apportion and encode information, synchronizing the different channels becomes crucial and is a big challenge (see [22]).
- **Robust Dialog Management in large domains:** An important challenge for providing conversational access to the Semantic Web is to be able to implement robust dialog management strategies that do not follow a fixed schema but allow, e.g., for clarification requests or repairments of misunderstanding. Thus, slot-filling techniques that have been developed in the dialog system community [25] are less suited in this context. Dialog management systems that can flexibly cope with arbitrary domain data are needed here. A first ap-

proach in this direction proposing to use a multimodal dialog system to access semantic data can be found in [15].

– **User acceptance:** One challenge is that embodied characters bear the risk of raising expectations with the users that the systems cannot live up to thus leading to annoyance or frustration from the side of the user. We argue that by smart design of the character and its interaction capabilities, and by a user-centered approach to developing character-based interfaces to the Web of Data, it may be possible to find the balance between what users demand from the system (e.g. full natural language conversation), what the character evokes by its appearance and behavior, and what it actually delivers.

While all of these issues are open research questions, we think that they are worthwhile to explore. We can conceive of a step-wise development towards the ultimate vision of an embodied virtual character that takes a query and answers it like a human expert in the respective field. At first, we will see agents that understand relatively simple requests and can automatically generate a simple discourse (possibly applying a limited set of templates). This first generation of agents might be already able to generate simple non-verbal output, synchronizing it with the speech modality and have basic mechanisms for conveying trust levels. The interaction with the user will be most likely text-based rather than via speech and there will be no mixed-initiative interaction, i.e, the agents will merely react to the input of a user. The technology for such characters is already available and building them is mainly a matter of system construction and attunement to the Semantic Web domain. Then, we will see agents that implement simple patterns of interaction and are able to engage in clarification dialogues. Simple mixed-initiative dialogs in selected domains have been realized already (see for example the project Gossip Galore which aims at developing conversational agents providing users access to pop trivia [35]). Finally, we might have reached a state that allows us to engage in conversation in selected domains and to receive multimodal information presentations from the character that are informative and tailored to the context. Robustness might be achieved by data-driven techniques which acquire script knowledge via games with a purpose [31], by observation, or via trial-and-error.

5. Conclusion

Providing meaningful interaction paradigms to access the Web of Data is an important topic for the Semantic Web community. We have suggested that providing access to the Semantic Web through multimodal interaction with embodied virtual characters is an interesting avenue to explore. Besides preliminary case studies, there has not been extensive research on this topic so far. As is clear from the challenges mentioned in this article, developing systems that provide conversational access to the Web of Data requires techniques and knowledge from a number of disciplines (dialog management, natural language processing, information retrieval, multimedia processing, virtual agents, etc.). Thus, we see it as a vision to which different research fields could (and should) contribute to and can cross-fertilize each other by doing so.

Certainly, "the one" interaction paradigm which fits all purposes and users does not exist. We thus think that it is only through appropriate user studies that we will be able to find out which (combination of) interaction paradigms are suited for which purpose. Conversational access to the Semantic Web might be one of them, possibly most suitable for casual users wanting to explore the Web of Data.

References

[1] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, **5**(2):58–71, 2007.

[2] K. Bergmann and S. Kopp. Gnetic – using bayesian decision networks for iconic gesture generation. In *Proceedings of Intelligent Virtual Agents (IVA09)*, pages 76–89. Springer-Verlag, 2009.

[3] T. Berners-Lee. Linked data. http://www.w3.org/DesignIssues/LinkedData.html, 2006.

[4] T. Berners-Lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets. Tabulator: Exploring and analyzing Linked Data on the Semantic Web. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.

[5] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, **5**(3):1–22, 2009.

[6] K. Bontcheva. Generating tailored textual summaries from ontologies. In *Proceedings of the European Semantic Web Conference (ESWC)*, pages 531–545, 2005.

[7] A. Breuing, T. Pfeiffer, and S. Kopp. Conversational interface agents for the semantic web – a case study. In *Proceedings of the 7th International Semantic Web Conference (ISWC2008)*, 2008.

[8] P. Buitelaar, P. Cimiano, P. Haase, and M. Sintek. Towards linguistically grounded ontologies. In *Proceedings of the 6th Annual European Semantic Web Conference (ESWC)*, pages 111–125, 2009.

[9] J. Carroll, C. Bizer, P. J. Hayes, and P. Stickler. Named graphs, provenance and trust. In *Proceedings of the World Wide Web (WWW) Conference*, pages 613–622, 2005.

[10] D.M. Dehn and S. van Mulken. The impact of animated interface agents: A review of empirical research. *Int. J. Human-Computer Studies*, **52**:1–22, 2000.

[11] Ian Dickinson. In favour of (more) intelligence in the semantic UI. In *Proceedings of the Semantic Web User Interaction Workshop (SWUI'09)*, 2009.

[12] R. Queiroz Dividino, S. Sizov, St. Staab, and B. Schüler. Querying for provenance, trust, uncertainty and other meta knowledge in RDF. *Journal of Web Semantics*, **7**(3):204–219, 2009.

[13] J. Golbeck, P. A. Bonatti, W. Nejdl, D. Olmedilla, and M. Winslett, editors. *Proceedings of the ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*. CEUR-WS.org, 2004.

[14] J. Golbeck and P. Mutton. Semantic web interaction on internet relay chat. In *Proceedings of the Workshop on Interaction Design on the Semantic Web*, 2004.

[15] Y. He, T.T. Quan, and S.C. Hui. A multimodal restaurant finder for semantic web. In *Proceedings of the IEEE International Conference on Computing and Communication Technologies (RIVF)*, 2006.

[16] T. Heath. How will we interact with the web of data? *IEEE Internet Computing*, **12**(5):88–91, 2008.

[17] T. Heath. A taskonomy for the Semantic Web. *Semantic Web Journal – Interoperability, Usability, Applicability*, **1**(1,2):75–81, 2010.

[18] T. Heath, J. Domingue, and P. Shabajee. User interaction and uptake challenges to successfully deploying semantic web technologies. In *Proceedings of the 3rd International Semantic Web User Interaction Workshop*, 2006.

[19] E.H. Hovy. Automated discourse generation using discourse structure relations. *Artif. Intell.*, **63**(1–2):341–385, 1993.

[20] E. Kaufmann and A. Bernstein. How useful are natural language interfaces to the Semantic Web for casual end-users? In K. Aberer, K.-S. Choi, N. Noy, D. Allemang, K.-I. Lee, L.J.B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudré-Mauroux, editors, *Proceedings of the 6th International and 2nd Asian Semantic Web Conference (ISWC/ASWC)*, volume 4825 of *Lecture Notes in Computer Science*, pages 281–294. Springer, 2007.

[21] M. Kimura and Y. Kitamura. Embodied conversational agent based on Semantic Web. In *Agent Computing and Multi-Agent Systems*, LNCS 4088, pages 734–741. Springer-Verlag, 2006.

[22] S. Kopp, K. Bergmann, and I. Wachsmuth. Multimodal communication from multimodal thinking – towards an integrated model of speech and gesture production. *Semantic Computing*, **2**(1):115–136, 2008.

[23] S. Kopp, L. Gesellensetter, N. Krämer, and I. Wachsmuth. A conversational agent as museum guide – design and evaluation of a real-world application. In *Intelligent Virtual Agents*, LNAI 3661, pages 329–345. Springer-Verlag, 2005.

[24] N. Krämer. *Soziale Wirkung virtueller Helfer*. Kohlhammer, 2008.

[25] O. Lemon, K. Georgila, J. Henderson, and M. Stuttle. An ISU dialogue system exhibiting reinforcement learning of dialogue policies: generic slot-filling in the talk in-car system. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstrations (EACL'06)*, pages 119–122. Association for Computational Linguistics, 2006.

[26] V. Lopez, M. Pasin, and E. Motta. Aqualog: An ontology-portable question answering system for the semantic web. In *Proceedings of the European Semantic Web Conference (ESWC)*, pages 546–562, 2005.

[27] V. Lopez, V.S. Uren, M. Sabou, and E. Motta. Cross ontology query answering on the Semantic Web: An initial evaluation. In Y. Gil and N. Noy, editors, *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP)*, pages 17–24. ACM, 2009.

[28] mc schraefel and D. Karger. The pathetic fallacy of rdf (position paper). In *Proceedings of the Semantic Web User Interaction Workshop (SWUI'06)*, 2006.

[29] K.R. McKeown. *Text generation: using discourse strategies and focus constraints to generate natural language text*. Cambridge University Press, New York, NY, USA, 1985.

[30] K. O'Hara, H. Alani, Y. Kalfoglou, and N. Shadbolt. Trust strategies for the semantic web. In *Proceedings of the ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*, 2004.

[31] J. Orkin and D. Roy. Automatic learning and generation of social behavior from collective human gameplay. In *Proceedings of the 8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, pages 385–392, 2009.

[32] M. Richardson, R. Agrawal, and P. Domingos. Trust management for the semantic web. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 351–368, 2003.

[33] T. Tran, H. Wang, S. Rudolph, and P. Cimiano. Top-k exploration of query candidates for efficient keyword search on graph-shaped (RDF) data. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 405–416, 2009.

[34] L. Wanner. On lexically biased discourse organization in text generation. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 369–378, 1994.

[35] F. Xu, P. Adolphs, H. Uszkoreit, X. Cheng, and H. Li. Gossip galore: A conversational web agent for collecting and sharing pop trivia. In *Proceedings of the International Conference on Agents and Artificial Intelligence (ICAART)*, pages 115–122, 2009.

Semantic Web 1 (2010) 137–141
DOI 10.3233/SW-2010-0009
IOS Press

Privacy in ontology-based information systems: A pending matter

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Michel Dumontier, Carleton University, Canada; Paulo Pinheiro da Silva, The University of Texas at El Paso, USA

Bernardo Cuenca Grau *
Oxford University Computing Laboratory, Oxford, United Kingdom
E-mail: berg@comlab.ox.ac.uk

Abstract. OWL ontologies are extensively used in the clinical sciences, with ontologies such as SNOMED CT being a component of the health information systems of several countries. Preserving privacy of information in ontology-based systems (e.g., preventing unauthorised access to system's data and ontological knowledge) is a critical requirement, especially when the system is accessed by numerous users with different privileges and is distributed across applications. Unauthorised disclosure, for example, of medical information from SNOMED-based systems could be disastrous for government organisations, companies and, most importantly, for the patients themselves. It is to be expected that privacy-related issues will become increasingly important as ontology-based technologies are integrated in mainstream applications. In this short paper, I discuss several challenges and open problems, and sketch possible research directions.

Keywords: Ontologies, semantic web, data privacy

1. Background

Ontologies expressed in the Web Ontology Language (OWL) or its revision OWL 2 are already being used for applications in fields as diverse as biomedicine, astronomy and defence. For example, OWL ontologies are extensively used in the clinical sciences, with ontologies such as SNOMED CT being a component of health information systems of several countries.

OWL ontologies can be used to formally describe the meaning of data (e.g., electronic patient records in the case of a medical application). Applications can then exploit ontologies to process the associated data in a more intelligent way. For example, a medical ontology describing patient record data may contain information such as "every patient with a mental disorder must be treated by a psychiatrist", "schizophrenia is a kind of psychosis", and "psychosis is a kind of mental disorder"; if John's medical record states that he suffers from schizophrenia, then an ontology can be used to conclude that he suffers from a mental disorder and must be treated by a psychiatrist.

Preserving privacy of the information in ontology-based systems (e.g., preventing unauthorised access to system's data and ontological knowledge) is a critical requirement, especially when the system is accessed by numerous users with different privileges and is distributed across applications. In particular, there may be multiple groups of users who want to access and retrieve information from the same ontology and its associated data sources. In this setting, different access rights may be granted to each of these groups and privacy preservation implies, for example, ensuring that users can only retrieve (either directly or indirectly via logical inference) the information they are allowed to access. The unauthorised disclosure, for example, of medical information from SNOMED-based systems (e.g., the identity of schizophrenic patients in

*The author is supported by a Royal Society University Research Fellowship.

a hospital) could be disastrous for government organisations, companies and, most importantly, for the patients themselves.

Data privacy in information systems is a long standing research area, which is particularly active in databases (DBs) (e.g., [3–5,10,20,23]). Very little is known, however, about privacy in the context of OWL ontologies and only recently has research been conducted in this direction [2,7,8,24].

Existing work on data privacy in databases focuses mainly on *complete* relational DBs [4,10,19,20]. Ontologies, however, are strongly related to *incomplete* DBs [18,22], with the difference that ontology languages are typically much more expressive than DB schema languages.

In contrast to complete DBs, query evaluation requires taking into account *all* models of the incomplete DB (or ontology) to compute the *certain answers* to the query formula (that is, the answers logically inferred by the union of the schema/ontology and the data). In our previous example, the fact that John suffers from a mental disorder and must be treated by a psychiatrist is not explicitly given; however, it can be deduced as a consequence of given information. These inferences may involve non-obvious interactions between different pieces of information in the system.

Data privacy in the context of incomplete or semi-structured DBs has only recently been investigated [5,12]. Furthermore, these works do not consider the presence of complex dependencies such as the ones present in OWL ontologies.

It is to be expected that privacy-related issues will become increasingly important as ontology-based technologies are integrated in mainstream applications. In the remainder of this paper, I discuss several challenges and open problems, and sketch possible research directions.

2. General challenges

In my discussion, I will focus on two general challenges for future research. The first one is related to the *design* of a privacy-preserving ontology-based system, whereas the second one concerns the *(re)use* of such system by external applications.

To illustrate the first challenge, consider the information system of a hospital whose privacy policy should prevent Bob from accessing the relationship between patients and their medical conditions. In DBs, access control has traditionally been achieved by presenting users with (relational) *views* that omit the sensitive information (e.g., the table relating patients to medical conditions) [1,14]. In the case of ontologies, however, providing a view that filters out such explicit statements may not be sufficient to ensure privacy.

Suppose that Bob knows that John has only been in the hospital once and, on that occasion, he was treated by both Dr. Smith (a gastroenterologist) and Dr. Andrews (a psychiatrist); from the ontology Bob knows that gastroenterologists only treat gastric diseases and psychiatrists only treat either mental disorders or psychosomatic illnesses; moreover, a disease cannot be both a mental disorder and a gastric disease and, if a disease is both psychosomatic and gastric, it must be a form of irritable bowel syndrome. Bob could then infer that John suffers from a kind of irritable bowel syndrome. Thus, restricted information can be leaked via logical inference.

Therefore, the first challenge is the development of the theoretical foundations and practical techniques necessary for the *design* of systems that provide provable privacy guarantees as well as to gain an understanding of the limitations of these guarantees.

The second challenge follows from the previous discussion, which suggests that access to information in an ontology-based system providing privacy guarantees should be *restricted* (i.e., the system's ontology and data cannot be published, at least not entirely). In the case of our previous example, to comply with the privacy requirements the system should not make public any information that would lead an external user to infer that John suffers from a kind of irritable bowel syndrome.

The system's owners may be reluctant to even publish the non-confidential information; for example, they may not be willing to distribute the contents of the ontology (even if data access is restricted), as doing so might allow competitors to plagiarise it; also, they might want to impose different costs for reusing parts of the ontology. This is the case with SNOMED CT, which is only available under a license agreement.

Currently, the only way for ontology-based applications to (re)use other ontologies and data sources is by means of OWL's *importing* mechanism [15]. OWL tools deal with imports by internally merging (i.e., constructing the union of the contents of) the relevant ontologies and the relevant data sources; hence the use of OWL's importing mechanism requires physical access to the entire contents of a system. If these contents are not available due to access limitations, the use of OWL's importing mechanism is clearly no

longer possible. As a consequence, further research is needed in order enable the effective (re)use of a privacy-preserving system by external applications.

Therefore, the second challenge is to investigate the conditions under which an application can effectively *(re)use* an ontology-based system to which access limitations have been imposed due to privacy considerations.

3. Design of a privacy-preserving system

In this section, I argue that the design of a privacy-preserving ontology-based system requires addressing the following issues:

1. *Policy representation*: How can system designers establish in a declarative way what information should be inaccessible to which users?
2. *Models of interaction*: What kinds of queries can users pose to the system?
3. *Formalisation of users' prior knowledge*: How could system designers take into account the knowledge that users may already have acquired when querying the system (e.g., the results obtained from previous queries) and which could be used to violate the policies?
4. *Notions of policy violation*: What does it mean for users to discover, by interacting with the system and using their prior knowledge, information that is confidential according to the policy applied to them?

A privacy policy specifies, in a declarative way, which information should not be accessible to which users (or group of users defined, for example, according to a role-based access model) [4]. An important issue is to establish the way in which policies are to be represented by the designer of the system.

In the database theory literature, policies are often represented using various types of data-centric *queries* (e.g., conjunctive queries) [11,19,20]. The representation of policies as (conjunctive) queries has been recently proposed by [25] in the context of ontologies. In the case of ontology-based systems, however, schema information plays a key role and hence policy languages should also take into account what schema information should be visible to a given user and hence typical data-oriented database queries may not suffice to specify suitable policies.

In the context of Web services, the languages WS-Policy [26] and XACML [21] have been used to specify policies. These languages provide sophisticated features that could also be useful for ontology-based systems. They are, however, not equipped with a logic-based semantics. In fact, although there have been attempts to formalise them (e.g., [6,27]), it is not clear how policies in these languages should be interpreted and evaluated w.r.t. the system's ontology. Therefore, the following questions can be an interesting starting point for future research:

- What policy languages are suitable in the context of ontology-based systems?
- How do such languages relate to those used in the context of databases and Web services?

The representation of complex policies leads to the problem of designing and maintaining them; that is, policy designers may have difficulties understanding the consequences of their policies as well as detecting errors. For example, a policy P (applied to managers) is more general than P' (applied to employees) if all the access restrictions in P also apply to employees. It would be useful to automatically check whether this is so if the system's ontology is taken into account. Therefore, an interesting research direction is to investigate reasoning problems for assisting system designers in writing high-quality policies. Preliminary results in this direction have been reported in [16].

Once the relevant policies have been designed, the next problem is to formally specify what it means for users to *violate* the policy assigned to them (i.e., to find out, by interacting with the system, information that is confidential according to the relevant policy). To this end, a first step is to formally describe the *interaction between users and the system*. It is reasonable to assume that users interact with the system by submitting queries in a given query language. Depending on the policy P assigned to each user, the system then decides whether to answer or reject the user's query Q and, in the former case, which answers to provide (for example, the true complete answer, an incomplete answer, or even an incorrect answer!).

In order for the system to make informed decisions, the *user's prior knowledge* (e.g., the answers to users' previous queries) should be considered. Formalising such prior knowledge and its provenance can be extremely difficult because information may come from many sources and/or from interactions between different users and so assumptions need to be made. In our example, Bob could query the system and learn that "John is treated by Dr. Andrews", or "Irritable bowel syndromes are gastric diseases"; also, he may access

other systems with overlapping information (e.g., the NHS website saying that "Dr. Andrews is a psychiatrist"). The formalisation of policy representation, user-system interaction and user's prior knowledge leads to the question of how to formalise the problem of policy verification, which can be informally described as follows:

> *Policy verification*: Given users prior knowledge and the corresponding policy P, does answering a given user's query Q violate the policy?

The notion of *policy violation* is open to many interpretations, and an interesting research problem is to investigate suitable semantics explaining what it is meant by violating a policy in this setting.

Once policy verification has been formalised, it remains to be seen how and when policies are verified by the system. Two scenarios are particularly worth investigating:

- *Online auditing*, where the system decides "on the fly" to answer or to reject users' queries.
- *Offline auditing*, where an auditor checks "a-posteriori" whether the answers given to a user might have compromised the policy.

In the former case, it seems reasonable to assume that users have only access to the system itself, and hence the only sources of relevant prior knowledge are the results of their previous queries. Indeed, in an online scenario it is virtually impossible to find out what other sources of information a user may have had access to or which users might have exchanged information. In the latter case, however, an auditor conducts an investigation which may reveal, for example, that Bob has had access to certain information in other systems, or has exchanged certain information with another user; the auditor then tries to determine whether Bob could be blamed for a particular privacy breach.

4. External use of a privacy-preserving system

Our second challenge was to study the situation where an external ontology-based application A wants to (re)use an ontology-based system S whose content is not available due to privacy considerations. The goal is to allow users of A to formulate queries and obtain the corresponding answers with respect to *the union of* the contents of *both* A and S, but taking into account that access limitations have been imposed to S. A central issue is how to model such access limitations, and only recently there has been research in this direction.

The authors of [7] have proposed to use data-centric *views* to formalise such access limitations. Views are represented as conjunctive queries, are given a priori, and must be compliant with the relevant policies. View extensions are computed as certain answers w.r.t. the ontology and data in S. The system makes sure that information from S not implied by the views remains hidden.

The authors of [17] have studied the situation in which the designers of S "hide" a subset of the vocabulary of S by publishing a so-called *uniform interpolant*. The interpolant can be seen as a "reusable projection" of the system's ontology and data that contains no "hidden" symbols that coincides with S on all logical consequences formed using the remaining "visible" symbols [17].

In our recent work, we have proposed an approach in which access limitations are imposed by making S accessible only via a limited query interface that we call an *oracle* [9,13]. The oracle can answer only a class of "allowed" queries over S. Under certain assumptions, a so-called *import-by-query* algorithm can reason over the union of the contents of A and S without having physical access to the content of S, by only posing queries to the oracle for S. In this situation, users may not even be aware of the existence of the privacy-preserving system.

The results in [7,9,13,17] have opened new areas of research. However, they have also left many open problems and further research is needed before they can be incorporated in practical systems.

5. Conclusion

In this short paper, I have discussed recent research on privacy-related issues in the context of ontology-based information systems.

I have identified several challenges and open problems for future research, and have sketched possible research directions. Many interesting related topics have been left out, which I believe will be (or continue to be) active areas of research within the next few years. These include, among others, privacy in the context of RDF data, issues related to trust and data provenance, and data and ontology anonymisation, among others.

References

[1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, 1995.

[2] J. Bao, G. Slutzki, and V. Honavar. Privacy-preserving reasoning on the semantic web. In *Proc. of WI-2007*, pages 791–797. IEEE Computer Society, 2007.

[3] E. Bertino and R. S. Sandhu. Database security-concepts, approaches, and challenges. *IEEE Trans. Dependable Sec. Comput.*, **2**(1):2–19, 2005.

[4] J. Biskup and P. A. Bonatti. Controlled query evaluation for enforcing confidentiality in complete information systems. *Int. J. Inf. Sec.*, **3**(1):14–27, 2004.

[5] J. Biskup and T. Weibert. Keeping secrets in incomplete databases. *Int. J. Inf. Sec.*, **7**(3):199–217, 2008.

[6] J. Bryans. Reasoning about XACML policies using CSP. In *Proc. of SWS*, 2005.

[7] D. Calvanese, G. D. Giacomo, M. Lenzerini, and R. Rosati. View-based query answering over description logic ontologies. In *Proc. of KR-2008*. AAAI Press, 2008.

[8] B. Cuenca Grau and I. Horrocks. Privacy-preserving query answering in logic-based information systems. In *Proc. of ECAI*. IOS Press, 2008.

[9] B. Cuenca Grau, B. Motik, and Y. Kazakov. Import-by-Query: Ontology Reasoning under Access Limitations. In *Proc. IJCAI*, pages 727–733. AAAI Press, 2009.

[10] A. Deutsch and Y. Papakonstantinou. Privacy in database publishing. In *Proc. of ICDT*, pages 230–245, 2005.

[11] A. V. Evfimievski, R. Fagin, and D. P. Woodruff. Epistemic privacy. In *Proc. of PODS-08*, pages 171–180. ACM, 2008.

[12] W. Fan, C. Y. Chan, and M. N. Garofalakis. Secure xml querying with security views. In *Proc. of SIGMOD*, pages 587–598, 2004.

[13] B. C. Grau and B. Motik. Pushing the Limits of Reasoning over Ontologies with Hidden Content. In *Proc. of the 12th Int. Conference on Principles of Knowledge Representation and Reasoning (KR 2010)*, Toronto, ON, Canada, May 9–13, 2010. AAAI Press. To appear.

[14] A. Y. Halevy. Answering queries using views: A survey. *VLDB J.*, **10**(4):270–294, 2001.

[15] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen. From $\mathcal{SHIQ}$ and RDF to OWL: The making of a Web ontology language. *J. of Web Semantics, Elsevier*, **1**(1):7–26, 2003.

[16] V. Kolovski, J. A. Hendler, and B. Parsia. Analyzing Web access control policies. In *Proc. of WWW*, pages 677–686, 2007.

[17] B. Konev, D. Walter, and F. Wolter. Forgetting and uniform interpolation in large-scale description logic terminologies. In *Proc. IJCAI*. AAAI Press, 2009.

[18] A. Y. Levy. Obtaining complete answers from incomplete databases. In *Proc. of VLDB*, pages 402–412, 1996.

[19] A. Machanavajjhala and J. Gehrke. On the efficiency of checking perfect privacy. In *Proc. of PODS 2006*, 2006.

[20] G. Miklau and D. Suciu. A formal analysis of information disclosure in data exchange. *J. Comput. Syst. Sci.*, **73**(3):507–534, 2007.

[21] T. Moses. Oasis Extensible Access Control Markup Language. Oasis Standard, 2005.

[22] B. Motik, I. Horrocks, and U. Sattler. Bridging the Gap Between OWL and Relational Databases. In *Proc. of WWW 2007*, pages 807–816. ACM Press, 2007.

[23] S. Rizvi, A. O. Mendelzon, S. Sudarshan, and P. Roy. Extending query rewriting techniques for fine-grained access control. In *Proc. of SIGMOD-04*, pages 551–562. ACM, 2004.

[24] P. Stouppa and T. Studer. A formal model of data privacy. In *Proc. of PSI-06*, volume 4378 of *LNCS*. Springer, 2007.

[25] P. Stouppa and T. Studer. Data privacy for knowledge bases. In *LFCS*, pages 409–421, 2009.

[26] A. Vedamuthu. Web Services Policy 1.5 – Framework. World Wide Web Consortium (W3C) Recommendation, 2007.

[27] N. Zhang, M. Ryan, and D. Guelev. Evaluating access control policies through model checking. In *Proc. of ISC*, 2005.

Semantic Web 1 (2010) 39–44
DOI 10.3233/SW-2010-0010
IOS Press

A reasonable Semantic Web

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA
Solicited review(s): Claudia d'Amato, Università degli Studi di Bari, Italy; Thomas Lukasiewicz, University of Oxford, UK
Open review(s): Aidan Hogan, DERI Galway, Ireland; Axel Polleres, DERI Galway, Ireland

Pascal Hitzler [a,*] and Frank van Harmelen [b]
[a] *Kno.e.sis Center, Wright State University, Dayton, Ohio, USA*
[b] *Vrije Universiteit Amsterdam, The Netherlands*

Abstract. The realization of Semantic Web reasoning is central to substantiating the Semantic Web vision. However, current mainstream research on this topic faces serious challenges, which forces us to question established lines of research and to rethink the underlying approaches. We argue that reasoning for the Semantic Web should be understood as "shared inference," which is not necessarily based on deductive methods. Model-theoretic semantics (and sound and complete reasoning based on it) functions as a gold standard, but applications dealing with large-scale and noisy data usually cannot afford the required runtimes. Approximate methods, including deductive ones, but also approaches based on entirely different methods like machine learning or nature-inspired computing need to be investigated, while quality assurance needs to be done in terms of precision and recall values (as in information retrieval) and not necessarily in terms of soundness and completeness of the underlying algorithms.

Keywords: Semantic Web, formal semantics, knowledge representation, automated reasoning, Linked Open Data

1. The Linked Data Web needs semantics

The Semantic Web community, in the course of its existence, has gone through an interesting swing concerning the emphasis between "data" and "knowledge."[1] Indeed, much of the talk (and research, and writing, and programming) in the early days of the Semantic Web was about ontologies as objects of study in their own right: languages to represent them, logics for reasoning with them, methods and tools to construct them, etc. Many of the research papers in the first half decade of Semantic Web research (say, 1999–2005) seemed to forget that ontologies are not made for their own sake, but that the purpose of an ontology (at least on the Semantic Web), is to help foster semantic interoperability between parties that want to exchange data. In other words, the knowledge in the ontologies (the T-box) is supposed to help interoperability of the data (the A-box).

This insight was at the birth of the Linked Open Data project [2], which put a renewed emphasis on publishing sets of actual data according to web principles. However, as it is often the case with "countermovements," it seems to us that (some of) the Linked Open Data work is erring on the other side, by only publishing just the data, and ignoring the value that can be had by annotating the data with shared ontologies.

Some of the problems that are plaguing the current Linked Open Data sets can be profitably solved by annotating data with ontologies. For example, knowing that some properties are inverse functional, knowing that certain classes are contained in each other, or that other classes are disjoint, all help to solve the instance unification problem.[2]

Similar arguments have been put forth regarding querying of Linked Open Data [19]: One of the main obstacles in querying over multiple Linked Open Data datasets is that severe information integration issues require solving. While having all data in

*Corresponding author. E-mail: pascal.hitzler@wright.edu.
[1] or, in Description Logic speak: between "A-box" and "T-box"

[2] The instance unification problem refers to the problem of determining when two differently named instances are in fact identical.

```
bills/h3962       dc:title          "H.R. 3962: ..." ;
                  usbill:hasAction  _:bnode0 .
_:bnode0          usbill:vote       votes/2009-887 .
votes/2009-887    vote:hasOption    votes/2009-887/+ .
                  dc:title      "On Passage: H.R. 3962 ..." ;
votes/2009-887/+ rdfs:label        "Aye" ;
                  vote:votedBy      people/P000197 .
people/P000197    usgovt:name       "Nancy Pelosi" .
```

Fig. 1. GovTrack triples encoding the knowledge that Nancy Pelosi voted in favor of the Health Care Bill. URIs have been abbreviated freely since the details do not matter for our discussion.

RDF syntax (Resource Description Framework [23]) solves the information integration issue on a syntactic level, the current state of querying over the Linked Open Data cloud exposes the fact that semantic integration is hardly present. Indeed, RDF language primitives which are actually reflected by the RDF formal semantics (such as `rdfs:subClassOf` or `rdfs:domain`) are relatively scarce in the cloud.[3] The only strong semantic language primitive used heavily is `owl:sameAs` from the Web Ontology Language OWL [15], and it has been observed frequently that its use is often rather abuse [6,13].

Another issue which points at a lack of semantics is the sometimes rather convoluted way of expressing knowledge in the Linked Open Data cloud. As just one example, let it be noted that the simple fact *Nancy Pelosi voted in favor of the Health Care Bill* is encoded in GovTrack[4] using eight RDF triples, two of which share a blank node (see Fig. 1). From this and other examples, it seems apparent that triplification for the Linked Open Data cloud is often done without deep contemplation of semantic issues,[5] or of usefulness of the resulting data.[6]

2. Semantics as shared inference

Semantic interoperability is usually defined in terms of a formal semantics. But what does it mean for two agents to agree on the formal semantics of a message? Although the primary definition of the semantics of formal languages is most often in terms of a denotational semantics, e.g. [14] and [24] for RDF and OWL, respectively, perhaps a more productive definition on the Semantic Web is to describe semantic interoperability in terms of shared inferences.

When an agent (a web server, a web service, a database, a human in a dialogue) utters a message, the message will often contain more meaning than only the tokens that are explicitly present in the message itself. Instead, when uttering the message, the agent has in mind a number of "unspoken," implicit consequences of that message. When a web page contains the message "Amsterdam is the capital of The Netherlands," then some of the unspoken, implicit consequences of this are that Amsterdam is apparently a city (since capitals are cities), that The Hague is not the capital of the Netherlands (since every country only has precisely one capital), that The Netherlands is a country, or a province, but not another city, since countries and provinces have capitals, but cities do not; a spatial implied fact is that the location of the capital city is inside the area covered by the country, etc.

If agent A utters the statement about Amsterdam to agent B, they can only be said to be truly semantically interoperating if B not only knows the literal content of the phrase uttered by A, but also understands a multitude of implicit consequences of that statement which are then shared by A and B. It is exactly these shared, implicit consequences which are made explicit in the form of a shared ontology.

We could say that the amount of semantic interoperability between A and B is measured by the number of new facts that they both subscribe to after having exchanged a given sentence: the larger and richer their shared inferences, the more semantically interoperable they are.[7]

A language such as RDF Schema [23] which contains (almost) no negation, allows agent A to enforce beliefs on the receiving agent B, e.g. by specifying the domain and range of a property like "is capital of." This puts a *lower bound* on the inferences to be made by agent B, i.e., it "enforces" inferences to be made by B when it subscribes to the shared semantics. A richer language such as OWL [15] also allows agent A to "forbid" agent B to make certain inferences. Stating that Amsterdam is the capital of The Netherlands, that "is capital of" is an inverse functional property, and that Amsterdam is different from The Hague will disallow the inference that The Hague is the capital of The Netherlands. This puts an *upper bound* on the inferences to be made by agent B. By making an ever richer ontology, we can move the upper and lower bounds of the shared inferences ever closer,

[3] "Scarcity," in this case, is a rather subjective matter. Let's just say that it currently seems too scarce to be really useful for reasoning.

[4] http://www.govtrack.us/

[5] See also [1,17,28] for further discussions.

[6] For an amusing critique on this practice, see [35].

[7] Ontology alignment issues obviously occur here, too.

hence obtaining ever finer-grained semantic interoperability through an ever more precisely defined set of shared inferences.

Of course, this perspective of semantics as "shared inference" is entirely compatible with the classical view of semantics as model theory, in the sense of the formal semantics of, e.g., RDF and OWL: Valid inferences are inferences which hold in all models, and invalid inferences are inferences that hold in no model. However, semantics as "shared inference" does not presuppose the use of model theory,[8] although the latter currently seems to be the most advanced method for capturing this kind of semantics. Essential to the "shared inference" perspective is that it facilitates communication (and, thereby, interoperability), while model theory is often construed[9] as "the defining of meaning in a unique way."

3. Semantics as a gold standard

The usual role of semantics is to define precisely how the meaning of a set of sentences in a logic is defined. In Section 2, we have already seen that it is also possible to think of semantics in terms of an ever narrowing gap of multi-interpretability (with an ever increasing set of axioms closing the gap between what must be derived (inferential lower bound) and what may not be derived (inferential upper bound) from a set of sentences.

The classical view on semantics is then that any properly defined system must precisely obey this semantics: it must be sound and complete, i.e., any consequence prescribed by the semantics must also be derived by the system, and vice versa. Only recently the Semantic Web community has begun to appreciate the value of incomplete systems [11]. It is often useful to build systems that do not manage to derive all required consequences, as long as they derive a useful subset of these.

Rather than regarding this as an unfortunate but perhaps inevitable sloppiness of such implementations with respect to their semantic specification, we would advocate a different perspective, namely to view the formal semantics of a system (in whatever form it is specified) as a "gold standard," that need not necessarily be obtained in a system (or even be obtainable). What is required from systems is not a proof *that* they satisfy this gold standard, but rather a precise description of the *extent to which* they satisfy this gold standard [29].

Notice that in other, related, fields this is already commonplace: in Information Retrieval, the measures of precision and recall correspond exactly to soundness and completeness, but with the crucial difference that nobody only expects systems where both of these values are at 100%. Instead, systems are routinely measured on the extent to which they approximate full precision (soundness) and recall (completeness), and both researchers and application builders have learned to live with imperfect systems, and with laws that tell us that increasing one of the measures typically decreases the other. In short, the logical model has perhaps confused the ideal with the realistic, and the theory and practice of information retrieval may well be more appropriate for Semantic Web reasoners.[10]

A wide misconception is that, even when incompleteness may be a worthy strategy, surely unsoundness is bad in all cases. Again, the perspective from Information Retrieval shows that this is simply false: depending on the use-case, one may have a preference for erring either on the side of incompleteness (e.g. finding just a few but not all matching products is fine as long as all answers do match the stated requirements) or on the side of unsoundness (e.g. finding all potential terrorist suspects, even when this possibly includes a few innocent people). Just as in Information Retrieval, a use-case specific balance will have to be struck between the two ends of the spectrum, with neither being always better than the other.

From this perspective (semantics as a, possibly unobtainable, gold-standard) systems with anytime behaviour also become a very natural object of study: they just happen to be systems that succeed in increasingly better approximations of the gold standard as time progresses. It turns out that many algorithms for deduction, query answering, subsumption checking, etc., have a natural anytime behaviour that can be fruitfully exploited from the perspective of "semantics

[8] We do not want to propose any particular approach at this stage, but let it be noted that even the notion of *formal semantics* does not necessarily rely on model theory. Semantics based on order theory or on metric spaces, as used in denotational semantics of programming languages, are just one example, and can be ported to the knowledge representation realm [16].

[9] it might be more accurate to say: misconstrued

[10] See [3] for some alternatives to precision and recall in a Semantic Web context. We restrict our discussion to precision and recall simply because they are well established. We do not claim that there are no good or better alternatives: future research will have to determine this.

as a gold standard" that need not be perfectly achieved before a system is useful.

4. Semantics as possibly non-classical

If we take the viewpoints that "semantics is a (possibly unobtainable) gold standard for shared inference," then we can also change our view on what form this semantics must take. Why would a shared set of inferences have to consist of conclusions that are held to be either completely true or completely false? Wouldn't it be reasonable to enforce a minimum (or maximum) degree of believe in certain statements? Or a degree of certainty? Or a degree of trust? This would amount to agent A and agent B establishing their semantic interoperability not by guaranteeing that B holds for eternally true all the consequences that follow from the statements communicated by A, but rather by guaranteering that B shares a degree of trust in all the sentences that are derivable from the sentences communicated by A.

A similar argument can be made for the handling of inconsistency. Shouldn't a semantics for "shared inference" be able to sort out inconsistencies and different perspectives on the fly? We know that classical model theory cannot deal with these issues. And what about default assumptions and the occurrence of exceptions to them? Classically, these lead to inconsistency, but in "shared inference" it should be dynamically resolvable.

While these perspectives, again, appear to be compatible with well-known knowledge representation approaches using, e.g., fuzzy or probabilistic logics [21, 31], paraconsistent reasoning [22], non-monotonic [7, 12,20,25], or mixed approaches [30], it is an open question whether they carry far enough for realistic use cases. While apparently promising as conceptual ideas, these logics have not yet been shown to be applicable in practice other than in simplified settings. How they could work on the open Semantic Web remains, to this date, unclear.

To us, it appears to be a reasonable perspective, that these issues need to be resolved, practically, in a different manner, as described below. Formal semantics, using non-classical logics, can probably still serve as a gold standard for evaluating inference system performances, but realistic data and applications will most likely force us to deviate from classical automated reasoning grounds for computing shared inferences.

5. Computing shared inferences

To summarize the train of thought we have laid out so far, we see that, in order to realize the interoperability required by the Semantic Web, we

- require shared ontologies which carry a formal semantics,
- formal semantics acts as a gold standard but does not need to be computed in a sound and complete way, and
- systems should be able to deal with noise, different perspectives, and uncertainty.

Traditionally, systems for computing inferences are based on logical proof theory and realize sound and complete algorithms on the assumption that input data is monolithic, noise-free, and conveys a single perspective on a situation or domain of applications. While this approach is certainly valid as such, it faces several severe challenges if ported to the Semantic Web. Two of the main obstacles are scalability of the algorithms, and requirements on the input data.

Concerning scalability, reasoning systems have made major leaps in the recent past [33,34]. However, it remains an open question when (and if[11]) these approaches will scale to the size of the web, and this problem is aggravated by the incorporation of non-classical semantics as discussed in Section 4, which inherently brings a rapid decrease in efficiency.

Concerning requirements on the input data, it is quite unrealistic to expect that data from the open Semantic Web will ever be clean enough such that classical reasoning systems will be able to draw useful inferences from them. This would require Semantic Web data to be engineered strongly according to shared principles, which not only contrasts with the bottom-up nature of the Web, but is also unrealistic in terms of conceptual realizability: many statements are not true or false, they rather depend on the perspective taken.

If we come to the conclusion that inference systems based on logical proof theory likely will not work on web-scale realistic Semantic Web data,[12] the discussion from Section 3 becomes of central importance: Formal semantics is required as a gold standard for evaluation of systems computing shared inferences, however, it is okay for such systems to deviate from

[11] Since the web keeps growing, they may never scale, even if they become much more efficient.

[12] This does, obviously, not preclude them from being very useful for smaller and/or more controled domains.

the gold standard, in a manner which can be qualitatively assessed in terms of precision and recall, if they scale better and/or are able to deal with realistic, noisy, data.

6. What is needed?

We have argued for the need of methods for computing shared inferences, which are not foremost based on the idea of producing sound and complete systems. We believe that there is a need for a concerted effort in the Semantic Web community to address this issue, both in terms of producing such systems, and in terms of pursuing use cases involving shared inference which employ reasoning methods which can scale up to web size.

Potential methods for establishing such inference systems can be found in other realms, where the need for approximate solutions is an accepted fact. Approximate algorithms, e.g., are commonly employed for NP-hard problems.[13] Approximate reasoning, understood in the same sense, has an established tradition. The development of according ideas for Semantic Web reasoning is indeed being pursued to a certain extent [18,26,27,32], and would benefit from a critical mass of further research.

Alternative approaches may employ methods which do not involve proof-theoretic aspects at all. From a bird's eye perspective, reasoning can be understood as a classification problem: classify a query as "true" or as "false." Machine learning, nature-inspired computing, or any method used in data mining or information retrieval are candidates for exploring new Semantic Web reasoning paradigms (see, e.g., [4,5,8–10]). These methods often have the pleasing property to be robust with respect to noise or contradictory input, and so there is reason to believe that they may simply render the difficulties identified in Section 4 to be void.

Let us close by emphasizing again that taking such approaches does not mean that we give up on formal semantics. It still serves as a gold standard for evaluation. It just means that we acknowledge that we need to rethink the role of semantics and the role of computation of semantics, provided we hope to make significant advances in the Semantic Web quest.

[13] Considering the fact that OWL reasoning is harder than NP, it is unfathomable why there should be any resistance against using approximate methods for OWL reasoning.

Acknowledgements

We thank Prateek Jain for digging out the example in Fig. 1. Pascal Hitzler acknowledges support by the Write State University Research Council.

References

[1] S. Auer, J. Lehmann. Making the Web a data washing machine – Creating knowledge out of interlinked data. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):97–104, 2010.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data – the story so far. *International Journal on Semantic Web and Information Systems*, **5**(3):1–22, 2009.

[3] C. d'Amato, N. Fanizzi, and F. Esposito. Query answering and ontology population: An inductive approach. In S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, editors, *The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1–5, 2008, Proceedings*, volume 5021 of *Lecture Notes in Computer Science*, pages 288–302. Springer, 2008.

[4] C. d'Amato, N. Fanizzi, and F. Esposito. Inductive learning for the Semantic Web: What does it buy? *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):53–59, 2010.

[5] C. d'Amato, N. Fanizzi, B. Fazzinga, G. Gottlob, and T. Lukasiewicz. Combining Semantic Web search with the power of inductive reasoning. In *Proceedings SUM 2010*, Lecture Notes in Computer Science. Springer, 2010. To appear.

[6] L. Ding, J. Shinavier, T. Finin, and D. L. McGuinness. An empirical study of owl:sameAs use in Linked Data. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line, April 2010, Raleigh, NC*, 2010. To appear.

[7] T. Eiter, G. Ianni, T. Lukasiewicz, R. Schindlauer, and H. Tompits. Combining Answer Set Programming with Description Logics for the Semantic Web. *Artificial Intelligence*, **172**(12–13):1495–1539, August 2008.

[8] N. Fanizzi, C. d'Amato, and F. Esposito. Statistical learning for inductive query answering on OWL ontologies. In A. P. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. W. Finin, and K. Thirunarayan, editors, *The Semantic Web – ISWC 2008, 7th International Semantic Web Conference, ISWC 2008, Karlsruhe, Germany, October 26–30, 2008. Proceedings*, volume 5318 of *Lecture Notes in Computer Science*, pages 195–212. Springer, 2008.

[9] B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz. Semantic Web search based on ontological conjunctive queries. In S. Link and H. Prade, editors, *Foundations of Information and Knowledge Systems, 6th International Symposium, FoIKS 2010, Sofia, Bulgaria, February 15–19, 2010. Proceedings*, volume 5956 of *Lecture Notes in Computer Science*, pages 153–172. Springer, 2010.

[10] B. Fazzinga and T. Lukasiewicz. Semantic search on the Web. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):89–96, 2010.

[11] D. Fensel and F. van Harmelen. Unifying reasoning and search to web scale. *IEEE Internet Computing*, **11**(2):96, 94–95, 2007.

[12] S. Grimm and P. Hitzler. Semantic matchmaking of web resources with local closed-world reasoning. *International Journal of e-Commerce*, **12**(2):89–126, 2008.

[13] H. Halpin and P. J. Hayes. When owl:sameAs isn't the same: An analysis of identity links on the Semantic Web. In *Proceedings of the WWW2010 workshop on Linked Data on the Web, LDOW2010*, 2010. To appear.

[14] P. Hayes, editor. *RDF Semantics*. W3C Recommendation, 10 February 2004. Available from http://www.w3.org/TR/rdf-mt/.

[15] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, editors. *OWL 2 Web Ontology Language: Primer*. W3C Recommendation 27 October 2009, 2009. Available from http://www.w3.org/TR/owl2-primer/.

[16] P. Hitzler and A. K. Seda. *Mathematics Aspects of Logic Programming Semantics*. CRC Press, 2011. To appear.

[17] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In *3rd International Workshop on Linked Data on the Web (LDOW2010) at WWW2010, Raleigh, USA, April 2010*, 2010. Available from http://events.linkeddata.org/ldow2010/.

[18] A. Hogan, A. Harth, and A. Polleres. Scalable authoritative OWL reasoning for the web. *Int. J. Semantic Web Inf. Syst.*, **5**(2):49–90, 2009.

[19] P. Jain, P. Hitzler, P. Z. Yeh, K. Verma, and A. P. Sheth. Linked Data is Merely More Data. In D. Brickley, V. K. Chaudhri, H. Halpin, and D. McGuinness, editors, *Linked Data Meets Artificial Intelligence*, pages 82–86. AAAI Press, Menlo Park, CA, 2010.

[20] M. Knorr, J. J. Alferes, and P. Hitzler. A coherent well-founded model for Hybrid MKNF knowledge bases. In M. Ghallab, C. D. Spyropoulos, N. Fakotakis, and N. Avouris, editors, *Proceedings of the 18th European Conference on Artificial Intelligence (ECAI-08)*, pages 99–103. IOS Press, 2008.

[21] T. Lukasiewicz and U. Straccia. Managing uncertainty and vagueness in description logics for the Semantic Web. *Journal of Web Semantics*, **6**(4):291–308, 2008.

[22] Y. Ma and P. Hitzler. Paraconsistent reasoning for OWL 2. In A. Polleres and T. Swift, editors, *Web Reasoning and Rule Systems, Third International Conference, RR 2009, Chantilly, VA, USA, October 25–26, 2009, Proceedings*, volume 5837 of *Lecture Notes in Computer Science*, pages 197–211. Springer, 2009.

[23] F. Manola and E. Miller, editors. *Resource Description Framework (RDF). Primer*. W3C Recommendation, 10 February 2004. Available from http://www.w3.org/TR/rdf-primer/.

[24] B. Motik, P. F. Patel-Schneider, and B. C. Grau, editors. *OWL 2 Web Ontology Language: Direct Semantics*. W3C Recommendation, 27 October 2009. Available from http://www.w3.org/TR/owl2-direct-semantics/.

[25] B. Motik and R. Rosati. Reconciling description logics and rules. *Journal of the ACM*, **57**(5), 2010.

[26] E. Oren, S. Kotoulas, G. Anadiotis, R. Siebes, A. ten Teije, and F. van Harmelen. Marvin: Distributed reasoning over large-scale Semantic Web data. *Web Semantics: Science, Services and Agents on the World Wide Web*, **7**(4):305–316, 2009.

[27] J. Z. Pan and E. Thomas. Approximating OWL-DL ontologies. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22–26, 2007, Vancouver, British Columbia, Canada*, pages 1434–1439. AAAI Press, 2007.

[28] A. Polleres, A. Hogan, A. Harth, and S. Decker. Can we ever catch up with the Web? *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):45–52, 2010.

[29] S. Rudolph, T. Tserendorj, and P. Hitzler. What is approximate reasoning? In D. Calvanese and G. Lausen, editors, *Web Reasoning and Rule Systems, Second International Conference, RR 2008, Karlsruhe, Germany, October 31–November 1, 2008. Proceedings*, volume 5341 of *Lecture Notes in Computer Science*, pages 150–164, 2008.

[30] T. Scharrenbach, R. Grütter, B. Waldvogel, and A. Bernstein. Structure preserving TBox repair using defaults. In V. Haarslev, D. Toman, and G. Weddell, editors, *Proceedings of the 2010 Description Logic Workshop (DL 2010)*, volume 573 of *CEUR Workshop Proceedings*, 2010.

[31] G. Stoilos, G. B. Stamou, J. Z. Pan, V. Tzouvaras, and I. Horrocks. Reasoning with very expressive fuzzy description logics. *Journal of Artificial Intelligence Research*, **30**:273–320, 2007.

[32] T. Tserendorj, S. Rudolph, M. Krötzsch, and P. Hitzler. Approximate OWL-reasoning with Screech. In D. Calvanese and G. Lausen, editors, *Web Reasoning and Rule Systems, Second International Conference, RR 2008, Karlsruhe, Germany, October 31–November 1, 2008. Proceedings*, volume 5341 of *Lecture Notes in Computer Science*, pages 165–180, 2008.

[33] J. Urbani, S. Kotoulas, J. Maassen, F. van Harmelen, and H. Bal. OWL reasoning with WebPIE: calculating the closure of 100 billion triples. In L. Aroyo, G. Antoniou, E. Hyvönen, A. ten Teije, H. Stuckenschmidt, L. Cabral, and T. Tudorache, editors, *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30–June 3, 2010, Proceedings, Part I*, volume 6088 of *Lecture Notes in Computer Science*, pages 213–227. Springer, 2010.

[34] J. Urbani, S. Kotoulas, E. Oren, and F. van Harmelen. Scalable Distributed Reasoning Using MapReduce. In A. Bernstein et al., editors, *Proceedings of the 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25–29, 2009*, volume 5823 of *Lecture Notes in Computer Science*, pages 634–649. Springer, 2009.

[35] D. Vrandečić, M. Krötzsch, S. Rudolph, and U. Lösch. Leveraging non-lexical knowledge for the Linked Open Data web. In R. Héliot and A. Zimermann, editors, *5th Review of April Fool's day Transactions*, pages 18–27, 2010. Available from http://vmgal34.deri.ie/~antzim/RAFT/afd2010.html.

Semantic Web 1 (2010) 127–130
DOI 10.3233/SW-2010-0011
IOS Press

Smart objects: Challenges for Semantic Web research

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Lora Aroyo, Free University of Amsterdam, The Netherlands; Boyan Brodaric, Geological Survey of Canada, Canada

Marta Sabou *
MODUL University Vienna, Austria
E-mail: marta.sabou@modul.ac.at

Abstract. The increased availability and robustness of sensors, the wide-spread use of the internet as a communication environment and the intensified adoption of semantic technologies foster the vision of embedding intelligence in physical objects. The race of realizing this vision is pervasive to a variety of research fields, most notably ambient intelligence and semantic web, and leads to the proliferation of several overlapping definitions and terminologies: smart products, semantic devices, semantic gadgets - to which we collectively refer to as *smart objects*. What exactly are smart objects? And what are the research challenges in realizing them? We hereby explore the answers to these questions.

Keywords: Smart objects, Semantic Web, ambient intelligence, challenges

1. Introduction

While the Semantic Web (SW) started out as an initiative for enhancing a Web of primarily textual documents, the technologies developed by this community have evolved and have been applied to major Web innovations such as the Web of services or the Social Web. With the advent of sensors, computationally enhanced physical devices, ubiquitous connectivity of objects (e.g., Internet of Things), the SW community naturally follows suit and an increased interest is now shown in extending the use of semantic technologies beyond the digital world into the realm of physical things and devices.

This novel orientation of the SW field complements longstanding efforts in AI to embed intelligence in the surrounding environments and physical objects in the context of research areas such as robotics and, more recently, ambient intelligence [14,15,21]. Obviously, the intention is not to compete against that large body of work but rather to complement it towards the realisation of a vision that, over the last two decades [21], has become a "melting pot" for various scientific disciplines. In particular, we note those differentiating characteristics of SW techniques which make them well suited in scenarios which involve a high number of heterogeneous devices: (i) they have been designed to work at Web scale; (ii) they foster interoperability between heterogeneous data sources; (iii) they rely on a stack of Web technology standards which allow for easy and large scale adoption.

While the quest for using SW technologies in pervasive computing application is currently intensifying, we can actually trace it back to almost a decade ago, in the area of Task Computing [11] where users can easily compose services based on semantic descriptions of devices. Then, ontology-based smart environments and devices were investigated by initiatives such as SOCAM [6] and CoBra [3]. However, an overall characteristic of these approaches was their centralised nature, where the intelligence of the individual devices depended on the processes handled by a central computer. The recent proliferation of intelligent devices,

*Part of the work reported in this paper has been performed by the author in the context of the FP7 SmartProducts project (231204) while working at the Knowledge Media Institute(KMi), The Open University, UK.

advances in sensor and communication technologies, all support a trend towards making devices more autonomous by embedding intelligence into them [19]. For example, the SoaM architecture relies on SW technologies to realise semantic gadgets [20]. The architecture allows both distributed and centralised topologies thus providing a smooth transition from centralised solutions towards autonomous smart objects. Indeed, the authors' experiments show that distributed topologies often rival centralised ones in terms of performance, thus providing an early proof for the feasibility of the distributed approaches.

In the next section we describe smart objects by summarising various, complementary definitions from fields as diverse as business studies, ambient intelligence and Semantic Web. Based on these definitions and on our experiences within the SmartProducts project[1], we discuss a set of research challenges in realising smart objects as well as current efforts towards solving those challenges. Our analysis complements a similar study in the area of the semantic sensor web published in this issue [4]: although smart objects rely on sensors and sensor networks, they provide a more focused application domain with its own challenges.

2. Defining smart objects

The notion of objects (products, devices, gadgets) that display some level of intelligence has been proposed in various research fields. Allmendinger and Lombreglia investigate the notion of smartness in a product from a business perspective [2]. They regard "smartness" as the product's capability to be *preemptive*, i.e., to be able to predict errors and faults thus "removing unpleasant surprises from [the users'] lives".

A recent notion introduced in the area of ambient intelligence is that of *smart products*. In 2008, Maas et al. [10] define smart products as *adaptive* to situations and users. This adaptivity is enabled by three main technologies: (i) sensing technologies which identify the global and the local context of a product (using global or local sensors respectively); (ii) communication infrastructures and (iii) IT services, in particular, "rich context representations, representations about product capabilities and domain knowledge" used "to infer how to learn from and adapt to users and situations". For Mühlhäuser [12], smart products are objects, software or services that have improved *simplicity* (in terms of user interaction) and *openness* (in terms of connecting to and communicating with other devices). These characteristics are achieved through *"context-awareness, semantic self-description, proactive behaviour, multimodal natural interfaces"* [12].

In the Semantic Web area, Lassila and Adler proposed the notion of *semantic gadget*, a device capable of performing *"discovery and utilisation of services without human guidance or intervention, thus enabling formation of device coalitions"* [8]. Vaquez et al. [19] extend this definition to that of a *semantic device*, a system that is *"spontaneously aware of surrounding context information, capable of reasoning and interpreting this information at a semantic level, and finally able to develop a reactive behaviour accordingly"*. Additionally, a semantic device is able to *"spontaneously discover, exchange and share context information with other fellow semantic devices"*. Some prototype semantic devices include SmartPlants (house plants paired with an intelligent artefact which sense lighting and temperature conditions and ask to be moved to the most suitable position) or the AwareUmbrella (umbrella which obtains weather information from local sensors and the Internet and alerts the owner to take it along when it is likely to rain).

The SmartProducts project combines research from the ambient intelligence and SW fields to provide an industry-applicable, lifecycle-spanning methodology with tools and platforms to support the construction of smart products. While using Mühlhäuser's definition [12] as a starting point, the project focuses on tangible objects (i.e., physical products) as smart products and not virtual products like software or services. Proactivity is a core characteristic of these products and is ensured by them being *"self-, situational-, and context-aware"*. Finally, the knowledge and functionalities of smart products can be shared with other products and evolve over time as a side effect of their interactions with users and other products.

While originating from diverse fields, the above definitions converge towards a set of core characteristics that a smart object (product, gadget, device):

- *context-awareness* – the ability to sense context;
- *proactivity* – the ability to reason upon and make use of this context and other information in order to proactively approach users and peers;
- *self-organization* – the ability to form and join networks with other products.

[1] http://www.smartproducts-project.eu/

In addition to these characteristics, smart products should support their entire life-cycle and should offer multimodal interaction with the users, in order to increase product simplicity [12]. Maas and colleagues highlight the need for using context information in order to support personalisation and adaptiveness [10]. They also see products as being aware of concrete business and legal constraints. The SmartProducts consortium identified some additional characteristics to those provided by Maas and colleagues in [10]. Most importantly, products are seen as capable of acting autonomously (by themselves) without the need of central control. The rest of the characteristics refer to aspects of the knowledge component that enables the smartness of the products. This knowledge has an important procedural component, it should evolve during the life-cycle of the product as a side effect of its interaction with users and products and, finally, it might need to be stored in a distributed fashion in order to overcome the resource limitations of some objects.

3. Challenges for semantic technologies

Knowledge technologies play a crucial role in embedding intelligence into physical objects, in particular, for semantically representing context information and providing reasoning mechanisms that underpin proactivity and product-to-product interactions. We hereby discuss some of the challenges that such technologies are likely to face:

Hardware resource limitations. In the process of moving from intelligent, centralised architectures towards autonomous objects with on-board intelligence, the hardware limitations of these objects present an important challenge. Although physical objects are heterogeneous in terms of their hardware resources for information storage and processing, even the most powerful objects will lag behind the resources characteristic of the computer machinery for which semantic technologies are currently built.

An important objective for the Semantic Web community is to adapt its technologies for use on objects with limited computational resources. Strategies in this area include reducing the storage space needed for semantic data [13] and optimising semantic tools in terms of resource consumption. For example, Ali and Kiefer [1] describe the μOR query answering and reasoning system for resource-constrained (mobile) devices which improves on the performance of two earlier reasoners specifically built for mobile devices, i.e., Bossam and Pocket KRHyper [7]. Alternatively, W. Tai et. al propose an automatically composable OWL reasoner which is customised automatically depending on the semantics of the ontology to be reasoned upon by selecting the required reasoning modules only [17]. They show that the approach reduces memory requirements while maintaining reasoning ability thus being well suited for resource constrained devices.

Complex reasoning algorithms. Smart objects use reasoning mechanisms on their rich knowledge bases in order to adapt to user needs, to perform personalisation and to proactively interact with users and other objects. This complex expected behaviour requires sophisticated reasoning mechanisms such as diagnosis or planning. Such reasoning is much more ambitious than current work in the area of sensor networks which primarily relies on subsumption matching (e.g., for matching between available resources and tasks [5]).

We expect that, given the proactive nature of smart objects, they will mostly rely on production rule-engines rather than DL reasoners. As a response to the increased interest in rule engines, the OpenRuleBench[2] benchmark has been established for analysing the performance and scalability of different rule engines and already used for comparing 11 systems [9]. While a good step towards understanding the capabilities of various rule-engines, this benchmark is not suited towards evaluating rule-engine performance on resource-constrained devices.

Tokmakoff et al. [18] argue that, in order to deal with ambiguities and uncertainties inherent in environments involving human beings, the reasoning mechanisms of smart products should not rely on two-valued logics but rather combine fuzzy, rough or probabilistic deduction methods. However, combing these methods is not trivial and still requires extensive research.

Suboptimal data quality. A fundamental characteristic of smart objects is that they rely on context information obtained from associated sensors which is then translated into higher level semantic information. However, as pointed out by Corcho and Garcia-Castro, ensuring sensor data quality is an important challenge and has to account for issues such as data unavailability or lack of accuracy [4]. Although they also describe a set of research efforts towards improving sensor data quality, it is reasonable to assume that sensor

[2] http://rulebench.projects.semwebcentral.org/

data will have a lower quality than manually authored and checked semantic information. For example, the derived data could be incomplete or, on the contrary, contain redundant elements. Therefore it is important a) to further develop fusion techniques that combine data from multiple sensors into meaningful semantic data and b) to build semantic techniques that are *robust* enough to be able to process such data.

Representing a variety of information. Researchers investigating semantic sensor webs generally agree that semantic models are needed for representing information about time, space and the domain relevant for the sensors [16]. From our analysis of smart objects and their characteristics, we can conclude that their representation needs are much richer and more diverse. Indeed, at a minimum, knowledge associated with smart objects should contain user models, task models (procedural knowledge), models to represent life-cycle stages and the main users (or communities of practice) involved in each stage, interaction models. Therefore, the employed semantic technologies should be able to cover all these representation needs.

Earlier studies from using semantic techniques in pervasive computing applications suggest new representational requirements for ontologies. For example, in [11], the authors report on using ontologies to enable task computing, i.e., easy composition of services provided by various devices in a room. The authors acknowledge that ontologies were not so much used for formal reasoning, but rather for making service composition easier for users. As such, ontology comments and labels played an important role.

Further challenges. It is envisioned that smart objects will continuously update their knowledge bases by deriving knowledge as a side effect of their interaction with users and other objects. Therefore, mechanisms for supporting the derivation and evolution of *emergent knowledge* need to be built. Further, given their close interaction with users, smart objects need to maintain a considerable amount of information about users including their likes, dislikes, their usage patterns, their personal information etc. It is therefore crucial to implement access rights mechanisms that can ensure the desired level of *trust and privacy* for user data distributed across multiple objects.

References

[1] S. Ali and S. Kiefer. μOR — A Micro OWL DL Reasoner for Ambient Intelligent Devices. In *Proc. of the Int. Conf. on Advances in Grid and Pervasive Computing*, 2009.

[2] G. Allmendinger and R. Lombreglia. Four Strategies for the Age of Smart Services. *Harvard Business Review*, **83**(10):131–145, 2005.

[3] H. Chen, T. Finin, and A. Joshi. Semantic Web in a Pervasive Context-Aware Architecture. In *A.I. in Mobile Systems*, 2003.

[4] O. Corcho and R. Garcia-Castro. Five Challenges for the Semantic Sensor Web. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):121–125, 2010.

[5] G. de Mel, M. Sensoy, W. Vasconcelos, and A. Preece. Flexible resource assignment in sensor networks: A hybrid reasoning approach. In *Proc. of SemSensWeb Ws.*, 2009.

[6] T. Gu, H.K. Pung, and D.Q. Zhang. A service-oriented middleware for building context-aware services. *J. Netw. Comput. Appl.*, **28**(1):1–18, 2005.

[7] T. Kleemann and A. Sinner. KRHyper – in Your Pocket, System Description. In *Proc. of Automated Deduction*, 2005.

[8] O. Lassila and M. Adler. *Spinning the Semantic Web*, chapter Semantic Gadgets: Ubiquitous Computing Meets the Semantic Web, pages 363–376. MIT Press, 2003.

[9] S. Liang, P. Fodor, H. Wan, and M. Kifer. OpenRuleBench: an Analysis of the Performance of Rule Engines. In *Proc. of the Int. Conf. on WWW*, pages 601–610, 2009.

[10] W. Maass and U. Varshney. Preface to the Focus Theme Section: 'Smart Products'. *Electronic Markets*, **18**(3):211–215, 2008.

[11] R. Masuoka, Y. Labrou, B. Parsia, and E. Sirin. Ontology-enabled pervasive computing applications. *Intelligent Systems, IEEE*, **18**(5):68–72, Sep/Oct 2003.

[12] M. Mühlhäuser. Smart Products: An Introduction. In *Constructing Ambient Intelligence Workshop*, 2008.

[13] D. Preuveneers and Y. Berbers. Encoding Semantic Awareness in Resource-Constrained Devices. *IEEE Intelligent Systems*, **23**(2):26–33, 2008.

[14] C. Ramos, J.C. Augusto, and D. Shapiro. Ambient intelligence—the next step for artificial intelligence. *IEEE Intelligent Systems*, **23**(2):15–18, 2008.

[15] N. Shadbolt. Ambient intelligence. *IEEE Intelligent Systems*, **18**(4):2–3, 2003.

[16] A. Sheth, C. Henson, and S. Sahoo. Semantic Sensor Web. *IEEE Internet Computing*, **12**(4):78–83, 2008.

[17] W. Tai, R. Brennan, J. Keeney, and D. O'Sullivan. An Automatically Composable OWL Reasoner for Resource Constrained Devices. In *Proc. of Semantic Computing*, 2009.

[18] A. Tokmakoff, X. Zhou, L. Holenderski, S. van Loo, and A. Sinitsyn. From lab to living-room: research challenges for AmI and Smart Products. In *Proc. of Smart Products: Building Blocks of Ambient Intelligence WS.*, 2009.

[19] J. I. Vazquez and D. Lopez de Ipina. Principles and experiences on creating semantic devices. In *Proc. of the Int. Symp. on Ubiquitous Computing and Ambient Intelligence*, 2007.

[20] J.I. Vazquez, D. Lopez de Ipina, and I. Sedano. SoaM: A Web-powered Architecture for Designing and Deploying Pervasive Semantic Devices. *Int. J. of Web Inf. Systems*, **2**(3/4):212–224, 2006.

[21] M. Weiser. The Computer for the Twenty-First Century. *Scientific American*, **265**(3):94–104, 1991.

Semantic Web 1 (2010) 11–15
DOI 10.3233/SW-2010-0012
IOS Press

Modeling vs encoding for the Semantic Web

Editors: Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited reviews: Thomas Lukasiewicz, Oxford University, UK; Giancarlo Guizzardi, Federal University of Espírito Santo, Brazil
Open review: Pascal Hitzler, Wright State University, USA

Werner Kuhn
Institute for Geoinformatics (ifgi), University of Münster, Weselerstr. 253, D-48151 Münster, Germany
E-mail: kuhn@uni-muenster.de

Abstract. The Semantic Web emphasizes encoding over modeling. It is built on the premise that ontology engineers can say something useful about the semantics of vocabularies by expressing themselves in an encoding language for automated reasoning. This assumption has never been systematically tested and the shortage of documented successful applications of Semantic Web ontologies suggests it is wrong. Rather than blaming OWL and its expressiveness (in whatever flavor) for this state of affairs, we should improve the modeling techniques with which OWL code is produced. I propose, therefore, to separate the concern of modeling from that of encoding, as it is customary for database or user interface design. Modeling semantics is a design task, encoding it is an implementation. Ontology research, for applications in the Semantic Web or elsewhere, should produce languages for both. Ontology modeling languages primarily support ontological distinctions and secondarily (where possible and necessary) translation to encoding languages.

Keywords: Ontology modeling and encoding, semantic engineering, expressiveness, algebra, functional languages, Haskell

1. Introduction

The Semantic Web transcends all previous attempts at enriching data with explicit semantics. Yet, the modeling languages brought to the task are weaker than those for conceptual modeling at smaller scales, such as databases or user interfaces. As a consequence, the Semantic Web rests on the premise that it is possible to produce and understand ontologies in OWL, using editors like Protégé [1]. Many of those who have tried this doubt the premise, especially if they have also done other kinds of conceptual modeling. Their experience in over three decades of conceptual modeling does not support the conclusion that description logic statements adorned with syntactic sugar and design patterns are sufficient (or even necessary) to capture what people mean when they use a vocabulary.

[1] Witness the online guide to Protégé: "The Protégé platform supports two main ways of modelling ontologies – frame-based and OWL" (http://protege.stanford.edu/doc/owl/getting-started.html).

Modeling semantics is a design task, encoding it is an implementation. With the former we explore how to constrain human and machine interpretations of vocabularies, with the latter we support automated reasoning. Expressiveness is an essential criterion for the former, decidability for the latter. Mixing the two concerns is harmful for both tasks, but routinely done. While there are complementary approaches to encode semantics (for example, through machine learning), the scope of this note is limited to conceptual modeling.

2. Can we support our own goals?

Architects do not start a design by constructing geometric figures, database administrators do not start a project by creating relational tables, and user interface designers do not model interfaces in a user interface toolkit. Each of these fields has its modeling languages and environments, allowing, for ex-

ample, to sketch a building, draw diagrams, or compose storyboards.

What does the Semantic Web offer in support of design? How does it assist modelers in choosing ontological distinctions, testing their implications, exploring their varieties, experimenting with applications?

Ontology engineers can choose from informal or semi-formal techniques, such as twenty question games, sorting tasks, concept mapping, or document analysis[2]. They can follow best practice[3] and get advice on pitfalls to avoid[4]. The more formally inclined designers can use methods like OntoClean [5], which ask key questions, for example, about identity and rigidity.

Yet, the gap between informal knowledge elicitation techniques, design patterns, and design methods on the one hand, and useful, tested OWL axioms on the other often remains too wide to jump across without breaking a leg or two. Evidence for this comes in the form of typical problems found in OWL ontologies:

- confusing instance-of with subclass-of;
- confusing part-of with subclass-of;
- leaving the range of an OWL property unspecified;
- introducing concepts and properties that are not sufficiently distinguished from others (a.k.a. "ontological promiscuity").

These and other well-known problems may just be attributed to sloppiness in modeling. However, if it is too easy to be sloppy without noticing it, the Semantic Web will have a serious quality and reputation problem. Also, some of these problems occur in prominent spots, such as Protégé's *Guide to Creating Your First Ontology*[5], which teaches us, for example, to model Côtes d'Or region as a class (!) and furthermore as a subclass-of Bourgogne region.

OWL ontologies cannot be expected to be directly written or understood by modelers, because OWL is optimized to support machine reasoning, not human thought. The Semantic Web is today at a stage of maturity that databases had passed in the 1970's and user interfaces in the 1980's, when they abandoned using a single paradigm for encoding and modeling. Relational algebra for database encoding and logical devices for user interfaces have been complemented by conceptual modeling languages like entity-relationship or state-transition diagrams, and subsequently by much more elaborate design techniques.

A likely consequence of the relatively immature state of modeling support is that today's Semantic Web contains assertions, whose implications have never been understood by anybody, and which may have been tested for satisfiability at best, but not for correctness or relevance.

In the face of this situation, some commonly encountered claims about the goals of the Semantic Web appear rather bold. For example, in a classical paper introducing OWL, it has been said that

"ontologies are expected to be used to provide structured vocabularies that explicate the relationships between different terms, allowing intelligent agents (and humans) to interpret their meaning flexibly yet unambiguously" [7].

The goal of *unambiguous interpretation* is a formidable one, and variants of the idea that ontologies contain "precisely defined meanings" are propagated throughout the Semantic Web literature and in countless project proposals. A recent paper co-authored by the Semantic Web's father, Tim Berners-Lee, even carries it forward to linked data, whose

"meaning is explicitly defined"

according to the authors [1]. Claims like these, if not taken with a very large grain of salt, vastly overstate the achievable goals of the Semantic Web and create expectations that are bound to be disappointed, at least with the currently available modeling support.

In the rest of this note, I will first suggest that the Semantic Web community should adopt a more modest *engineering view* of semantics. Then, I will argue why OWL is too weak for modeling. Finally, I will propose to use and develop modeling languages to complement today's encoding languages.

3. An engineering view of semantics

Neither linguists nor philosophers have so far been able to define *meaning* as an object of scientific study in a way that would capture what people mean when they use vocabularies. Thus, specifying particular "meanings", or targeting unambiguous interpretations, rests on shaky grounds, no matter

[2] see http://www.semanticgrid.org/presentations/ontologies-tutorial/GGFpart4.ppt for an excellent overview

[3] for example, http://www.w3.org/2001/sw/BestPractices/OEP

[4] http://www.ontologyportal.org/Pitfalls.html

[5] http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

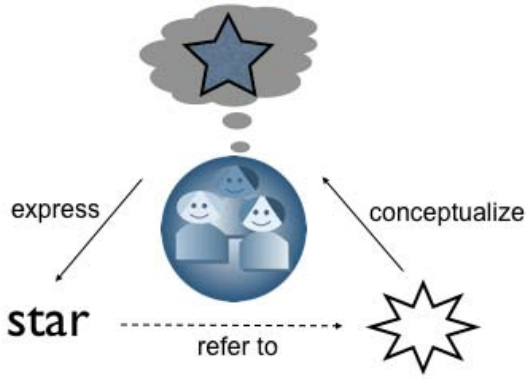

Fig. 1. The meaning triangle.

how it is attempted. Yet, while philosophers figure out what meaning really means, information scientists and engineers can use ontologies pragmatically to *constrain interpretations*.

Ontology engineers have recommended striving for *minimal* ontological commitments [3], rather than for any kind of completeness in ontological specifications. In this spirit, I have recently proposed a pragmatic view of concepts and their specifications [8]. It treats *meaning as a process* to be constrained, rather than as an object to be defined. As the saying "words don't mean, people do" expresses, it is people who mean something when they use a vocabulary, rather than the words having a fixed meaning. Such a process view of meaning can be captured by the threefold view of concepts in the meaning triangle (Fig. 1), involving people using

- words to
- express conceptualizations and to
- refer to *something* in the world.

For example, the community of English speakers uses the word "star" to refer to shapes like that in the bottom-right corner of Fig. 1. Note that a speaker's or listener's idea of a star may not exactly match the instance referred to.

If one adopts such a triadic view of concepts, ontologies do not need to specify "meanings", much less the existence of something in reality. They simply provide humans or machines with constraints on how to apply and interpret vocabularies. These constraints support reasoning, while not removing all ambiguities.

The Semantic Web does not need to and probably cannot raise the bar higher than this semantic engineering goal. Traditional formal semantics is compatible with it, as long as truth is not put before meaning. Truth is a consequence of meaning, just as much as it is a cause, as the cyclic nature of the triangle implies. Truth conditions do not "define" meaning, but constrain interpretations of vocabularies to the ones that make sentences "true". For natural or informal technical languages, where there are no logical truth criteria, this is equivalent to sentences being correctly interpreted in the language community. For example, English speakers need to be able to correctly interpret the word "star" when it is used to refer to the bottom-right shape in Fig. 1.

4. Does OWL support modeling?

Naked OWL code does not convey much insight. Consider how hard it is to understand what somebody else's OWL statements say about a collection of concepts. OWL editors like Protégé provide class and property hierarchies as useful overviews, but do not show how the stated properties interact. Some context and rationale for the statements may be guessed from annotations, but the processes in which concepts are used remain informal and often invisible. Graph representations, showing classes as nodes and relations as edges, can be produced with Protégé plug-ins, but these tend to work only one-way and show only parts of the story. Even professional (and costly) development environments provide only limited and fractioned support for exploring, communicating, and evaluating – in other words for *modeling* before, while, and after encoding.

In addition to these usability and understandability issues, OWL and the current tools around it are often *not expressive enough* for modeling. For example, they

- treat properties and relations as the same, though these are two rather different ideas in modeling;
- limit us to binary relations, though interesting relations often start out as ternary or more;
- provide a well-defined primitive for taxonomies (subclass-of), but not for partonomies and other formal ontological relations;
- make it hard to encode processes and events, though these are often essential elements of semantics.

OWL's expressiveness may be sufficient (or even overkill) for eventual encoding and machine reasoning; but human understanding and reasoning require

more expressive languages and we do not seem to understand yet what these should be. Web searches for "ontology modeling languages", for example, lead either to OWL or to UML or to requirements for better modeling languages with experimental implementations at best. While it is possible and valuable to introduce ontological distinctions into diagrammatic modeling languages like UML [6], automated reasoning support at the modeling stage requires more formal languages. Note that a call for more powerful modeling languages does not contradict the idea of lightweight approaches in implementation. On the contrary, better modeling tends to produce leaner, simpler encodings.

5. Toward formal modeling languages

Modeling ontologies involves tasks like

- *finding out* what should be said,
- *understanding* what has been said,
- *conveying* that understanding to others,
- *checking* whether it is what was intended,
- *spotting errors* in what has been said, and
- *testing* whether what has been said is relevant and useful.

These are human reasoning and communication challenges that are unlikely to be met by reasoning and visualization at the encoding level. As Nicola Guarino proposed in [4], they require ontological distinctions built into modeling languages and automated reasoning support during modeling. What exactly the ontological distinctions should be remains an important research question. The formal ontological distinctions proposed by Guarino (essentially, those of OntoClean [5]) are immensely useful, but appear difficult to build into modeling languages.

Here, I will propose some distinctions that are already available in existing languages. Their choice has been motivated by the work of Joseph Goguen on algebraic theories [2] and 25 years of conceptual modeling applying these ideas, 15 years of which using the functional language Haskell (http://haskell.org), into which many of them are built. Similar ideas and arguments with a larger scope can be found in [9].

A fundamental ontological distinction is that between objects, events, and properties (these labels vary, but the basic idea remains the same). It is quite natural for humans to think in terms of objects and events with properties, rather than just in terms of predicates. Description logics do not allow for this distinction. Functional languages offer *formal* distinctions, based on kinds of functions. For example, properties (say, temperature) can be modeled as functions mapping objects (say, air masses) to values; events (say, a storm) can be modeled as functions mapping between objects (say, air masses again).

With a distinction between objects and events comes the need and opportunity to model the *participation* relation (say, of air masses in storms). Since participation is fundamental for semantics (witness the idea of thematic roles), having it as a modeling primitive would be very useful. *Types* in any language provide it. Data types and operation signatures capture the possible participation of objects in operations. Thereby, typing also provides a theory for instantiation, which is an undefined primitive in description logics: instances of a type are the individuals who can participate in its operations.

Distinguishing *properties* from *relations* is straightforward through unary functions for properties and n-ary (Boolean) functions for n-ary relations. Any interesting ontological relations (such as part-of or location) can then be specified through equational axioms, well known from algebraic specifications. For example, one can specify that an object that was put into another object is in that object as long as it has not been taken out again. Function composition allows for defining semantic constraints involving such sequences of events, which abound in practice. Some recent examples of these and other modeling capabilities of Haskell can be found in [11].

Modeling *roles* is much harder. Their anti-rigidity may be captured through so-called dependent types, whose instances depend on their values. Haskell type classes (generalizations over types) offer a useful model without explicit typing. For example, the types Person and University can be declared as belonging to a type class which provides the functions of enrolling and unenrolling. Students are then modeled by terms like enroll (person, university), rather than by explicit typing.

Haskell's most useful support for modeling, however, stems from the fact that, as a programming language, it allows for *simulation*: ontological specifications can be tested through their constructive model, while and after being developed [12].

6. Conclusions

As Einstein is said to have pointed out, everything should be made as simple as possible, but not simpler. Conceptual modeling in the Semantic Web has been made to look simpler than it is. This carries the risk of yet another turn against Artificial Intelligence in some of its latest incarnations (not only the Semantic Web, but also Linked Data).

Yet, the idea of allowing information sharing with minimal human intervention at run time is too good to be discredited. Therefore, our challenges as researchers in this field are to

- promise only what we honestly believe we can achieve;
- work hard to achieve what we promised;
- validate what we have achieved;
- improve our theories and tools.

These challenges give us some re-thinking and re-tooling to do, and the new journal is a welcome place to report progress. It would be a pity if encoding-biased views about admissible approaches to ontology engineering precluded alternatives that support conceptual modeling.

Acknowledgments

Among the many colleagues who helped shape these ideas over decades in discussions (without necessarily agreeing) are Andrew Frank, Joseph Goguen, Nicola Guarino, Krzysztof Janowicz, Ronald Langacker, David Mark, and Florian Probst, as well as the three reviewers.

References

[1] Bizer, C., Heath, T., Berners-Lee, T., 2009. Linked Data – The Story So Far, http://tomheath.com/papers/bizer-heath-berners-lee-ijswis-linked-data.pdf (preprint of a paper to appear in: Heath, T., Hepp, M., and Bizer, C. (eds.). *Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS)*.

[2] Burstall, R.M., Goguen, J.A., 1977. Putting theories together to make specifications, *Proceedings of the 5th International Joint Conference on Artificial intelligence*, Cambridge, USA: 1045–1058.

[3] Gruber, T.R., 1995. Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, **43**(5–6): 907–928.

[4] Guarino, N., 2009. The Ontological Level: Revisiting 30 Years of Knowledge Representation. *Conceptual Modelling: Foundations and Applications. Essays in Honor of John Mylopoulos*. Edited by A. Borgida, V. Chaudhri, P. Giorgini, E. Yu. Springer-Verlag: 52–67.

[5] Guarino, N. and Welty, C., 2002: Evaluating Ontological Decisions with OntoClean. *Communications of the ACM*, **45**(2): 61–65.

[6] Guizzardi, G., 2005: Ontological Foundations for Structural Conceptual Models. *Telematica Instituut Fundamental Research Series* No. 015.

[7] Horrocks, I., Patel-Schneider, P.F., and Harmelen, F. van, 2003. From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*, **1**(1).

[8] Kuhn, W., 2009. Semantic Engineering. In G. Navratil (Ed.): Research Trends in Geographic Information Science. Springer-Verlag, *Lecture Notes in Geoinformation and Cartography*: 63–74.

[9] Lüttich, K., T. Mossakowski, and B. Krieg-Brückner, 2005. Ontologies for the Semantic Web in CASL. In J. Fiadeiro, editor, WADT 2004, Springer-Verlag, *LNCS* **3423**: 106–125.

[10] Ogden, C.K., Richards, I.A., 1923. *The Meaning of Meaning*. Harcourt Brace Jovanovich.

[11] Ortmann, J., and Kuhn, W., 2010. Affordances as Qualities. *6th International Conference on Formal Ontology in Information Systems (FOIS 2010)*, IOS Press: 117–130.

[12] Raubal, M. and W. Kuhn, 2004. Ontology-Based Task Simulation. *Spatial Cognition and Computation* **4**(1): 15–37.

Semantic Web 1 (2010) 143–147
DOI 10.3233/SW-2010-0013
IOS Press

Digital heritage: Semantic challenges of long-term preservation

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Pascal Hitzler, Wright State University, USA
Open review(s): Krzysztof Janowicz, Pennsylvania State University, USA

Christoph Schlieder
University of Bamberg, Germany
E-mail: christoph.schlieder@uni-bamberg.de

Abstract. The major digital preservation initiatives are about as old as the idea of the Semantic Web but the research areas only had little effect upon each other. This article identifies connections between the two research agendas. Three types of ageing processes are distinguished which affect digital records: media ageing, semantic ageing, and cultural ageing. It is argued that a period of 100 years constitutes an appropriate temporal frame of reference for addressing the problem of semantic ageing. Ongoing format migration constitutes currently the best option for temporal scaling at the semantic level. It can be formulated as an ontology matching problem. Research issues arising from this perspective are formulated that relate to the identification of long-term change patterns of ontologies and the long-term monitoring of ontology usage. Finally, challenges of cultural ageing are discussed.

Keywords: Digital preservation, format migration, ontology matching, ontology change

Introduction

Scalability is widely considered a core objective of the Semantic Web, but it is mainly looked at from a quantitative data perspective, considering, for instance, the number of RDF triples that can be handled. In digital preservation, the focus lies on finding solutions that scale well along the temporal dimension [1]. Memory institutions such as museums care very much whether the documents and/or the data they publish will be accessible in 50 years from now. Considerable effort was invested, for instance, at Emory University, Atlanta, to make the 20-year old digital material of the writer Salman Rushdie accessible by recreating the software environment he used via emulation [3]. Increasingly, less famous individuals ask what sort of digital legacy they will be able to leave with today's technologies. The problem of digital preservation has moved from a concern of specialists to mainstream awareness [15]. In the following, we will explore the challenges of temporal scalability.

In the pre-digital world, the preservation of written records over long periods of time depended on several prerequisites which are rarely made explicit. Firstly, the record needs to be preserved physically. Secondly, the semantic capabilities to read and interpret the records have to persist. A reader of a clay tablet, for instance, has to master a particular form of cuneiform writing and the Acadian language. Thirdly, there must be a community that (still) shows interest in the record. Only an interested community will mobilize the resources required to teach its members complex semantic skills or to even engage in the deciphering of extinct languages. In many respects, the preservation of digital records faces similar problems.

The ageing of digital records

One reason for which digital preservation can fail is media ageing. Any medium that carries a digital encoding will physically deteriorate until it is no longer possible to recover the original bit stream. This process of media ageing has received much attention from memory institutions but it seems less critical for the Web with its capacity to easily replicate data.

Like written records before the computer, digital content is affected by *semantic ageing*, that is, the evolution of data formats and the fact that knowledge about data semantics quickly disappears if not specified explicitly. Finally, there is a process which may be called *cultural ageing*. This process is rarely discussed in connection with digital preservation. Gradually, the community loses interest in some particular content. The corresponding documents are no longer retrieved, the data is no longer used in inferences. Knowledge about the semantics of digital records may persist for a while after the community loses interest in their content. However, as the semantic knowledge is not maintained and transmitted any more, its loss is almost unavoidable.

Choosing a temporal frame of reference

Before identifying the semantic challenges of digital preservation it is important to determine at which temporal scale to address the issue. At the short-term end there is the time frame which the legal regulations of many countries provide for the preservation of business documents, namely 10 years. The market offers a number of archiving solutions which handle digital preservation at this scale by using archiving formats with are maintained for at least a decade. The preservation problem may be considered solved at this scale.

Probably, the most ambitious temporal frame of reference considered for digital preservation is the formidable period of 10.000 years promoted by the Long Now Foundation [2]. Without doubt, it is intellectually challenging to look at ten millennia, the relevant unit of analysis for a number of global problems such as climate change. It is less clear, however, in what way such a very long-term perspective fosters the emergence of technological solutions (e.g. format repositories) radically different from those currently discussed in digital preservation.

For the purpose of this article a much more modest frame of reference is chosen, a period of 100 years, which is more accessible to empirical evaluation as well as closer to personal experience. Centering this frame of reference upon the present sets a double agenda: (1) finding strategies to access digital contents from the past 50 years in spite of media ageing and semantic ageing, (2) planning the preservation of currently accessible digital content for future use during the 50 years to come.

A major consequence of this specific planning horizon consists in the fact that the problem of semantic ageing cannot be solved anymore by simply agreeing on a standard format for digital archiving. Half a century is just plenty of time for requirements to evolve beyond any standard. This holds even for plain text as the chronology of character formats illustrates. The year 1963 witnessed the first edition of the ASCII standard which ceased to evolve with a last update in 1986. In the same year, the Latin-1 character set was published which became part of the ISO 8859 series of standards. ISO ceased maintenance of these standards in 2004 to concentrate its resources on Unicode. The example shows that even for data of little semantic complexity only a sequence of standards was able to bridge a period of almost 50 years. Note also that the end of a standard´s evolution does not imply the end of its usage.

Digital preservation and the Semantic Web

Digital preservation has been a very real concern of memory institutions who addressed the problem long before the problem of an impending “digital dark age” [10] became known to a wider audience. Public funding of the major research initiatives started around the turn of the millennium, notably the National Digital Information Infrastructure and Preservation Program (NDIIPP) established in 2000 by the US congress and comparable European research initiatives. In other words, the mainstream of digital preservation research is about as old as the Semantic Web. Unfortunately, both strands of research have only interacted in rather limited ways so far.

The digital preservation initiatives basically explored two families of approaches for the problem of semantic ageing: migration and emulation – as well as combinations of both. *Migration* is especially interesting for document-centered workflows, in-

cluding those used in the humanities and in cultural-historic research. The ideal target format for migration is published under an open source license, comes with an explicit account of its semantics, and possesses a large community of users.

Emulation constitutes the best solution for archives of highly interactive media, e.g. interactive art or video games. Emulation strives for authenticity, for a reenactment of a user experience from the past. However, being able to run the software which created the data does not per se make it interoperable with present day technology. Migration, on the other hand, aims at the integration of past content into future knowledge-based workflows. Because of the focus on data and interoperability, migration seems to blend more easily with the different flavors of the Semantic Web – definitely with the idea of a Web of semantically interoperable knowledge bases but to a certain extent also with the more recent idea of a Web of Linked Data.

At least two levels can be distinguished at which migration strategies are currently supported by Semantic Web technologies: the preservation planning level and the semantic transformation level. A major result of the digital preservation initiatives was to conceive preservation as an ongoing process based on an appropriate digital curation lifecyle model (e.g. [4,16]). Preservation planning is a central element of such a life cycle model. At the *preservation planning level*, migration strategies are implemented by services that monitor data formats and data access mechanisms on the one hand and available migration tools on the other hand. Emerging risks are assessed (obsolescence detection) and recommendations for migration pathways are generated. A first link between the world of digital preservation and the world of the Semantic Web exists at this planning level. Preservation services have been described as Semantic Web services, for instance, in the PANIC system [9] which uses an extension of the OWL-S ontology to describe preservation-specific services and computes semantic matches to support service discovery.

Services such as format transformations are based on the preservation metadata that comes with the digital records. This works best for atomic single media records but becomes more difficult for composite multimedia records. For highly complex data objects such as those produced in the architecture, engineering, and construction (AEC) industry by special-purpose CAD systems ready-to-use migration services simply do not exist [5]. In such cases, before addressing migration at the preservation planning level, it has to be implemented at the *semantic transformation level*. For specialized domains such as architectural drawings, archival formats start to emerge although they have some limitations from the point of view of digital preservation [14].

However, many projects in the AEC industry have documentation needs that require some sort of application specific semantic modeling. It is near at hand to use ontological modeling languages such as OWL for describing those application ontologies and for relating them to the domain ontology [8]. This is a second link that has been established between the world of digital preservation and the world of the Semantic Web.

Once that data semantics is captured by ontological modeling, the problem of migrating from one data format to another can be described as an ontology matching problem which transforms a source ontology into a target ontology [6]. Format migration is thus closely related to ontology change as defined in [7]. In adopting such an approach, we must, however, be aware of the general limitations of ontological modeling. While many aspects of data semantics are easily captured by modeling formalisms such as description logics, some aspects of the semantics of natural languages are difficult to render. The same holds for epistemic drifts that are not reflected by a change of the logical modeling of data.

The challenges of semantic ageing

Within this setting – digital preservation based on ongoing format migration modeled as a sequence of ontology changes – a number of challenges arise. In one way or the other, they are all related to the issue of how well solutions scale over the chosen temporal frame of reference of 100 years, or rather 50 years, if only forward preservation is considered.

Identifying long-term patterns of ontology change

Only by looking at periods that are significantly longer than the 10 years handled by current technology, it can be determined how the changes in the ontologies which cause semantic ageing distribute over time. There seem to be change processes with an almost constant rate of change. However, in many cases, changes occur in bulk. Open research questions relating to ontology change include:

- What different change patterns are there? Do they depend on the type of ontology (top-level, domain, task, application)?
- Is it possible to predict impending bulk changes by analyzing the time series of changes and the structural complexity of the ontology?
- Media ageing is studied by artificial ageing processes. Can similar simulation approaches be designed for semantic ageing?

Monitoring long-term usage of ontologies

Software tools with rich functionality (e.g. special-purpose CAD systems), tend to generate data with complex semantic relationships. Often, however, only a fraction of the functionality is used to actually create a digital record (e.g. a CAD document with only 2D geometries). Migration at the semantic transformation level would be greatly simplified if for a collection of digital records it is known whether there are parts of the source ontology that are not used by any of the records, or only used by very few. The long-term evolution of ontology usage has not been studied so far. Issues to be addressed in this context include:

- How does the population of classes with instances change over long time intervals?
- Which instances are actually used in queries and inferences? Do usage patterns change over time?
- How can information about ontology usage patterns help to improve ontology matching?
- How is ontology usage monitoring best integrated into preservation life-cycle management?

The challenges of cultural ageing

The creation of meaning by communities is an ongoing process which is inevitably accompanied by an antagonistic process in which meaning is lost. Pre-computer history is full of examples for this process of cultural ageing which affects natural languages and their writing systems as well as complex belief systems such as religions. Cultural ageing has at least a technical benefit. Only the records that a community still shows interest in will be migrated which reduces the semantic translation workload by orders of magnitude. The downside is also evident. Processes of cultural renewal ("renaissances") which generate interest in content that was considered uninteresting for generations are not possible.

At present, cultural ageing does not constitute a focus of research on digital preservation. Probably, this is due to the fact that there are sufficiently many other problems that seem more pressing. On the other hand, it is difficult to imagine a satisfactory solution to digital preservation which does not take the mechanisms of cultural ageing into account. This is not so much a matter of trying to prevent cultural ageing – a hopeless task – rather than to monitor the community's access to the digital records and to identify content that will become vulnerable to semantic ageing because of the community's loss of interest.

Web archiving provides a good example of how the digital version of cultural ageing operates. Defining a selection policy for the Web sites that are going to be preserved constitutes the crucial first step in the design of any Web archive [13]. Different computational methods have been proposed to determine the relevance of a web page such as Google's PageRank or the HITS algorithm [11]. The underlying problem – measuring visibility in a large-scale communicative processes – has been studied from many disciplinary perspectives including social network analysis and bibliometrics. Simulation studies can show how visibility evolves over time [12].

Monitoring cultural ageing

A similar type of selection has been effective in the pre-computer era archives. What is new, is the quantitative scale, that is, the number of records for which choices need to be made. The choice is not necessarily one of inclusion or exclusion but rather a decision about the quality level at which semantic ageing is dealt with. An online journal with a high impact factor will probably enjoy a premium migration process involving human intervention which even preserves, for instance, interactive 3D-models and HD videos while the automatic standard migration process for less visible journals is going to concentrate on preserving just text, tables, and images. A number of research problems are connected to the selection process triggered by cultural ageing.

- Which is the appropriate way to quantify the visibilty of documents and/or data in the relevant communicative processes, e.g. in scientific communication in the humanities?
- What is an adequate formal model of the loss of semantics triggered by cultural ageing?
- How can digital preservation reflect the plurality of interests that different communities show?
- What type of preservation planning will permit or even encourage the rediscovery of documents or data?

Conclusions

Although the digital dark age is a menace of the present, the processes of media ageing, semantic ageing, and cultural ageing have been effective since pre-computer times. In a world of distributed digital data, however, semantic ageing constitutes a bigger problem than media ageing. The best way to overcome the effects of semantic ageing is by migrating digital records into new formats. We have seen how Semantic Web technologies support migration at the preservation planning level as well as at the semantic transformation level. Research challenges have been formulated for semantic ageing and for cultural ageing.

Long-term preservation constitutes an application field that forces the Semantic Web research community to adopt a much longer temporal frame of reference. By doing so it places the study of ontology change under a new perspective which focuses, among others, on the changes of ontology use and on the changing population of classes with instances (categorization patterns of instance data).

Taking cultural ageing seriously means to abandon the idea that digital preservation operates like a time capsule. The picture of content that is enclosed in a digital capsule to be opened at some moment in the future is misleading because it is not the past that sends messages to the future. Rather, it is the present that makes choices, selecting content from the past and linking to it. This ongoing process of linking from the present into the past makes up digital heritage.

References

[1] Borghoff, U., Rödig, P., Scheffczyk, J., and Schmitz, L. (2006) Long-term Preservation of Digital Documents: *Principles and Practices*, Springer, Berlin.

[2] Brand, S. (2000) *Clock of the Long Now: Time and Responsability*, Basic Books, New York, NY.

[3] Cohen, P. (March 16, 2010) Fending Off Digital Decay, Bit by Bit. The New York Times, C1.

[4] Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., et al. (2009) DCC&U: An Extended Digital Curation Lifecycle Model. *The International Journal of Digital Curation*, **4**(1), Article 3.

[5] Doyle, J., Viktor, H., and Paquet, E. (2009) A Metadata Framework for Long Term Digital Preservation of 3D Data. *International Journal of Information Studies*, **1**(3), 165–171.

[6] Euzenat, J. and Shvaiko, P. (2007) *Ontology Matching*, Springer, Berlin.

[7] Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., and Antoniou, G. (2008) Ontology Change: Classification and Survey. *The Knowledge Engineering Review*, **23**(2), 117–152.

[8] Freitag, B. and Schlieder, C. (2009) MonArch – Digital Archives for Monumental Buildings. *KI*, **23**(4), 30–35.

[9] Hunter, J. and Choudhury, S. (2006) PANIC – An Integrated Approach to the Preservation of Composite Digital Objects using Semantic Web Services. *International Journal on Digital Libraries*, **6**(2), 174–183.

[10] Kuny, T. (1997) A Digital Dark Ages? Challenges in the Preservation of Electronic Information. Paper at the 63rd IFLA Council and General Conference, Workshop on Preservation and Conservation, http://archive.ifla.org/IV/ifla63/63kuny1.pdf (2 Apr 2010).

[11] Langville, A. and Meyer, C. (2006) *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, Princeton, NJ.

[12] Malsch, T., Schlieder, C., Kiefer, P., Lübcke, M., Schmitt, M., and Stein, K. (2007) Communication Between Process and Structure: Modelling and Simulating Message Reference Networks with COM/TE. *Journal of Artificial Societies and Social Simulation*, **10**(1), Article 9.

[13] Masanès, J. (2006) *Selection for Web Archives in Web Archiving*, Masanès, J. (ed), Springer, Berlin, pp. 71–90.

[14] Smith, M. (2009) Curating Architectural 3D Models. *The International Journal of Digital Curation*, **4**(1), Article 8.

[15] Solvberg, I. and Rauber, A. (2010) *Digital Preservation in Digital Preservation*, Solvberg, I. and Rauber, A. (eds), European Research Consortium for Informatics and Mathematics, pp. 12–13.

[16] Strodl, S., Becker, C., Neumayer, R., and Rauber, A. (2007) How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure in *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries. JCDL 2007; Vancouver, British Columbia, Canada, June 18–23, 2007*, Rasmussen, E. and Larson, R. (eds), ACM Press, New York, NY, pp. 29–38.

Semantic Web 1 (2010) 33–37
DOI 10.3233/SW-2010-0014
IOS Press

Preventing ontology interoperability problems instead of solving them

Editors: Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited reviews: Giancarlo Guizzardi, Federal University of Espírito Santo (UFES), Brasil; Martin Raubal, University of California, Santa Barbara, USA

Eero Hyvönen
Semantic Computing Research Group (SeCo), Aalto University and University of Helsinki, P.O. Box 15500, FI-00076 Aalto, Finland
E-mail: Eero.Hyvonen@cs.Helsinki.fi

Abstract. A major source of interoperability problems on the Semantic Web are the different vocabularies used in metadata descriptions. This paper argues that instead of *solving* interoperability problems we should focus more effort on *avoiding* the problems in the first place, in the spirit of Albert Einstein's quote "*Intellectuals solve problems, geniuses prevent them*". For this purpose, coordinated collaborative development of open source vocabularies and centralized publication of them as public vocabulary services are proposed. Methods, guidelines, and tools to facilitate this have been developed on a national level in the Finnish FinnONTO initiative, and are now in pilot use with applications and promising first results

Keywords: Interoperability, vocabularies, ontology libraries, ontology services

1. Interoperability of vocabularies

Much of the power of the Web comes from the freedom for anybody to publish and link his/her own content as the Web of Pages. When moving into the era of the Semantic Web, the Web of (Linked) Data, content is being linked on the level of ontological concepts and metadata underlying the pages[1] [3]. This leads to interoperability problems, especially interoperability regarding metadata schemas and vocabularies used for filling element values in the schemas. Approaches to schema interoperability include the dumb-down principle, as suggested in the Dublin Core (DC) community[2], and using a shared schema ontology onto which other metadata representations can be transformed, as suggested by the CIDOC CRM and FRBR communities[3]. In contrast, this paper focuses on interoperability problems due to *domain vocabularies* (ontologies of hierarchically organized domain-specific concepts) used in annotations, not to schema models such as DC of CIDOC CRM that are also sometimes called "vocabularies" or "ontologies".

Content aggregated in semantic portals, or on the web scale in the Linked Data initiative, comes from actors and organizations that produce content for their own purposes and come from different disciplines, cultures, and countries. As a result, lots of different, partly overlapping vocabularies are used in metadata descriptions. To approach the interoperability problems, various techniques of ontology matching (mapping) [5] are used. For example, lots of mappings based on the *owl:sameAs* relation have been created for the resources in the Linked Data cloud. There are, for example, mappings between the place resources of DBPedia[4] and GeoNames[5]. A

[1] http://linkeddata.org/
[2] http://dublincore.org/
[3] http://cidoc.ics.forth.gr/
[4] http://dbpedia.org/

key problem here is how to deal with situations, where multiple entity names and identifiers are used for a single real world object [4], and where different objects have the same name or identifier. The same problem is encountered in Web 2.0 sites, where tagging using literals without identified meaning is causing more and more semantic confusion as more and more tags are being created (e.g. "jaguar" as an animal, or a car or an airplane model).

2. Coordinated collaboration for vocabulary creation

The mess of meaning references on the metadata level on the Semantic Web creates lots of interesting research problems to study. Most research on interoperability issues seems to be focusing on developing methods and tools for obtaining interoperability between heterogeneous annotations (e.g. the datasets of the Linked Data initiative). However, from a non-academic practical viewpoint, this is a problem that should be avoided in the first place as far as possible. Obviously, more research effort should be focused on developing methods, tools, and practices by which metadata could be produced on a larger scale in an interoperable way at the time of creating it. Instead of solving interoperability problems we should rather try to prevent them by better ontology services, coordination, and collaboration in ontology development and content creation.

FinnONTO[6] 2003–2012 is a research project and a Living Laboratory experiment [6,7], where the idea is to establish a collaboration framework for vocabulary development and services on a national level for the Semantic Web. The main goal of FinnONTO is to create an open source, national level cross-domain "content infrastructure" for the Semantic Web, aligned with international vocabularies, standards, and practices. This infrastructure and network of concepts can be paralleled, on a conceptual level, with the construction of railroad, electrical, or telephone networks in the past.

The work is based on the domain independent Semantic Web standards[7] of the W3C, such as RDF, SKOS, OWL, and SPARQL, but the heart of the system is domain-specific ontologies. While standardization work at W3C focuses on defining general principles of ontological structuring and reasoning, such as subsumption and inheritance, the general goal of FinnONTO is to facilitate cross-domain interoperability of metadata descriptions on a domain-specific vocabulary level. The idea is that when content is published on the web, it should be possible to connect it semantically with other related (cross-domain) contents based on a system of mutually aligned domain ontologies.

The vocabulary infrastructure has been built by transforming nationally used traditional keyword thesauri [1] into lightweight ontologies, which makes the ontologies interoperable with already indexed content in databases. A key goal in the work is to encourage collaboration between ontology developers of different domains by proving a general FinnONTO ontology framework in which new vocabularies can be aligned with existing ones already during the ontologization process, instead of afterwards. The kernel of the FinnONTO system [7] is the General Finnish Ontology YSO developed from the widely used General Finnish Thesaurus YSA that consists of some 25,000 general concepts and that is maintained by the National Library of Finland. The corresponding ontology YSO has been extended by various domain-specific daughter ontologies, based on other national thesauri used in domains such cultural heritage, agriculture and forestry, applied arts, geography, photography, and others. These ontologies create together virtually one ontology, the Collaborative Holistic Ontology KOKO, that now has over 70,000 general concepts, not including ontology-like datasets, such as places, persons, mammal and bird species of the world, and historical events.[8]

Figure 1 illustrates the structure of KOKO, with the top ontology YSO on top, and overlapping domain ontologies AFO (agriculture and forestry), MAO (cultural heritage), TAO (applied art), and VALO (photography) extending its concept hierarchies.

3. Commandments for social vocabulary development

The key idea in the ontologization process is to aim at a system of vocabularies that are *born interoperable* with each other. To facilitate this, a new

[5] http://geonames.org/

[6] http://www.seco.tkk.fi/projects/finnonto/

[7] http://www.w3.org/standards/semanticweb/

[8] http://www.seco.tkk.fi/ontologies/

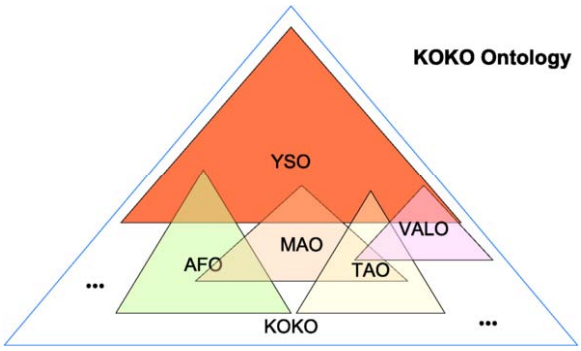

Fig. 1. KOKO system of overlapping aligned ontologies [6].

thesaurus is first matched with the general YSO top ontology in order to identify potential overlaps. The result is a Protégé editor[9] project that includes YSO and the new thesaurus concepts. This structure is then corrected and maintained manually. (Alignment with other overlapping ontologies is also possible in a similar way.) In this way, the work already done in YSO can be reused in daughter ontologies and, at the same time, interoperability is enhanced by collaboration.

Vocabulary work in our view is as much a social process as it is a technical challenge. The work is guided by the following principles or "ten commandments":

1. Add machine semantics to legacy vocabularies. Start transforming thesauri [1] into machine interpretable lightweight ontologies in order to boost their usage on the Semantic Web.
2. Think cross-domain. Consider not only your own micro world, but also cross-domain usage of concepts when making ontological decisions.
3. Establish collaboration networks of domain expert groups. Nobody masters the whole universe.
4. Reuse others' work.
5. Maintain interoperability with the past and other ontologies. Otherwise benefits of collaboration are lost.
6. Proceed in small steps. Adding even little semantics can be very useful (and keeps e.g. the funding agencies happy).
7. Respect different ontological views. It is not possible to come up with only one ontological view of the world.
8. Accept imperfect models. The ontology will never be fully perfect.
9. Minimal ontological commitment. Keep ontological structures simple and generic in order to facilitate cross-domain reuse.
10. Coordinate the work and add new commandments if needed. This is done now by the FinnONTO research project but later, if the project is successful, by another coordinating organization.

[9] http://protege.stanford.edu/

4. Vocabulary services for legacy systems

Another key component of the FinnONTO infrastructure is the National Ontology Service ONKI[10] [15] hosting currently over 80 ontologies and vocabularies. The idea is provide the vocabularies as a free open source service for both human *and* machine users to utilize. ONKI ontology services such as concept finding, browsing, fetching, and query expansion [13,14] can be integrated with legacy systems through REST, Web Service, or AJAX APIs in a way that is analogous to using Google Maps as an external service in applications. We hope that by making vocabulary services available and usable in an economically motivating way, organizations and people start using shared ONKI vocabularies and URIs, preventing interoperability problems rising form using local or depreciated vocabularies, and ambiguous literal terms in annotations. Other ontology servers on the web with the goal of publishing and sharing ontologies in public include Cupboard [2] and BioPortal [11].

5. Evaluation

The feasibility of the FinnONTO approach is tested and demonstrated in practice by applications, such as the collaborative semantic portals MuseumFinland[11], HealthFinland[12] [12], and CultureSampo[13] [8] that makes use of the whole KOKO system aligned with some international vocabularies, such as the Getty vocabularies[14] AAT, TGN, and ULAN.

[10] http://www.onki.fi/
[11] http://www.museosuomi.fi/
[12] http://www.tervesuomi.fi/
[13] http://www.kulttuurisampo.fi/
[14] http://www.getty.edu/research/conducting_research/vocabularies/

In summer 2009, 150 organizations in Finland and abroad had been registered to use ONKI services, and new ontologized vocabularies in the system have been developed by external organizations, e.g. an ontology for maritime terms (MERO) and for literature content (KAUNO). The latter one that has been used, based on the ONKI services, for annotating over 50,000 pieces of Finnish novels and short stories in a Web 2.0 fashion by Finnish librarians for the semantic literature portal Kirjasampo[15]. In HealthFinland metadata is being created using the ONKI ontologies and services by a variety of national health organizations, and the system is in use[16] and maintained by the National Institute for Health and Welfare since 2009 [12].

Our own experience suggests that gaining semantic interoperability in terms of vocabularies is a very tedious task and hinders fast publication cycle form legacy databases to the Web. In CultureSampo, for example, the content is harvested from tens of museums, libraries, archives, media companies, and web sources producing heterogeneous content. The vocabulary interoperability problem should in our mind definitely be addressed seriously at the time and place of content creation, rather than after harvesting the content, and we hope that the FinnONTO infrastructure is a step towards facilitating this in practice.

6. Discussion

Changing the established practices of vocabulary development, and adapting software in legacy systems to use ontologies cannot happen instantly but only over time. However, we believe there is now a promising road ahead to go based on the collaborative FinnONTO approach, although many problems of interoperable ontology development need to be addressed in the future.

A concern is the management of changes in the evolving ontologies and their alignments. Ontology versioning is needed because 1) the underlying real world or 2) our conceptualization about it may change [10], or 3) the underlying vocabulary standards evolve. Here one faces the problem that old content has been annotated using an old vocabulary while the end-user or applications may use a modern vocabulary or different old vocabularies. To address the problem, alignments between vocabulary versions along the temporal dimension are needed. An approach to modeling temporal ontology changes was developed in the Finnish Spatio-temporal Ontology SAPO[17] modeling over 1000 geographical changes of Finnish counties (e.g. boundaries and names) since 1865 [9].

An important question in sharing ontologies is application specificity or point of view dependency. An ontology developed from one point of view may not be usable from another perspective. To pursue application independence, the FinnONTO vocabularies are kept lightweight with as little ontological commitment to applications as possible. The vocabularies provide only little more than the skeletal RDFS subsumption hierarchy of concepts, and it is left up to the applications to build more domain specific semantics based on that.

End-users, domain experts, and ontology engineers may have different views to a domain. In such cases, different separate ontologies for the same domain may be needed, aligned with each other. For example, in the HealthFinland portal, the content is annotated using domain expert vocabularies, such as Medical Subject Headings (MeSH)[18], but the vocabularies provided for the citizen end-users in the faceted search engine are based on layman's concepts extracted using a card-sorting technique [12].

Still another concern is whether two vocabularies sharing the same concepts should share the same ontological structure, too. Since there can be different views and opinions to modeling the real world, the modeling choices in a vocabulary in FinnONTO can be made independently from those in other overlapping vocabularies. The FinnONTO framework only makes the different vocabularies and views visible to all parties, encouraging but not forcing to sharing structures.

It is our hope that supporting collaboration in distributed ontology development facilitates cost-efficient creation of large cross-domain vocabularies with better interoperability than using a centralized approach or distributed development without coordination. It is also our hope that supporting social collaboration will lead to ontologies of better quality. By using the ONKI service for ontology publishing, the results of the joint efforts can be utilized in practical applications easily as ready-to-use services—both by human and machine end-users.

[15] http://www.kirjasampo.fi/
[16] http://www.tervesuomi.fi/
[17] http://www.seco.tkk.fi/ontologies/sapo/
[18] http://www.nlm.nih.gov/mesh/

Acknowledgements

Thanks to Martin Raubal, Giancarlo Guizzardi, and Krzysztof Janowicz for fruitful comments concerning an earlier version of this article. Tens of researchers have been working in the various parts of the FinnONTO 2003–2012 project. The research has been funded by the Finnish Funding Agency for Technology and Innovation (Tekes), and a consortium of over 40 companies and public organizations.

References

[1] Jean Aitchison, Alan Gilchrist, David Bawden: *Thesaurus Construction and Use: A Practical Manual*. Routledge, 2000.

[2] M. d'Aquin, H. Lewen: Cupboard—A Place to Expose Your Ontologies to Applications and the Community. In: *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, Springer-Verlag, 2009.

[3] C. Bizer, T. Heath, T. Berners-Lee: Linked Data—The Story So Far. In: *Journal of Semantic Web and Information Systems*, Vol. 5, No. 3, 2009.

[4] Paolo Bouquet, Heiko Stoermer, Claudia Niederee, Antonio Mana: Entity Name System: The Backbone of an Open and Scalable Web of Data. In: *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2008)*. IEEE Computer Society, 2008, pp. 554–561.

[5] J. Euzenat, P. Shvaiko: *Ontology Matching*. Springer-Verlag, 2007.

[6] Eero Hyvönen: Developing and Using a National Cross-Domain Semantic Web Infrastructure. In P. Sheu et al. (eds): *Semantic Computing*. IEEE Wiley, 2010.

[7] Eero Hyvönen, Kim Viljanen, Jouni Tuominen, Katri Seppälä: Building a National Semantic Web Ontology and Ontology Service Infrastructure—The FinnONTO Approach. *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)*, Springer-Verlag, 2008.

[8] Eero Hyvönen, Eetu Mäkelä, Tomi Kauppinen, Olli Alm, Jussi Kurki, Tuukka Ruotsalo, Katri Seppälä, Joeli Takala, Kimmo Puputti, Heini Kuittinen, Kim Viljanen, Jouni Tuominen, Tuomas Palonen, Matias Frosterus, Reetta Sinkkilä, Panu Paakkarinen, Joonas Laitio, Katariina Nyberg: CultureSampo—Finnish Culture on the Semantic Web 2.0. Thematic Perspectives for the End-user. In: J. Trant, D. Bearman (eds): *Museums and the Web 2009: Proceedings. Archives & Museum Informatics*, Toronto.

[9] Tomi Kauppinen, Eero Hyvönen: Modeling and Reasoning about Changes in Ontology Time Series. In: Rajiv Kishore, Ram Ramesh, Raj Sharman (eds): *Ontologies: A Handbook of Principles, Concepts and Applications in Information Systems*. Springer-Verlag, 2007, pp. 319–338,

[10] M. Klein: Change Management for Distributed Ontologies. PhD thesis, Free University of Amsterdam, The Netherlands, 2004.

[11] M. Musen, N. Shah, N. Noy, B. Dai, M. Dorf, N. Griffith, J. Buntrock, C. Jonquet, M. Montegut, D. Rubin: BioPortal: Ontologies and Data Resources with the Click of a Mouse. In: *Proceedings of the Annual Symposium Proceedings/AMIA Symposium*, 2008.

[12] Osma Suominen, Eero Hyvönen, Kim Viljanen, Eija Hukka: HealthFinland—a National Semantic Publishing Network and Portal for Health Information. *Journal of Web Semantics*, vol. 7, no. 4, 2009, pp. 271–376.

[13] Jouni Tuominen, Matias Frosterus, Kim Viljanen, Eero Hyvönen: ONKI SKOS Server for Publishing and Utilizing SKOS Vocabularies and Ontologies as Services. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*. Springer-Verlag, 2009.

[14] Jouni Tuominen, Tomi Kauppinen, Kim Viljanen, Eero Hyvönen: Ontology-based Query Expansion Widget for Information Retrieval. *Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009). CEUR Workshop Proceedings,* Vol. 449, http://ceur-ws.org/, 2009.

[15] Kim Viljanen, Jouni Tuominen, Eero Hyvönen: Ontology Libraries for Production Use: The Finnish Ontology Library Service ONKI. *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*. Springer-Verlag, 2009.

Semantic Web 1 (2010) 3–10
DOI 10.3233/SW-2010-0015
IOS Press

Theoretical foundations and engineering tools for building ontologies as reference conceptual models

Editors: Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited Reviews: Oscar Corcho, Universidad Politécnica de Madrid, Spain; Pascal Hitzler, Wright State University, USA

Giancarlo Guizzardi
Ontology and Conceptual Modeling Research Group (NEMO), Computer Science Department, Federal University of Espírito Santo (UFES), Vitória, Espírito Santo, Brazil
E-mail: gguizzardi@inf.ufes.br

Abstract. Perhaps the most fundamental notion underlying the desiderata for a successful Semantic Web is *Semantic Interoperability*. In this context, ontologies have been more and more recognized as one of the enabling technologies. This paper defends the view that an approach which neglects the role of ontologies as *reference conceptual models* cannot meet the requirements for full semantic interoperability. The paper starts by offering an engineering view on ontology engineering, discussing the relation between *ontologies as conceptual models* and *ontologies as codification artifacts*. Furthermore, it discusses the importance of foundational theories and principles to the design of ontology (conceptual) modeling languages and models, emphasizing the fundamental role played by *true ontological notions* in this process. Finally, it elaborates on the need for proper tools to handle the complexity of ontology engineering in industrial scenarios and complex domains. These tools include *ontological design patterns* as well as well-founded computational environments to support ontology creation, verification and validation (via model simulation).

Keywords: Conceptual modeling, foundational ontology, methodological and computational tools for ontology engineering

1. Introduction

Perhaps the most fundamental notion underlying the desiderata for a successful Semantic Web is *Semantic Interoperability*. To a large extent, the Semantic Web is about offering support for complex information services by combining information sources that have been designed in a concurrent and distributed manner. In this context, ontologies have been more and more recognized as one of the enabling technologies.

In general, in computer science, ontologies have been used either as a *reference model of consensus* to support semantic interoperability, or as an explicit, declarative and machine processable artifact coding a domain model to enable automated reasoning. This duality, however, points to different (and even conflicting) sets of quality criteria that should be met by the representation languages employed to construct these ontologies.

On one hand, ontologies considered as reference conceptual models for semantic interoperability should be constructed in manners that maximize, on one hand, the expressivity in capturing fundamental aspects of the underlying domain and in making explicit the underlying *ontological commitments*. On the other hand, they should also be designed to maximize conceptual clarity (or pragmatic efficiency) to afford the tasks of communication, domain understanding, problem-solving and meaning negotiation among human users. In contrast, ontologies as reasoning artifacts for the semantic web, should be

built in a way that supports decidable and computationally tractable automated reasoning.

The first idea defended in this article can be summarized in the following manner. If Ontology Engineering is to become a mature engineering discipline, able to construct and manage artifacts in a range of complex domains, it must incorporate a number of lessons learned from other closely-related engineering disciplines. This starts with the acknowledgement that *there is no Silver Bullet*! From a language point of view, this means that we should not attempt to produce one single representation system (with associated methodological tools). In contrast, we should recognize that different representation systems of different nature are needed in different phases of an ontology engineering process. This idea is articulated in Section 2.

The second point I want to make is that, in order to meet the quality criteria outlined above for producing ontologies as reference conceptual models (i.e., ontological expressivity and conceptual clarity), we cannot eschew *truly ontological* questions. In other words, we need an ontology conceptual modeling language that assists modelers in: (i) making explicit the ontological commitment assumed in that conceptualization; (ii) producing representation structures that do justice to the nature of the underlying reality. The design of such a language can greatly benefit from theories produced in disciplines such as Formal Ontology in Philosophy, Cognitive Science, Linguistics and Philosophical Logics. This point is elaborated in Section 3.

Finally, for us to be able to count on a systematic engineering discipline that can be used to establish full and successful semantic interoperability in heterogeneous and complex real-world scenarios, we need proper methodological and computational tools to handle that complexity. This is the third topic of this paper which is discussed in Section 4.

These three points are non-orthogonal, in the sense that, in the way I have presented in this article, the solutions outlined in Section 4 dependent on the acceptance and implementation of the views advocated in Sections 2 and 3. For this reason, the latter sections can also be understood as background knowledge for the former. Furthermore, this accounts for a certain unbalance in length between these sections.

2. An engineering view to ontology engineering

A domain ontology is a special kind of conceptual model, i.e. an engineering artifact with the additional requirement to represent a model of consensus within a community. This model is designed to facilitate individuals to share information about that domain by conforming to some standard set of constructs. For this reason, this activity should be structured in process paths that are analogous to the ones practiced in other disciplines that also support the transition from a representation of a conceptualization to some coding artifact. In this transition path, the process must take into account that the produced coding artifact should preserve not only the real-world semantics of the original representation but it should also typically comply with a number of non-functional requirements particular to a specific computational environment.

In disciplines such as Software and Information Systems Engineering, there is a clear distinction between Conceptual Modeling, Design and Implementation. In Conceptual Modeling, a solution-independent specification is produced with the aim to make a clear and precise description of the domain elements. In the Design phase, this conceptual specification is transformed into a logical design specification (e.g. a relational database schema or an object class model) by taking into consideration a number of issues ranging from architectural styles, non-functional quality criteria to be maximized (e.g., performance, adaptability), target implementation environment, etc. The same conceptual specification can potentially be used to produce a number of (even radically) different logical designs. Finally, in the Implementation phase, a physical design is coded in one or more target languages to be then deployed in a computational environment. Again, from the same logical design, a number of different implementations can be produced. Design, thus, bridges Conceptual Modeling and Implementation.

The same reasoning should be applied to the discipline of Ontology Engineering [8]. Firstly, in a conceptual modeling phase in Ontology Engineering, the main requirements for the resultant models (and, hence, for the modeling languages) are *domain appropriateness* and *comprehensibility appropriateness* [7]. These requirements mean that on the one hand, the models should be truthful to the phenomena being represented. And on the other hand, it should be clear for users of the language to understand what elements of the universe of discourse are represented by elements of the model, as well as what problem-solving operations are to be performed on these elements. Consequently, the features of a modeling language that maximize these quality attributes should not be sacrificed in favor of issues such as decidabil-

ity and computational efficiency for automatic reasoning (which are design concerns, not conceptual ones).

Secondly, as a conceptual model of reference, an ontology can then be used to produce several different alternative implementations in different codification languages (e.g., OWL DL, RDF, F-Logic, DLR_{US}, Haskell [1], Relational Database languages, CASL, among many others). The choice of each of these languages should be made to favor a specific set of non-functional requirements. Moreover, within the solution space defined by these codification languages, we have a multitude of choices regarding, for instance, decidability, completeness, computational complexity, reasoning paradigm (e.g., closed versus open world, adoption of a unique name assumption or not), expressivity (e.g., regarding the need for representing modal constraints, higher-order types, relations of a higher arity), verification of finite satisfiability, among many others. The point here is that the choice of a particular codification language can only be justified as a *design choice*. To put it baldly, the question is not whether, for instance, OWL is good or not for representing ontologies. The question is whether OWL is *justifiable as an adequate design choice in a specific design scenario*. At this point, I would like to echo the historical report of Janis Bubenko regarding an analogous discussion taking place in the conceptual modeling community in the 70's between supporters of *Conceptual Data Models* (e.g. ER diagrams) and those of the *Relational Data Model* [3]. As summarized by Bubenko, "*[t]oday the battle is settled: conceptual data models are generally used as high-level problem oriented descriptions. Relational models are seen as implementation oriented descriptions*".

To complete the view outlined above, between the phases of *Ontology Conceptual Modeling* and *Ontology Codification*, we need a phase of *Ontology Design* that provides methodological supports for: (i) systematic exploration of the solution space, hence, supporting reasoning with possible choices of codification technology as well as their ability to satisfy a specific set of non-functional requirements; (ii) mapping from the conceptual to a selected codification language with the goal of preserving as much as possible the real-world semantics of the original model while still attempting at *satisficing* the non-functional requirements at hand.

[1] See the paper "Modeling vs. Encoding for the Semantic Web" by Werner Kuhn in this inaugural issue.

This rationale has received much attention in the context of the OMG's MDA (Model-Driven Architecture) initiative which aims at improving model-reuse via separation of concerns. In that scenario, due to recognition of the elevated costs of producing high-quality domain representations, there is a clear understanding that these representations should be independent of computational concerns (hence the term *Computational Independent Model*). The idea is to prevent these models from becoming deprecated due to changes which are purely related to technological choices.

Finally, there is an important additional aspect which I would like to draw attention to and which directly comes to mind when thinking about reference models. The role of a domain reference model is to provide a *frame of reference*, i.e., to serve as a conceptual tool for mastering the complexity and harmonizing possibly heterogeneous viewpoints and terminologies regarding a domain. Such a reference model is commonly used as a frame for producing implementations (including ones with automated reasoning). However, it can also be used in an off-line manner in a multitude of other meaning negotiation tasks. In other words, a Reference Conceptual Model has a value in itself, independent of the implementations that can be derived from it.

3. Revisiting the ontological level

In this section, I focus on ontologies as Reference Conceptual Models. Given the nature of possible applications of an ontology in this sense, a conceptual modeling language for producing high-quality ontologies should be able to: (i) allow the conceptual modelers and domain experts to be explicit regarding their ontological commitments, which in turn enables them to expose subtle distinctions between models to be integrated and to minimize the chances of running into a *False Agreement Problem* [5]; (ii) support the user in justifying their modeling choices and providing a sound design rationale for choosing how the elements in the universe of discourse should be modeled in terms of language elements.

Regarding (i), in order for a conceptual modeling language to be able to produce truthful specifications of a domain conceptualization, it must offer modeling primitives which are able to capture the nuances and subtleties involving the very *essence* of the elements constituting that domain. As recognized in the Harvard Business Review report of October 2001:

"*one of the main reasons that so many online market makers have foundered [is that]the transactions they had viewed as simple and routine actually involved many subtle distinctions in terminology and meaning*"[2]. Corroborating this point, [7] demonstrates a number of semantic interoperability problems that can arise when integrating even simple lightweight ontologies. Additionally, [6] elaborates on cases of semantic overload involving concepts which are central to a domain (e.g., the concept of Petroleum for a Petroleum company!) that pass undetected even within the same organization. In both these cases, the problems are related to the inability of the modeling approach used in giving support for establishing precise meaning agreements.

Regarding (ii), I would like to revisit a classification put forth by Nicola Guarino is his seminal paper "*The Ontological Level*" [4]. As discussed there, *Logical-Level* languages (e.g., FOL) are "flat" in the sense that they put all predicative terms (e.g., Apple and Red) in the same footing; *Epistemological-Level* languages (e.g., UML, ER, OWL) provide ways for elaborating structures which differentiate these terms. For instance, in UML: (a) we can define a Class of Apples with an attribute *color=red*; or (b) we can define a Class of Red with an attribute *type=apple*. What an Epistemological-Level language does not give us is a precise criterion for explaining why structure (a) is better than (b). As discussed in that paper, structuring decisions, such as this one, should not be the result from heuristic considerations, but they should rather reflect important *ontological distinctions* that should be motivated and explained. For instance, in this case, the choice between these modeling alternatives reflects a choice between sorts of object types of completely different nature, and which entails radically different consequences both in theoretical and practical terms [7].

In summary, in order to meet the *desiderata* in (i) and (ii), we need the support of a system of truly *Ontological Categories.* This system should comprise a body of formal (i.e., domain independent) theories postulating ontological distinctions, as well as a rich axiomatization prescribing how these distinctions can be related. Moreover, this system of categories should be embedded in a language system, i.e., we need a modeling language with a set of constructs that honor these ontological distinctions.

A language designed with the specific purpose of addressing these issues for the case of *Structural Conceptual Models* is the version of UML 2.0 proposed in [7] and latter dubbed OntoUML. This language reflects a system of categories postulated by an underlying reference ontology of endurants (objects), based on a number of theories from Formal Ontology, Philosophical Logics, Philosophy of Language, Linguistics and Cognitive Psychology. As a result, the language offers a rich set of primitives capturing fine-grained distinctions among, for example: (i) part-whole relations; (ii) object types, (iii) properties; (iv) forms of ontological dependence, etc. In the next section, we refer to OntoUML to illustrate some of the points discussed there.

[2] I thank Nicola Guarino for bringing this text to my attention.

4. The humble ontologist

In his ACM Turing Award Lecture entitled "*The Humble Programmer*" [11], E.W. Dijkstra discusses the sheer complexity one has to deal with when programming large computer systems. His article represented an open call for an acknowledgement of the complexity at hand and for the need of more sophisticated techniques to master this complexity.

I believe that we are now in an analogous situation with respect to conceptual modeling, in general, and ontology construction, in particular. We will experience an increasing demand for building and using reference ontologies in subject domains in reality for which sophisticated ontological distinctions are demanded. As discussed in the previous section, we need ontologically sound representation languages. However, for the sake of scalability and separation of concerns, the ontology engineering practitioner should not be required to deal with all the intricacies of the theories underlying the language. In other words, on the one hand, we need to offer to the working ontologist, theories and modeling distinctions as expressive as possible. On the other hand, we need as much as possible to shield this practitioner from the complexity of these conceptual tools.

In the sequel, I discuss three of the possible kinds of tools that can be used to master the inherent complexity of this process.

4.1. Ontological design patterns

In software engineering, design patterns have become a way to capture in a standard form a solution to a recurrent problem. As recognized by the community of pattern languages, patterns are actually not only a means for reusing expert's knowledge. More than that, they define a language to talk about design,

having become part of the area's jargon. In other words, people exchange patterns as signs with specific and shared semantics within that community as opposed to having to repeatedly explain the situation that motivated their creation.

In ontological engineering, there are obvious opportunities to take advantage of a similar approach. Due to space limitations, I will comment here on just two classes of these patterns, namely, *modeling patterns* and *transformation patterns.* For an example of *analysis patterns* proposed to identify the scope of transitivity of parthood, one can refer to [7].

Firstly, we need patterns that can be used to represent domain-independent solutions to modeling problems that can be manifested in several domains. These patterns shall be motivated by formal ontological reasons and (also because of that) I predict that they will hardly be identified in an approach that neglects formal ontological categories. Examples of patterns in this class have been proposed, for instance in [7], to address modeling problems such as: (i) role modeling with disjoint admissible types; (ii) modeling of material relations and their truth-makers (relational properties); (iii) separating entities from their constitutions; (iv) representation of qualities with alternative associated quality spaces; (v) harmonizing alternative notions of roles, among others. It is important to highlight that, in the case of all these patterns, the modeling solutions proposed result from an ontological analysis of the problems at hand in terms of the categories and theories of an underlying foundational ontology.

More than collecting a number of useful Modeling Patterns, we should pursue the construction of ontology modeling languages which are *pattern languages*. OntoUML is a language which has such a feature to a large extent. In that language, it is common that the choice of modeling a domain element using a particular construct causes a whole pattern to be manifested [7]. This opens the possibility for an editor that supports the user in modeling with elements of higher-granularity and cohesion, i.e., instead of simply using isolated primitives such as classes, associations and attributes, the models would be constructed with pattern blocks instantiating formal relations from a foundational theory.

Secondly, we need *transformation patterns* capturing standard solutions to problems of mapping ontologically rich models to languages which are less expressive or have specific characteristics. A number of examples of patterns in this category which aim at supplanting the limitations of OWL have been collected in ODP Portal[3]. However, other patterns in this class have also been proposed considering radically different paradigms. For instance, [10] proposes a pattern which captures a solution to the problem of preserving the basic semantics of mereological relations in traditional Object-Oriented implementations. In fact, there are many opportunities for employing standard OO Patterns such as Composite, Delegation, State and Observer to propose standard solutions for implementing ontology-related issues such as transitive propagation of properties, multiple and anti-rigid classification, and existential dependency. Having the source model represented in an ontologically rich language provides a direct guidance for when and how to apply these patterns.

4.2. Model-driven editors

As previously discussed, the OntoUML metamodel contains: (i) elements that represent ontological distinctions prescribed by an underlying foundational ontology; (ii) constrains that govern the possible relations that can be established between these elements. Let us illustrate these points by using the distinction between the object type *Kind* and *Roles*. In a simplified view we can state that: a Kind is a type that congregates all the essential properties of its instances and, for that reason, all instances of a Kind cannot cease to instantiate it without ceasing to exist; a Role, in contrast, represents a number of properties that instances of a Kind have contingently and in a relational context. A stereotypical example of this distinction can be appreciated when contrasting the Kind Person and the Role Student [7]. Regarding (i), OntoUML incorporates constructs that represent both of these ontological categories. Regarding (ii), the metamodel embeds constraints such as: a role must be a subtype of exactly one ultimate Kind; a role cannot be a supertype of a Kind.

Because these distinctions and constraints are explicitly and declaratively defined in the metamodel, they can be directly implemented using metamodeling architectures such as the OMG's MOF (Meta Object Facility). Following this strategy, [1] reports on an implementation of OntoUML graphical editor by employing a number of basic Eclipse-based frameworks such as the ECore (for metamodeling purposes), MDT (for the purpose of having automatic verification of OCL constraints) and GMF (for the purpose of building a model-based graphical interface). An interesting aspect of this strategy is that, by

[3] http://ontologydesignpatterns.org/

incorporating ontological and semantic constraints in the metamodel (i.e., the abstract syntax) of the language, it mimics a process which also takes place in natural language.

As an example of the latter point, take the two sentences: *(i) (exactly) five mice were in the kitchen last night; (ii) the mouse which has eaten the cheese has been in turn eaten by the cat.* If we have the patterns *(exactly) five X...* and *the Y which is Z...*, only the substitution of X,Y,Z by common nouns will produce sentences which are grammatical. To see that, one can try the replacement by the adjective Red in the sentence *(i): (exactly) five red were in the kitchen last night.* Now, the reason for why this is the case is an ontological one [7]. The interesting aspect here is that the competent user of this natural language does not need to know that! In other words, one can (as most language speakers do) abstract from the ontological reasons behind a grammatical constraint.

We should pursue the same ideal in ontology conceptual modeling languages. For example, one does not need to be fully aware of the reason why *a Role cannot be a supertype of Kind*. Actually, following the strategy adopted for the OntoUML tool editor, the user does not even have to be aware of the syntactic rule either: if one tries to produce a model violating this rule, this will be identified by the embedded OCL constraint checker of the tool, and the modeler will be promptly notified about the forbidden action.

Another important advantage of having an ontology language with an explicitly defined metamodel is the possibility of implementing multiple transformations from an ontology conceptual model to different codification schemes. Again, metamodel transformation is a widespread practice by the followers of OMG's MDA initiative (refer to Section 2). In this spirit, once we have a transformation model defined between, for example, the OntoUML and the OWL metamodels, every model in the first language can be automatically transformed into a specification in the second one. For example, [2] implements a transformation model from OntoUML to the constraint language Alloy by using the ATL language (a popular implementation of the OMG's QVT model). This mapping enables the creation of the model simulator discussed in the next section.

4.3. Model simulators

Having a modeling language whose metamodel incorporates the ontological constraints of a foundational theory directly eliminates the representation of *ontologically non-admissible states of affair*. However, it cannot guarantee that only *intended states of affair* are represented by the domain model at hand [21]. This is because the admissibility of domain-specific states of affair is a matter of factual knowledge (regarding the world being the way it happens to be), not a matter of *consistent possibility*.

To illustrate this point, suppose a medical domain ontology representing the procedure of a transplant. In this domain, we have concepts such as Person, Transplant Surgeon, Transplant, Transplanted Organ, Organ Donor, Organ Donee, etc. A *transplant conceptual model* which places Organ Donor (role) as a supertype of Person (kind), or one that represents the possibility of a Transplant (event) without participants clearly violates ontological rules. However, these two cases can be easily detected and proscribed by an editor such as the one discussed in the Section 4.2. The issue here is that in this case one can still produce a model which does not violate any of these rules but which still admits unintended states of affair as valid instances. One example is a state of affair in which the Donor, the Donee and the Transplant Surgeon are one and the same Person. Please note that this state of affair is only considered inadmissible due to domain-specific knowledge of social and natural laws. Consequently, it cannot be ruled out *a priori* by a domain independent system of categories.

Guaranteeing the exclusion of unintended states of affair without a computational support is a practically impossible task for any relevant domain. In particular, since many fundamental ontological distinctions are modal in nature [7], in order to validate a model, one would have to take into consideration the possible valid instances of that model in all possible worlds.

In [2], an extension to the OntoUML editor was presented that offers a contribution to this problem. On the one hand, it aims at proving the satisfiability of a given ontology by presenting a valid instance (logical model) of that ontology. On the other hand, it attempts to exhaustively generate instances of the ontology in a branching-time temporal structure, thus, serving as a visual simulator for the possible dynamics of entity creation, classification, association and destruction. The snapshots in this world structure confront a modeler with states of affair that are deemed admissible by the ontology's current axiomatization. This enables modelers to detect unintended states of affair and to take the proper measures to rectify the model. The assumption is that the example world structures support a modeler in this validation process, especially since it reveals how states of af-

fair change in time and how they may eventually evolve in counterfactual scenarios.

5. Final considerations

As we argue throughout this paper, one of the key aspects of the Semantic vision is Semantic Interoperability. If conceptual modeling is about "*the construction of models of reality that promote a common understanding of that reality among their human users*" [11], then successful semantic interoperability is about harmonizing different viewpoints reflected in different conceptualizations of that same reality. In any case, reality cannot be left out of the loop. As a consequence, an approach which neglects the role of ontologies as reference conceptual models cannot meet the requirements for semantic interoperability.

The Semantic Web vision puts forth an undoubtedly inspiring challenge. Moreover, it brought us a number of interesting results from serious and talented researchers working on the field. However, in that context, there has been an unbalanced focus on developing representation techniques to support efficient reasoning. In contrast, the very task of domain representation, i.e., the task of constructing principled conceptual structures that represent with truthfulness and clarity the underlying domain, has been left to the user. Another negative aspect that must be brought to attention regarding the Semantic Web is that, due to its popularity, the hype wave it has generated also brought us a lot of *noise*. I believe this seriously harmed the establishment of a clear view of the sheer complexity involved in the problem at hand.

Since ontology engineering is a young discipline, there are many lessons to be learned from closely related areas such as Software Engineering, Information Systems and Databases. One of these is that the quality of any implementation artifact based on a model is ultimately bound by the quality of that model. Another one is that the area must properly define its problem and solution spaces as well as to bridge them effectively. The latter bears strong ties with the topic of *Ontology Education*, a subject which has gained much interest in the international community recently[4]. I am afraid until we have a minimally agreed curriculum or body of knowledge[5] to guide the education of ontologists, many of our discussions will still be carried out by engineers that lack true *ontological* knowledge as well as formal ontologists that lack basic industrial experience and sensitivity to the need of engineering tools.

This paper elaborates on a research program that addresses exactly the conceptual modeling phase of Ontology Engineering, focusing on the development of foundational theories, modeling languages and methods, design patterns and supporting computational environments that aim at supporting the construction of ontologies as reference models. The paper reflects on a limited list of items and there are many other fundamentally important issues regarding ontology engineering which I did not deal with here.

There are still many challenges to be met before we can have a mature discipline of Ontology Engineering. The road ahead of us is both challenging and exciting and, once more paraphrasing Dijkstra, we should do a much better ontology engineering job in the future, "*provided that we approach the task with a full appreciation of its tremendous difficulty*".

[4] See the discussions on the "Ontology Summit 2010: Creating the Ontologists of the Future" (http://ontolog.cim3.net/cgi-bin/wiki.pl?OntologySummit2010).

[5] For a contrast, see the IEEE Guide to the Software Engineering Body of Knowledge (http://www.swebok.org/).

Acknowledgements

This research is supported by FAPES (Grant# 45444080/09) and CNPq (Grant# 481906/2009-6).

References

[1] Benevides, A.B., Guizzardi, G., A Model-Based Tool for Conceptual Modeling and Domain Ontology Engineering in OntoUML, *Proceedings of the 11th ICEIS, Milan*, 2009.

[2] Benevides, A.B., Guizzardi, G., Braga, B.F.B., Almeida, J.P.A., Assessing Modal Aspects of OntoUML Conceptual Models in Alloy, *Proc. of the 1st ETheCom, Gramado*, 2009.

[3] Bubenko Jr., J.A., From Information Algebra to Enterprise Modelling and Ontologies – a Historical Perspective on Modelling for Information Systems, *Conceptual Modeling in Information Systems Engineering*, Springer-Verlag, 2007.

[4] Dijkstra, E.W., The Humble Programmer, *Communications of the ACM*, **15**:10, October 1972.

[5] Guarino, N., Formal Ontology and Information Systems, *Proceedings of the 1st FOIS, Trento, Italy, June 6–8*. IOS Press, Amsterdam: pp. 3–15, 1998.

[6] Guarino, N., The Ontological Level, In R. Casati, B. Smith and G. White (eds.), *Philosophy and the Cognitive Science*. Holder-Pivhler-Tempsky, Vienna: pp. 443–456, 1994.

[7] Guizzardi, G., *Ontological Foundations for Structural Conceptual Models*, Telematica Instituut Fundamental Research Series No. 15, ISBN 90-75176-81-3, The Netherlands, 2005.

[8] Guizzardi, G., Halpin, T., Ontological Foundations for Conceptual Modeling. *Applied Ontology*, v. 3, p. 91–110, 2008.

[9] Guizzardi, G., Lopes, M., Baião, F., Falbo, R., On the impor tance of truly ontological representation languages, *International Journal of Information Systems Modeling and Design (IJISMD)*, IGI-Global, 2010.

[10] Guizzardi, G., Falbo, R.A., Pereira Filho, J.G., Using objects and patterns to implement domain ontologies, *Journal of the Brazilian Computer Society*, ISSN 0104-6500, **8**:1, July 2002.

[11] Mylopoulos, J., Conceptual modeling and Telos. In P. Loucopoulos & R. Zicari (eds.), *Conceptual Modeling, Databases, and CASE* (Chapter 2, pp. 49–68). Wiley, 1992.

Semantic Web 1 (2010) 45–52
DOI 10.3233/SW-2010-0016
IOS Press

Can we ever catch up with the Web?

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA
Solicited review(s): Martin Raubal, University of California, Santa Barbara, USA; Andreas Hotho, University of Würzburg, Germany
Open review(s): Pascal Hitzler, Wright State University, Dayton, Ohio, USA

"The truth is rarely pure and never simple." – Oscar Wilde

Axel Polleres [a,*], Aidan Hogan [a], Andreas Harth [b] and Stefan Decker [a]
[a] *Digital Enterprise Research Institute – DERI, National University of Ireland, Galway, Ireland*
E-mail: aidan.hogan@deri.org, stefan.decker@deri.org
[b] *Institut of Applied Informatics and Formal Description Methods – AFIB, Karlsruhe Institute of Technology, Germany*
E-mail: harth@kit.edu

Abstract. The Semantic Web is about to grow up. By efforts such as the Linking Open Data initiative, we finally find ourselves at the edge of a Web of Data becoming reality. Standards such as OWL 2, RIF and SPARQL 1.1 shall allow us to reason with and ask complex structured queries on this data, but still they do not play together smoothly and robustly enough to cope with huge amounts of noisy Web data. In this paper, we discuss open challenges relating to querying and reasoning with Web data and raise the question: can the burgeoning Web of Data ever catch up with the now ubiquitous HTML Web?

Keywords: Web of data, reasoning, querying, rules, ontologies, RDF, OWL2, SPARQL, RIF

Introduction

We finally find ourselves at the tipping point for a Web of Data [45]: through efforts such as the Linking Open Data initiative [6,8], resources like Wikipedia, movie and music databases, news archives, online citation indexes, social networks, product catalogues and reviews, etc., are becoming available in structured form as RDF, using common ontologies mostly in the form of lightweight vocabularies like FOAF [11], SIOC [9], YAGO [52], etc.

In an idealised world, Linked Data promises to expose the knowledge items published on the Web as one big graph of networked knowledge. Leaving all implied problems aside, such an idealised view means:

- Besides publishing or dynamically generating HTML, everybody exposes their knowledge directly as RDF/XML [5], embedded in HTML as RDFa [1], or even makes their database accessible behind SPARQL endpoints.
- HTTP URIs are used as names and are dereferenceable. Data publishers use the same distinct URIs to reference the entities they talk about, be it individual instances, or classes and properties: that is, data is linked.
- Where different properties and classes are used, relations between those are declared in some ontology: that is, also ontologies are linked.

Emerging standards such as OWL 2 [29], RIF [10] and SPARQL 1.1 [21] subsequently allow for reasoning and elaborate queries on the resulting huge RDF graph, but still this novel Web of Data is brittle.

The alert reader will recognise that particularly the first two items in the above list just paraphrase the original Linked Data principles [6], but we call these "idealised" since in fact the current status of the vast ma-

*Corresponding author. E-mail: axel.polleres@deri.org.

jority of datasets in the Linking Open Data "cloud"[1] is still far from this ideal. For instance, reuse of identifiers across datasets is still sparse in Linked Data; in the absence of a centralised "URI mint" – which in any case would be against the ad-hoc nature of the Web – publishers continue to use locally defined URIs: in fact, Linked Data principles could be seen as encouraging such practice where publishers mint URIs which dereference to their local description of the referent resource. Services like Sig.ma [53] provide initial entity-search facilities to help here, but still the usage of such services can't be enforced in an open structure such as the Web; although co-referent identifiers are sometimes subsequently identified across sources using `owl:sameAs`, this is not sufficient and more fine-grained notions of similarity or contextualised equality may be necessary (as argued in [19]).

Additionally, the chaotic Web will not provide one clean graph, but noisy and conflicting information will be published, meaning that the formal semantics of OWL or RIF have to be applied with care to make sense out of this data – in fact, it may be more accurate to think of Linked Data as a collection of interlinked graphs, each with its own contextual (and possibly fuzzy) interpretation of truth than the simplified view of one global, homogeneous knowledge base: see also [28] in this issue for more discussion.

Thus, rather than operating on an ideally structured, global knowlege base, we have to deal with Linked Data as it is currently published, where we face the following three main challenges. On the one hand, (i) we still have *too little* Linked Data out there to answer complex queries that extend beyond the coverage of single datasets (Section 1). Also, (ii) Linked Data is of largely varying *quality*: publishing errors and (deliberate or accidental) inconsistencies arise naturally in an open environment such as the Web (Section 2). On the other hand, (iii) we may have *too much* data to deal with efficiently given current technologies and standards (Section 3). In this paper, we will discuss these three challenges, along with current approaches and possible solutions. We conclude with a deliberately speculative outlook on what might be next – i.e., challenges on the horizon – charting possible evolutions on the Web of Data.

[1] cf. http://richard.cyganiak.de/2007/10/lod/

1. Too little linked data

Common Semantic Web enthusiasts are quickly humbled when they try to answer basic queries over the Web of Data. A lack of both data and links becomes especially evident when one wants to pose queries that combine information from several sources. Imagine a query such as "give me information about bands my friends recently listened to or blogged/twittered about": it is likely that the information you need to answer that query is on the Web, but is (i) not available as RDF; (ii) only partially available as RDF; (iii) in RDF, but not sufficiently linked.[2]

Although the Web of Data is growing and covering a broader range of topics, it is unclear whether data published in structured formats such as RDF will ever be able to compete with prose documents in that regard. Clearly, expressing information in prose is highly flexible and allows publishers to easily specify 'niche' or 'nuanced' claims about the world such that is easily understandable by a speaker of the language. However – and not denying the inherent flexibility of RDF – it is certainly more difficult to express such claims in RDF and in a manner such that machines can appropriately exploit the resulting data.

For example – and again although the coverage of vocabularies is growing – the necessary terms may not yet exist, may not perfectly fit the meaning intended by a given publisher, or may not be easy to find.[3] Thus, a simple prose claim such as "Andreas was disappointed by the 'James Blunt' gig he recently attended" may not be possible with the available vocabulary terms, and modelling such a claim using RDF(s)/OWL may require complex modeling, and thus experience (see [28] for a more detailed example of "awkward triplification"). If one invents a novel vocabulary for such a claim, then ideally other publishers with similar claims could re-use the terms and follow precedent: however, encouraging broad re-use of vocabulary terms currently requires a large community-driven effort, as has been demonstrated by the Herculean efforts in and around SIOC [9] and FOAF [11]. Both of these examples have shown that enabling adoption of an ontol-

[2] Inadvertently, we also raise privacy issues which Semantic Web technologies are not well poised to address; if we continue to shirk privacy issues, we may risk losing potential early adopters and applications involving personal or sensitive data.

[3] For discussion of an approach to better structure the development and re-use of vocabularies, see also [37] in this issue. Whether "Coordinated Collaboration for Vocabulary Creation" as promoted in this approach is feasible at Web scale has yet to be proven.

ogy requires more in terms of community effort (incorporating feedback from users, building tools and exporters, spreading the word) than in terms of technical design: both ontologies consist of only a minimalistic bunch of classes, properties and axioms.

Despite the Linked Data community's enthusiasm, the vast majority of day-to-day Web developers still ignores semantic technologies. Thus, we will have to pick developers up where they are, incorporating RDF in widely used tools in an unobtrusive, easy to learn manner. Starting points in this direction exist: Triplify [3], or RDF in Drupal [12]. Yet, more are needed to "catch up" with the speed of growth and diversity of the HTML Web.

Again, more vocabularies and terms are needed – reciprocally, more infrastructure and support is required to lower the barriers-to-entry for creating agreed-upon vocabularies. Efforts such as the Neologism tools [4] for vocabulary creation and maintenance, VoCamp meetings[4] to create ad-hoc vocabularies, or ontology term search services such as the one sketched in [12], are trying to address this need.

Finally, on the Web of Data, there is too little inter-dataset linkage on the instance level to allow for elaborate queries or machine-learning applications [7]. Most current exporters use disparate identifiers (usually for reasons of dereferenceability) for the same entities, say DBPedia (e.g. http://dblp.l3s.de/d2r/page/authors/Tim_Berners-Lee) vs. DBLP (http://www4.wiwiss.fu-berlin.de/dblp/page/person/100007) vs. FOAF profiles (http://www.w3.org/People/Berners-Lee/card#i); even though explicit `owl:sameAs` links are appearing in more and more abundance – and even leaving aside the problems with respect to how they are currently used – they alone are still not enough. Tackling the lack of such links, Silk [56] offers a publishing-centric means of creating links – possibly `owl:sameAs` – between related datasets.

From a data-consumer perspective, OWL reasoning can provide a richer set of `owl:sameAs` relations – e.g., by exploiting (inverse) functional properties such as `foaf:homepage` – to align identifiers [31]. However, such approaches still run into problems when fired on real Web data because (i) suitable information on which to align may not exist, and (ii) erroneous information leads to aligning too much [31–33]. Thus, probabilistic, fuzzy, or statistical approaches – cf. the preliminary results of [35] – may prove more promising (or complementary) for deriving same-as links between datasets.

But linkage does not end at the instance level; as certain vocabularies become established, links between vocabularies by "bridging ontologies" or mappings may become necessary to link ontologies. As discussed in [39], users may wish to query over information aggregated from multiple sources using disparate schemata – they propose an upper-level ontology as a possible solution, though this in our opinion would be in direct conflict with the ad-hoc bottom–up approach at the very heart of Linked Data's success. As OWL and RDFS alone do not provide the means to describe complex mappings, one may envision using SPARQL as a mapping language [47] or W3C's new Rule Interchange Format (RIF) [10], yet no best practice is agreed as of yet to publish and share mappings on the Web, nor how to process them at Web scale. Efforts such as the Ontology Alignment Evaluation Initiative[5] – and more generally, the well-established Ontology Matching research community behind it [18] – are just starting to discover Linked Data as a field of application, and have yet to prove that their methods apply over the loose conglomerate of lightweight ontologies found online.

Certainly more plumbing is needed, but a much wider range of data would open up if additional means of "mappings" to/from non-RDF data – such as relational or XML sources which serve as the backbone of the vast majority of Web-based information systems – became available. Efforts, such as D2RQ – linking to relational Databases and forming one of the starting points of W3C's recently started RDB2RDF working group – or XSPARQL [46] – a combined query language which we proposed to ease transformations from and to XML by merging XQuery and SPARQL – and similar efforts should allow the Semantic Web to interact with existing sources of structured and semi-structured data.

2. Linked data quality

With respect to the RDF currently published on the Web – mostly exports of legacy structured or semi-structured data – there are still many issues which inhibit consumer applications from fully exploiting that data. Firstly, although RDF theoretically offers excel-

[4] http://vocamp.org/

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

lent prospects for automatic data integration assuming re-use of identifiers and strong inter-dataset linkage, such an assumption currently only weakly holds (as already outlined in the previous Section). Secondly, publishers are prone to making errors which impinge on the quality of the resulting data.

In [32], we provided discussion and illustrative statistics relating to the current quality of Linked Data publishing: besides HTTP-level issues relating to content-type reporting and dereferenceability of URIs, we reported that applying reasoning over the Web of Data can be problematic. For instance, undefined classes and properties – those without a formal RDFS or OWL description – are commonly instantiated in Web data. Similarly, for example, datatype clashes – e.g., lexically invalid datatype literals – are common under D-entailment [26]. Finally, we discovered various examples of inconsistencies relating to instance membership of disjoint classes.

Note that in the future, as more data gets published, we will probably have to expect a lot more inconsistencies, be they accidental or deliberate in nature. Accidental inconsistencies often arise when data publishers are ignorant of or mis-interpret certain ontology terms. For example, data publishers may use the `foaf:img` property to relate an arbitrary resource with an image, missing the fact that the domain of `foaf:img` is `foaf:Person`; performing inference over such data, a reasoner infers that the resource is of type `foaf:Person`, which could cause an inconsistency if the resource's explicit class and `foaf:Person` are defined as disjoint. Inconsistencies can also occur due to incompatible naming across sources: for example, we found two Linked Data exporters which used LastFM profile page URIs to identify users and documents respectively, taken together resulting in inconsistencies [32]. Deliberate inconsistencies may also occur, expressing genuine disagreement amongst data publishers: for example, imagine ontologies by different providers that define *vegetables disjoint from fruit*, *tomatoes are fruits* and *tomatoes are vegetables*, which, when taken in combination, result in an inconsistent knowledge base.

We can broadly distinguish four strategies reported in the literature for dealing with inconsistencies. First, inconsistencies can be simply ignored: RDFS/OWL (2 RL) rule-based reasoning approaches can detect some inconsistencies, but will not suffer the explosive consequences of *ex contradictione quodlibet*.

Second, the Web community at large takes care of resolving the inconsistencies in a social discourse: for example by working with data publishers to resolve inconsistencies that arise by accident. An example for such an initiative is the Pedantic Web group[6], which comprises of loosely organised volunteers that are concerned with erroneous data on the Web – the group points out mistakes to data publishers and actively supports them to fix the issues.

Third, algorithms can be used to resolve inconsistencies. For example, model-based revision operators can be used to resolve inconsistencies by removing axioms that cause the inconsistency [48]. Approaches advocating para-consistent reasoning on the Web (cf. for instance [36,42]) could also help to draw valid inferences even in the face of inconsistencies. Although such methods work on small ontologies, adapting these methods to scale to the Web is an open area for research. These methods attempt to choose a consistent model from inconsistent data, e.g., based on distance metrics or probability functions. Alternatively, ranking [16,23,30] of statements and inferences may be used to weigh contradicting inferences against each other.

Fourth, in the case of deliberate inconsistencies, users might need to decide which point of view to take for contentious topics – deliberate disagreements are not so much an issue so far, but this may become a bigger issue as soon as data publishers use their logical understanding of OWL & Co to express different opinions. Such different points of view, as found over and over in current Web content, and although expressible formally in OWL, still miss an agreed way of being handled in terms of standards. How to distinguish deliberate from accidental inconsistencies is also an open question.

Returning more generally to data quality – and no matter what solutions are proposed – the Web of Data will always contain noise and inconsistencies; thus, tracking the provenance of data is hugely important. For example, SPARQL includes the notion of named graphs (butt we are still missing a formal framework for reasoning over those named graphs). Recent research has looked at including consideration of the source of data in algorithms for ranking (cf. [16,23,30]) and reasoning (cf. [14,33]) over Linked Data. A generic framework for querying and reasoning over annotations (including, e.g., provenance or trust values) of RDF [41,50] may also serve as a useful starting point.

[6] http://pedantic-web.org/

Data quality could also be improved through usage: i.e., leveraging explicit or implicit feedback loops in systems (search engines, browsers, etc.) operating over Linked Data to determine data quality or rely on end users to fix issues.

Again, we refer the interested reader to [32] for a more detailed discussion of noise and inconsistency in current Linked Data, including proposals of solutions.

3. Too much data

In contrast to Section 1, many challenges relating to scalability arise from the increasing volume of RDF data being published on the Web. First of all, consumers of RDF need to be able to locate and interact with structured data of interest. Linked Data principles encourage the use of dereferenceable URIs – URIs which, upon lookup, return some interesting data about the referent. However, relying solely on simple dereferencing to locate data requires that publishers use dereferenceable URIs and that consumers know the URI(s) of the entity of interest. Also, such an approach mitigates the data-integration potential of RDF, ignoring the related and relevant contribution of remote publishers.

Thus, data warehouse approaches (take for example SWSE [34] or Sindice [43]) which provide mechanisms for locating and interacting with structured information are necessary for many applications. Data warehouses can offer lookups for relevant sources of structured information – somewhat emulating current HTML-centric Web search engines – or can also allow users to pose queries and tasks over a locally indexed version of the Web of Data. Perhaps the most obvious challenge for such systems is scalable storage of data and query-processing: for example, supporting arbitrary SPARQL queries at scale quickly becomes both computationally [44] and economically cost prohibitive. Scalable triple/quad stores are now appearing in the literature, some of which are based on native or IR-based RDF storage solutions (cf. [15,24]) and some which use underlying databases (cf. [17,20]); importantly, each system can only demonstrate scalability and efficiency for a subset of SPARQL.

Besides storage and query-processing, such systems often incorporate data curation and analysis components to improve precision, recall and/or usability of the systems. Such curation often involves scalable techniques inspired by the Semantic Web standards, as well as more traditional Information Retrieval techniques including: (i) data integration: e.g., applying entity consolidation to canonicalise co-referent identifiers and thus merge the contribution of independent publishers for a given entity (cf. [31,35]); (ii) reasoning: inferring new knowledge given the semantics of terms described in OWL/RDFS (cf. [14,33]); (iii) ranking: scoring the importance and relevance of given data artefacts for prioritisation of results (cf. [16,23,30]). Although such data-warehouses can borrow from existing information retrieval techniques known to scale – such as crawling, ranking and indexing techniques – the unique nature of the Web of Data mandates deviation from well-understood approaches, and also the additional challenges relating to entity consolidation, reasoning and querying.

The current RDF publishing standards do not lend themselves naturally to scalable processing. For example, OWL (2) Full reasoning is well-known to be undecidable, and OWL (2) DL is not naturally suited to reasoning over the inconsistent, noisy and potentially massive Web of Data; a starting point in this direction is to cautiously narrow down inferences to "safe terrains" by deliberately incomplete approaches that avoid non-authoritative statements during inference [14,33] – again in [28], Hitzler et al. argue that soundness and completeness wrt. the formal semantics are often infeasible goals for practical reasoning systems, and that precision-/recall-type measures should be adopted as more realistic evaluation metrics.

For all such scalability challenges, distribution plays an important role. Although distribution is not, per-se, a 'magic bullet' – a task that is not scalable on one machine will likely not scale either over multiple – appropriate parallel execution of data processing, indexing, and query processing allows for faster indexing of source data and faster responses from the system. Distributed indexing [17,20,24], querying [34,51] and reasoning [14,34,54,55,57] is currently being investigated in various incomplete/approximative approaches, but still not in a manner that can handle dynamic data, or live queries that retrieve data directly from the sources [25]. When going as far as combining dynamic data with dynamic inferences, that is, querying the data under dynamically changing inference regimes and with different (versions of) ontologies, even rule-based approaches can so far only be handled at relatively small scale [38]; distribution of such fully dynamic reasoning and querying has, to our knowledge, not yet been investigated.

Closely related to distribution is query federation: that is, distributed querying over closed endpoints,

each of which provides a query interface and potentially a self-description of its capabilities/dataset; due to the schema-less nature of the Semantic Web, the task of query federation – which is highly intractable without restrictions for the traditional relational setting already (cf. for instance [27,40]) – becomes even harder. We currently see only few works going in this direction [49], none of which yet demonstrate scale suitable for the Web.

Indeed, predominant data warehousing techniques have two inherent and significant disadvantages: (i) some segment of the data indexed must necessarily become stale; and (ii) privacy becomes an issue as such warehouses take control of data – and how it is used, offered, and presented to the public – away from publishers. A sweet spot between (distributed) data warehouse approaches, fully fledged query federation and live lookups has yet to be determined. As a first step in the direction of tackling (i), we are currently exploring data-summaries such as QTrees for on-demand queries over Linked Data [22].

Conclusions and speculative outlook

The Semantic Web is rapidly approaching its teens. The expectations for the Semantic Web are constantly in flux. Here we have aimed to discuss what we believe to be the most pending challenges relating to the RDF Web data that is out there now – the so called Web of Data – relating to how it can be extended, improved, interpreted and exploited. Still, one could argue that in doing so, we have been myopic by focusing on obvious challenges and directions for the Semantic Web only.

There are, of course, other streams of research within the auspices of Semantic Web research which have promising futures. Methods from Semantic Web Services – which have suffered in the past from being tackled at a conceptual level only with in fact no real services on the Web to integrate – might regain attention in another disguise as a next evolution step away from the current mostly static data sources. Newer fields, such as the emergence of sensor data in an Internet of Things, the Mobile Web, or the Smart Energy Grid, may lead to new applications and dramatic shifts in requirements for the Semantic Web – for example, the need for temporal and spatial annotations and support for highly dynamic data streams [13]. New perspectives, such as from the young Web Science discipline, may be poised to exploit RDF Web data in novel and interesting ways. A tremendous amount of data readily available to data management, machine learning and visual analytics communities might enable new insights into humans behaviour, help to meet ambitious targets for making power generation and traffic flows more efficient, lead to more transparent governments, and in general may have a similarly profound impact on our lives as the Web had. In order to get there, Linked Data and the related Semantic Web technologies seem to be the right ingredients.

However, many promises of the Semantic Web are not only alluring, but at the moment also entirely ethereal; many challenges – some of which we have discussed and sketched possible solution paths for in this paper – have yet to be overcome. Given the recent (and very non-ethereal) growth of RDF data published on the Web as Linked Data, the Semantic Web community should be fostering significantly more applied research to demonstrate what's possible on the data that's out there now.[7] We should be a little more hesitant to complain that there is "too little data" or "too much useless data" or "the data is too noisy" or "not well linked" or "too simplistic", and should be a little more resolute to get our hands dirty and demonstrate applications over this data – only by eagerly researching and demonstrating and understanding what's possible or not possible on the Web of Data that's out there now can we credibly hold an opinion on what direction the Semantic Web (in the original sense of the term) should take in the future.

Acknowledgements

The work of Axel Polleres, Aidan Hogan and Stefan Decker is supported by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2). Aidan Hogan is supported by an IRCSET Postgraduate scholarship. The work of Andreas Harth is supported by the Europen Commission under the PlanetData project.

References

[1] B. Adida, M. Birbeck, S. McCarron, and S. Pemberton. RDFa in XHTML: Syntax and Processing, Oct. 2008.

[2] S. Auer and J. Lehmann. Making the Web a data washing machine – Creating knowledge out of interlinked data. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):97–104, 2010.

[7] The interested reader may also want to have a look at a related article [2] in this issue, which poses similar challenges on dealing with Linked Data from a slightly different perspective.

[3] S. Auer, S. Dietzold, J. Lehmann, S. Hellmann, and D. Aumüller. Triplify – lightweight Linked Data publication from relational databases. In *WWW 2009*, 2009. ACM Press.
[4] C. Basca, S. Corlosquet, R. Cyganiak, S. Fernández, and T. Schandl. Neologism – Easy Vocabulary Publishing. In *4th Workshop on Scripting for the Semantic Web*, June 2008.
[5] D. Beckett and B. McBride (eds.) RDF/XML Syntax Specification (Revised), February 2004. W3C Recommendation.
[6] T. Berners-Lee. Linked Data – Design Issues, July 2006. http://www.w3.org/DesignIssues/LinkedData.html.
[7] F. Biessmann and A. Harth. Analysing dependency dynamics in Web data. In *Linked AI: AAAI Spring Symposium*, 2010.
[8] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data – the story so far. *Int.'l Journal on Semantic Web and Information Systems*, **5**(3):1–22, 2009.
[9] U. Bojārs, J.G. Breslin, D. Berrueta, D. Brickley, S. Decker, S. Fernández, C. Görn, A. Harth, T. Heath, K. Idehen, K. Kjernsmo, A. Miles, A. Passant, A. Polleres, L. Polo, and M. Sintek. SIOC Core Ontology Specification, June 2007. W3C member submission.
[10] H. Boley and M. Kifer. RIF Basic Logic Dialect, June 2010. W3C Recommendation.
[11] D. Brickley and L. Miller. FOAF Vocabulary Specification 0.97, Jan. 2010. http://xmlns.com/foaf/spec/.
[12] S. Corlosquet, R. Delbru, T. Clark, A. Polleres, and S. Decker. Produce and consume Linked Data with drupal! In *ISWC 2009*, vol. 5823 of *LNCS*, p. 763–778, Oct. 2009. Springer.
[13] S. Decker and M. Hauswirth. Enabling networked knowledge. In *12th Int.'l Workshop on Cooperative Information Agents (CIA)*, vol. 5180 of *LNCS*, p. 1–15, Sept. 2008. Springer.
[14] R. Delbru, A. Polleres, G. Tummarello, and S. Decker. Context dependent reasoning for semantic documents in Sindice. In *4th Int.'l Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS 2008)*, Karlsruhe, Germany, Oct. 2008.
[15] R. Delbru, N. Toupikov, M. Catasta and G. Tummarello. A Node Indexing Scheme for Web Entity Retrieval. In *ESWC 2010*, 2010. Springer.
[16] R. Delbru, N. Toupikov, M. Catasta, G. Tummarello and S. Decker. Hierarchical Link Analysis for Ranking Web Data. In *ESWC 2010*, 2010. Springer.
[17] O. Erling, I. Mikhailov. RDF support in the Virtuoso DBMS. In *CSSW 2007*, 2010.
[18] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer, 2007.
[19] H. Halpin and P. Hayes. When owl:sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In *Linked Data on the Web Workshop (LDOW)*, 2010.
[20] S. Harris, N. Lamb and N. Shadbolt. 4store: The Design and Implementation of a Clustered RDF Store. In *Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS)*, 2009.
[21] S. Harris and A. Seaborne (eds.) SPARQL Query Language 1.1. W3C Working Draft, Jun. 2010. http://www.w3.org/TR/sparql11-query/.
[22] A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data summaries for on-demand queries over Linked Data. In *WWW2010*, Apr. 2010. ACM Press.
[23] A. Harth, S. Kinsella and S. Decker. Using Naming Authority to Rank Data and Ontologies for Web Search. In *ISWC*, 2009. Springer.
[24] A. Harth, J. Umbrich, A. Hogan, and S. Decker. YARS2: A federated repository for querying graph structured data from the Web. In *ISWC2007*, p. 211–224, 2007. Springer.
[25] O. Hartig, C. Bizer, and J.-C. Freytag. Executing SPARQL queries over the Web of Linked Data. In *ISWC2009*, 2009. Springer.
[26] P. Hayes. RDF semantics. W3C Recommendation, Feb. 2004.
[27] D. Heimbigner and D. McLeod. A federated architecture for information management. *ACM Trans. Inf. Syst.*, **3**(3):253–278, 1985.
[28] P. Hitzler and F. van Harmelen. A reasonable Semantic Web. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):39–44, 2010.
[29] P. Hitzler, M. Krötzsch, B. Parsia, P.F. Patel-Schneider, and S. Rudolph (eds.) OWL 2 Web Ontology Language primer. W3C Recommendation. Oct. 2009.
[30] A. Hogan, A. Harth, and S. Decker. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS)*, 2006.
[31] A. Hogan, A. Harth, and S. Decker. Performing Object Consolidation on the Semantic Web Data Graph. In *WWW2007 Workshop I^3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web*, 2007.
[32] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the Pedantic Web. In *3rd Int.'l Workshop on Linked Data on the Web (LDOW2010)*, Apr. 2010.
[33] A. Hogan, A. Harth, and A. Polleres. Scalable Authoritative OWL Reasoning for the Web. *Int.'l Journal on Semantic Web and Information Systems*, **5**(2), 2009.
[34] A. Hogan, A. Harth, J. Umbrich, S. Kinsella, A. Polleres, and S. Decker. Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine. Technical Report DERI-TR-2010-07-23, Digital Enterprise Research Institute (DERI), 2010.
[35] A. Hogan, A. Polleres, J. Umbrich, and A. Zimmermann. Some entities are more equal than others: Statistical methods to consolidate Linked Data. In *Workshop on New Forms of Reasoning for the Semantic Web: Scalable & Dynamic (NeFoRS)*, 2010.
[36] Z. Huang, F. van Harmelen, and A. ten Teije. Reasoning with inconsistent ontologies. In *IJCAI2005*, p. 454–459, 2005.
[37] E. Hyvönen. Preventing ontology interoperability problems instead of solving them. *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):33–37, 2010.
[38] G. Ianni, T. Krennwallner, A. Martello, and A. Polleres. Dynamic querying of mass-storage RDF data with rule-based entailment regimes. In *ISWC2009*, volume 5823 of *LNCS*, p. 310–327, Oct. 2009. Springer.
[39] P. Jain, P. Hitzler, P.Z. Yeh, K. Verma and A.P. Sheth. Linked Data is Merely More Data. In *AAAI Spring Symposium "Linked Data Meets Artificial Intelligence", AAAI Press*, March 2010.
[40] D. Kossmann. The State of the Art in Distributed Query Processing. *ACM Computing Surveys (CSUR)*, **32**(4):422–469, Dec. 2000.
[41] N. Lopes, A. Zimmermann, A. Hogan, G. Lukacsy, A. Polleres, U. Straccia, and S. Decker. RDF needs annotations. In *W3C Workshop on RDF Next Steps*, June 2010.
[42] Y. Ma and P. Hitzler. Paraconsistent reasoning for OWL 2. In *RR2009*, p. 197–211, Oct. 2009. Springer.

[43] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: a document-oriented lookup index for open Linked Data. *Int.'l Journal of Metadata, Semantics and Ontologies*, **3**(1):37–52, 2008.

[44] J. Pérez, M. Arenas, and C. Gutierrez. Semantics and complexity of sparql. In *ISWC2006*, p. 30–43, 2006.

[45] A. Polleres and D. Huynh, editors. *Journal of Web Semantics, Special Issue: The Web of Data*, volume **7**(3). Elsevier, 2009.

[46] A. Polleres, T. Krennwallner, N. Lopes, J. Kopecký, and S. Decker. XSPARQL Language Specification, Jan. 2009. W3C member submission.

[47] A. Polleres, F. Scharffe, and R. Schindlauer. SPARQL++ for mapping between RDF vocabularies. In *ODBASE 2007*, vol. 4803 of *LNCS*, p. 878–896, Nov. 2007. Springer.

[48] G. Qi and J. Du. Model-based revision operators for terminologies in description logics. In *Proceedings of the 21st Int.'l Joint Conference on Artificial Intelligence*, p. 891–897, 2009.

[49] B. Quilitz and U. Leser. Querying distributed RDF data sources with SPARQL. In *ESWC2008*, p. 524–538, June 2008. Springer.

[50] U. Straccia, N. Lopes, G. Lukácsy, and A. Polleres. A general framework for representing and reasoning with annotated semantic Web data. In *AAAI 2010, Special Track on Artificial Intelligence and the Web*, Atlanta, Georgia, USA, July 2010.

[51] H. Stuckenschmidt, R. Vdovjak, J. Broekstra, and G.-J. Houben. Towards distributed processing of RDF path queries. *Int.'l Journal of Web Engineering and Technology*, **2**(2/3):207–230, 2005.

[52] F.M. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *WWW 2007*, 2007. ACM Press.

[53] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. Sig.ma: Live views on the Web of Data. *Journal of Web Semantics*, 2010. To appear.

[54] J. Urbani, S. Kotoulas, E. Oren, and F. van Harmelen. Scalable distributed reasoning using MapReduce. In *ISWC2009*, vol. 5823 of *LNCS*, p. 634–649, Oct. 2009. Springer.

[55] J. Urbani, S. Kotoulas, J. Maassen, F. van Harmelen and H.E. Bal. OWL reasoning with WebPIE: Calculating the closure of 100 billion triples. In *ESWC2010*, p. 213–227, 2010. Springer.

[56] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the Web of data. In *ISWC2009*, vol. 5823 of *LNCS*, p. 650–665, Oct. 2009. Springer.

[57] J. Weaver and J.A. Hendler. Parallel materialization of the finite RDFS closure for hundreds of millions of triples. In *ISWC2009*, vol. 5823 of *LNCS*, p. 682–697, Oct. 2009. Springer.

Semantic Web 1 (2010) 1–2
DOI 10.3233/SW-2010-0017
IOS Press

Editorial

Semantic Web – Interoperability, Usability, Applicability

Pascal Hitzler [a] and Krzysztof Janowicz [b,*]
[a] *Kno.e.sis Center, Wright State University, USA*
[b] *GeoVISTA Center, Pennsylvania State University, USA*

The Semantic Web is here to stay.

While this statement seems obvious, this has not been so a few years ago, when basic research funding seemed to be running out, and industrial uptake was hardly happening. In the meantime, we do not only see sustained funding for Semantic Web related research (in particular by the European Commission), but also significant investment by industry, including major IT and venture capital companies. The Semantic Web is here to stay – and to grow.

The Semantic Web is multidisciplinary and heterogeneous. Many Semantic Web researchers maintain close ties to neighboring disciplines which provide methods or application areas for their work. However, the Semantic Web has now established itself as a research field in its own rights. Consequently, a growing number of researchers, in particular those of the second or third generation, seem to identify themselves with the Semantic Web as their primary field of work. The growing number of top quality events dedicated to Semantic Web topics is also a clear indication of this trend. Another indicator is the increasing interweavement of Semantic Web methods into related disciplines leading to research topics such as geospatial-semantics, the Semantic Sensor Web, semantic desktop, or work on cultural heritage.

The Semantic Web journal is set up to be a forum for highest-quality research contributions on all aspects of the Semantic Web. Its scope encompasses work in neighboring disciplines which is motivated by the Semantic Web vision. Besides the publishing of research contributions, it is also an outlet for reports on tools, systems, applications, and ontologies which *enable* research, rather than being direct research contributions.[1] The journal also publishes top-quality surveys which serve as introductions to core topics of Semantic Web research.

The journal's subtitle – *Interoperability, Usability, Applicability* – reflects the wide scope of the journal, by putting an emphasis on enabling new technologies and methods. *Interoperability* refers to aspects such as the seamless integration of data from heterogeneous sources, on-the-fly composition and interoperation of Web services, and next-generation search engines. *Usability* encompasses new information retrieval paradigms, user interfaces and interaction, and visualization techniques, which in turn require methods for dealing with context dependency, personalization, trust, and provenance, amongst others, while hiding the underlying computational issues from the user. *Applicability* refers to the rapidly growing application areas of Semantic Web technologies and methods, to the issue of bringing state-of-the-art research results to bear on real-world applications, and to the development of new methods and foundations driven by real application needs from various domains.

The primary modern purpose of a scientific journal is quality control and visibility. Fair quality control in scientific publishing directly depends on the quality

*Corresponding author. E-mail: jano@psu.edu.

[1] See http://www.semantic-web-journal.net/authors/ for information on different types of papers accepted for publication.

of the underlying review process and further editorial choices. Today, however, reviewing and publishing is inflationary, which increases potential conflicts of interest and substantially reduces the quality of the typical paper – and of the typical review. While we cannot simply reverse this trend, we can take advantage of the World Wide Web to counteract these developments and improve quality and transparency by bringing the review process out into the public space.

The Semantic Web journal thus relies on an open and transparent review process.[2] All submitted papers as well as the corresponding solicited reviews are made publicly available. All researchers can additionally contribute public reviews and submit comments. Reviewers and editors are publicly known by name. Discussions between reviewers and authors can (and should) happen in public. Reviewers and editors are acknowledged by name in the published versions of the papers.

Reviewers put more effort into providing constructive reviews if their work and contribution to the final manuscript becomes visible. Editors can document their choice of reviewers. Authors can receive additional feedback to ensure that their submission is of sufficient maturity for an archival journal. Public discussions on controversial submissions minimize errors in the decision making and thus result in a fairer procedure.

The success which ensues the setup of the journal is highly encouraging. Until September 2010, we received more than 50 paper submissions; this does not include the vision statement papers contained in this issue. We have several special issues from various domains lined up, some of which have not been publicly announced yet. Researchers have contributed open reviews without us asking them to. So far, there have only been very few occasions where a solicited reviewer has asked to remain anonymous.

This very first issue of the Semantic Web journal contains vision statements by the members of the Editorial Board. While these contributions were essentially invited, they nevertheless underwent the full, open, and transparent review process of the journal in order to improve quality and clarity. Their publication on the journal's webpage already led to comments and open reviews from external researchers. The resulting collection is an impressive compilation of topics of core concern to the Semantic Web community. While it cannot possibly be exhaustive in terms of the many aspects of Semantic Web research, its breadth of coverage is indicative of the breadth of scope of the journal.

We thank all contributors, authors and reviewers alike. If you would also like to contribute to the journal in any way, you can find us at http://www.semantic-web-journal.net/.

Acknowledgements

This journal wouldn't be here without Arnoud de Kemp. We thank him for taking the initiative. This journal wouldn't be sustainable without all the hard work put into it by the editorial board members, and by the responsible staff at IOS Press. Thank you for your support.

[2] See http://www.semantic-web-journal.net/reviewers#review.

Semantic Web 1 (2010) 131–135
DOI 10.3233/SW-2010-0018
IOS Press

Building an effective Semantic Web for health care and the life sciences

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Rinke Hoekstra, Universiteit van Amsterdam, The Netherlands; Kunal Verma, Accenture, USA

Michel Dumontier
Department of Biology, Institute of Biochemistry, School of Computer Science, Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, Canada, K1S5B6
E-mail: michel_dumontier@carleton.ca

Abstract. Health Care and the Life Sciences (HCLS) are at the leading edge of applying advanced information technologies for the purpose of knowledge management and knowledge discovery. To realize the promise of the Semantic Web as a framework for large-scale, distributed knowledge management for biomedical informatics, substantial investments must be made in technological innovation and social agreement. Building an effective Biomedical Semantic Web will be a long, hard and tedious process. First, domain requirements are still driving new technology development, particularly to address issues of scalability in light of demands for increased expressive capability in increasingly massive and distributed knowledge bases. Second, significant challenges remain in the development and adoption of a well founded, intuitive and coherent knowledge representation for general use. Support for semantic interoperability across a large number of sub-domains (from molecular to medical) requires that rich, machine-understandable descriptions are consistently represented by well formulated vocabularies drawn from formal ontology, and that they can be easily composed and published by domain experts. While current focus has been on data, the provisioning of Semantic Web services, such that they may be automatically discovered to answer a question, will be an essential component of deploying Semantic Web technologies as part of academic or commercial cyberinfrastructure.

Keywords: Semantic Web, health care, life sciences, digital libraries, cyberinfrastructure, ontology

1. Introduction

The vision of the Semantic Web (SW) outlines that common standards for all aspects of knowledge management will facilitate the development of an interoperable ecosystem of data and services so that it becomes easier to publish, find, and re-use information in ways that go beyond their original design (Berners-Lee, Hendler, & Lassila, 2001). As a major consumer of information technologies, the Health Care and Life Sciences (HCLS) has traditionally placed demanding requirements to support activities related to knowledge management and knowledge discovery. While HCLS data is highly heterogeneous and growing at an unprecedented rate, SW technologies offer a salient solution to accurately publish this diverse knowledge in so that it becomes a major resource for research and development. In fact, the W3C Semantic Web HCLS Interest Group is specifically chartered to develop, advocate and support SW technologies for HCLS communities (HCLS, 2005). Our experience maintains that in order to build an effective Semantic Web for the HCLS, significant efforts still have to be made towards the coordinated development of high quality vocabularies, well thought out protocols for data sharing and publication, and scalable, cohesive cyberinfrastructure.

Coordinated efforts by a wide range of communities to promote a coherent representation of data will foster commoditization of information and create

entirely new commercial opportunities and public-good efforts devoted to provisioning data, in-depth analysis and effective visualization. There is little doubt that by making biomedical data available through the Semantic Web, we will dramatically improve overall productivity, increase investment returns, decrease the cost of research, create new economic activity and augment the outcomes of basic and applied research. The challenge then is to assess the vision for the Semantic Web with respect to the state-of-the art in knowledge representation and technology.

2. State of the art

The SW positions itself as a platform for information exchange between intelligent agents. Interoperability is achieved by ensuring that the information is consistently encoded (syntax) and uses symbols that have a formally defined meaning such that they can be consistently interpreted (semantics). An effective Semantic Web will ensure interoperability between cyberinfrastructure components including i) capacity to capture knowledge, ii) infrastructure to publish and share information, iii) efficient middleware for question answering and knowledge discovery.

2.1. RDF and linked data

The Resource Description Framework (RDF) is a core SW language that offers a lightweight mechanism to describe entities in term of their types, attributes and relations to other entities. Entities are identified by International Resource Identifiers (IRIs) which includes web based identifiers (HTTP URIs) that can be resolved on the Web. Statements about these entities captured as subject-predicate-object "triples", and are described using vocabularies from domain-specific ontologies. RDF Schema (RDFS) makes it possible to specify simple type and relation hierarchies using the "is a" relation. RDF can be queried using the SPARQL query language.

A number of life science projects are using RDF as their core language of representation and publishing the information so that information about the entities can be queried and visualized. Bio2RDF[1] is at the forefront of generating and provisioning ~40 billion triples of linked life science data from over 40 high profile databases. Bio2RDF normalizes the data IRIs so as to facilitate linking of datasets (Belleau, Nolin, Tourigny, Rigault, & Morissette, 2008). Each dataset is deployed as its own SPARQL endpoint, which allows original data provider to actively participate in the network while decentralization of resource offerings provides web-scalability. Bio2RDF offers specialized federated query services across its global mirrors (Ottawa, Quebec City, Guelph and Brisbane). The Linking Open Drug Data (LODD) [2] and Chem2Bio2RDF[3] projects are generating linked data to support chemical-based investigations including drug discovery. These projects provision RDF data from relational databases using D2R. LinkedLifeData[4] consists of a diverse array of life science datasets provisioned through cluster-based data warehouse solution using the commercial BigOWLIM engine. Yet all of these projects largely involve information retrieval in the most basic sense, without making full use of the background knowledge provided by ontologies.

2.2. Ontologies

Initially driven by the need to query gene and gene product annotation across a number of model organisms, the Gene Ontology (GO) has emerged as an vast controlled vocabulary of biological processes, molecular functions and cellular components (GO Consortium, 2008). Since its inception, GO strives to more accurately describe their 20,000+ terms principally organized via an "is a" axis, but also augmented with other relations (e.g. parthood). Following GO, there are now over 150 Open Biomedical Ontologies (OBO) listed at the National Center for Bio-Ontology (NCBO) BioPortal, which now spans molecular, anatomical, physiological, organismal, health, experimental information. Yet significant overlap exists between ontologies, as a search yielding 20 different terms for "protein" will attest. Towards developing a set of orthogonal ontologies, the OBO Foundry (Smith et al., 2007) promotes development over basic categories drawn from the Basic Formal Ontology (BFO) and encourages the use of reuse basic, domain-independent relations from the Relational Ontology (RO). Well defined relations should make it clear when the relations are to be used, and what inferences, if any, may be drawn from them.

[1] http://bio2rdf.org

[2] http://esw.w3.org/HCLSIG/LODD

[3] http://chem2bio2rdf.org

[4] http://linkedlifedata.org

2.3. OWL and linked knowledge

Drawing from the well understood area of Description Logics, the Web Ontology Language (OWL) provides a substantially more expressive vocabulary to axiomatically describe entities for enhanced reasoning. Building these kinds of ontologies not only requires domain expertise to properly define describe the entities, but also requires a keen understanding of formal knowledge representation so that knowledge is properly captured and becomes intuitive to query using an information system.

Several projects have now demonstrated the use of OWL-based information systems. The HCLS knowledge base contains a collection of instantiated ontologies used to identify interesting molecular agents in the treatment of Alzheimer's (Ruttenberg et al., 2007). With consideration of how genetics plays a role in effective drug treatment, the Pharmacogenomics Knowledge Base (PGKB) offers depression-related pharmacogenomic information that facilitates additional knowledge curation beyond the PharmGKB database (Dumontier & Villanueva-Rosales, 2009). Thus, ontologies can play an important role both in semantic data integration as well as guide curation activities with well established use cases towards populating a specialized knowledge base.

2.4. Semantic Web services

Web services define application programming interfaces by structuring messages and content with the Web Services Description Language (WSDL). HCLS web services may be registered and annotated using the Web 2.0 inspired BioCatalogue (Goble, Stevens, Hull, Wolstencroft, & Lopez, 2008). Workflow application tools like Taverna facilitate chaining of services, to obtain and logically consume content (Oinn et al., 2004). Yet, the pairing of services still remains rather difficult because the inputs are generally datatypes as opposed to semantic types that can be reasoned about. SADI, a new Semantic Web services framework project, uses OWL ontologies to formally describe services, in which the Semantic Health And Research Environment (SHARE) query system undertakes service matchmaking and invocation through a SPARQL query (Vandervalk, McCarthy, & Wilkinson, 2009). This has been put to use in CardioSHARE, a system that integrates patient data with analytical services so as to identify *bone fide* cardiovascular health indicators.

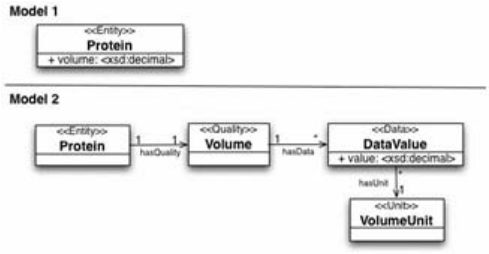

Fig. 1. Two models for representing a physical attribute.

3. Challenges

3.1. Scalable Semantic Web technologies

Requirements of Semantic Web technologies have been drawn from extensive analysis of domain requirements, technical feasibility and vendor capabilities. While these including HCLS centric concerns, they do not reflect the enormous amounts of data (trillions of facts), nor the widespread and decentralized nature of databases (thousands of indirectly connected databases) that would have to be accommodated. Current stand-alone solutions appear to scale up into hundreds of millions of triples, while cluster-based solutions (Virtuoso Cluster Edition; BigOWLIM; BigData) appear to scale into the tens or hundreds of billions of statements, but with highly restricted capability to reason about OWL data. New and sustained efforts into large-scale reasoning and possibly incomplete reasoning may be required, as recently demonstrated (Urbani, Kotoulas, Maaseen, van Harmelen, & Bal, 2010).

3.2. From linked data to linked knowledge

RDF linked data efforts currently employ a simple model for representing knowledge: entities are either related to other entities or related to valued attributes through a single relation. Model 1 (Fig. 1) exemplifies a typical linked data model for representing the volume of a protein using a decimal datatype. Such a model does not express the unit of measure, and no statements can be made as to how or under what conditions the value was obtained. In contrast, Model 2 overcomes these limitations by explicitly representing the entity, quality, measurement value, and the unit as distinct entities. However, moving from 2 triples in Model 1 to the 8 triples required in Model 2 translates to a 4x increase in the storage requirements

and requires more sophisticated query to retrieve all the relevant information. The benefit increasing our capacity to make meaningful statements about any one of these entities, which cannot (easily) be done in Model 1, is nevertheless substantial.

3.3. Consistent knowledge representation

If Model 2 is deemed desirable, then the challenge lies not only in getting scalable systems to accommodate this influx of triples (possibly by devising customizable indexes), but also in getting users to learn about and deploy standard patterns which they can apply to their own data. The patterns should be coherent, intuitive and well specified such that non-experts can read, understand and apply the guidelines found therein. Importantly, these patterns should specify the relations that hold between instances, and for this reason having a coherent, well founded set of types and basic relations supported by formal ontology is of critical value. While BFO+RO combination provides guidance for instantiable types, it lacks the capacity to handle all elements of scientific discourse (Dumontier & Hoehndorf, 2010), specifically with types that may be hypothesized (putative agents of disease), predicted (genes and proteins from genomic sequences), or simply do not occur (perpetual motion). This necessitates significantly more effort in developing a foundational ontology (types + relations) to represent a more diverse array of knowledge, including that which is *already* found in linked data.

Recent work by the W3C HCLS subgroup on translational medicine has produced a knowledge base composed of the Translational Medicine Ontology, which provides 75 core classes mapped to 223 classes from 40 ontologies, and acts as a global schema over a set of fake patient data and linking open data (LOD) resources (Dumontier et al., 2010). They featured queries that span bedside to bench by not only matching patients to clinical trials, but also in finding trials for which their drugs had different mechanisms of action so as to potentially avoid common side effects. Here, the integration of electronic health records with public data provides new avenues for clinical research and improved health care. With increased interest in building smarter health care systems using electronic health records, Semantic Web technologies can play a pivotal role in incentivizing interoperability between health care providers by linking valuable to public data.

3.4. The need for axiomatic description of classes

Until recently, OBO ontologies have been largely crafted using the OBO language, an ad-hoc language with its own (non-XML) syntax and lacking formal semantics. OBO ontologies differ enormously in terms of their development status, expressivity, and overall quality. While the standard transformation to OWL involves fixed semantics, more recent work demonstrates how more flexible semantics can be assigned as patterns associated with well defined relations such as the RO (Hoehndorf et al., 2010). Axiomatic description of classes should improve ontology quality by forcing ontology designers to be explicit about the necessary conditions for class membership, as opposed to relying on potentially vague descriptions using natural language. Such formalization can make use of automated reasoners to find errors and provide explanations for unexpected inferences.

3.5. Provenance and attribution

Provenance and the corresponding attribution of knowledge is normal practice in science. Several approaches (Research Objects, Provenance Ontology, Provenir Ontology, SWAN-SIOC provenance) have now been articulated, and must now be unified. Importantly, contributions to community-based ontologies need to be acknowledged. Further, the wholesale provenance of data need to be specified, and while RDF reification or OWL axiom annotations supports this, they generate significantly higher overhead (4 triples per statement). In contrast, TriX/TRiG/RDF Named Graphs may be more effective and needs to go down the path of standardization.

3.6. User interfaces

Despite a decade of research and development around Semantic Web technologies, significant gaps still remain in tools that facilitate data management and knowledge discovery. User interfaces are still developed "close to the metal", forcing a model that is not meant for human consumption. New innovative approaches need to consider FreeBase's Parallax[5], but for the Semantic Web. Impressively, the sig.ma[6] Mashup tool uses the Sindice Semantic Web Search engine to provide an enhanced view of indexed RDF triples, including those provided by Bio2RDF and

[5] http://www.freebase.com/labs/parallax/
[6] http://sig.ma/

DBpedia. For OWL knowledge bases, SMART (Battista, Villanueva-Rosales, Palenychka, & Dumontier, 2007) offers a way to craft queries as class expressions using the Manchester OWL syntax. Significantly more research in human-computer interaction is required to identify effective ways to work with with hyper-dimensional data from multiple (and possibly untrustworthy) sources.

4. Conclusion

Building an effective Semantic Web for HCLS is clearly a long term effort that needs coherent representations along with simple tools to create, publish, query and visualize generic Semantic Web data. With hundreds of bioinformatics web services, thousands of biological databases and millions of unrecorded facts in waiting, significant effort will also have to be placed in training the next generation of application developers to correctly use Semantic Web technologies. HCLS communities can then be served by custom portals, and ultimately act as a key component of cyberinfrastructure for both textual and semantically annotated data and services.

References

[1] Battista, A.D., Villanueva-Rosales, N., Palenychka, M., & Dumontier, M. (2007). *SMART: A Web-Based, Ontology-Driven, Semantic Web Query Answering Application*. Busan, South Korea: International Semantic Web Conference.

[2] Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., & Morissette, J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Semantics*, **41**(5), 706–716.

[3] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, **284**, 34–43.

[4] Dumontier, M., & Hoehndorf, R. (2010). Realism for Scientific Ontologies. In *6th International Conference on Formal Ontology in Information Systems* (p. 12). Toronto, Canada.

[5] Dumontier, M., & Villanueva-Rosales, N. (2009). Towards pharmacogenomics knowledge discovery with the semantic web. *Briefings in Bioinformatics*, **10**(2), 153–163.

[6] Dumontier, M., Andersson, B., Batchelor, C., Denney, C., Domarew, C., Jentzsch, A., et al. (2010). The Translational Medicine Ontology: Driving personalized medicine by bridging the gap from bedside to bench. In *Proceedings of Bio-Ontologies 2010: Semantic Applications in Life Sciences* (p. 4). Boston.

[7] GO Consortium. (2008). The Gene Ontology project in 2008. *Nucleic Acids Research*, **36**(Database issue), D440-4.

[8] Goble, C., Stevens, R., Hull, D., Wolstencroft, K., & Lopez, R. (2008). Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, **9**(6), 506–517.

[9] HCLS. (2005). Semantic Web Health Care and Life Sciences Interest Group. Retrieved from http://www.w3.org/blog/hcls.

[10] Hoehndorf, R., Oellrich, A., Dumontier, M., Kelso, J., Herre, H., Rebholz-Schuhmann, D., et al. (2010). OWLDEF: Integrating OBO And OWL. In *Proceedings of the 7th International OWL: Experiences and Directions*. San Francisco.

[11] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., et al. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics (Oxford, England)*, **20**(17), 3045–3054.

[12] Ruttenberg, A., Clark, T., Bug, W., Samwald, M., Bodenreider, O., Chen, H., et al. (2007). Advancing translational research with the Semantic Web. *BMC bioinformatics*, **8** *Suppl. 3*, S2.

[13] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, **25**(11), 1251–1255.

[14] Urbani, J., Kotoulas, S., Maaseen, J., van Harmelen, F., & Bal, H. (2010). OWL reasoning with WebPIE: calculating the closure of 100 billion triples. *Proceedings of the 2010 Extended Semantic Web Conference*.

[15] Vandervalk, B.P., McCarthy, E.L., & Wilkinson, M.D. (2009). Moby and Moby 2: creatures of the deep (web). *Briefings in bioinformatics*, **10**(2), 114–128.

Semantic Web 1 (2010) 97–104
DOI 10.3233/SW-2010-0019
IOS Press

Making the Web a data washing machine

Creating knowledge out of interlinked data

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Claudia d'Amato, Università degli Studi di Bari, Italy; Rinke Hoekstra, Universiteit van Amsterdam, The Netherlands
Open review(s): Prateek Jain, Wright State University, USA

Sören Auer * and Jens Lehmann
Universität Leipzig, Institut für Informatik, PF 100920, 04009 Leipzig, Germany
E-mail: lehmann@informatik.uni-leipzig.de

Abstract. Over the past 3 years, the Semantic Web activity has gained momentum with the widespread publishing of structured data as RDF. The Linked Data paradigm has therefore evolved from a practical research idea into a very promising candidate for addressing one of the biggest challenges in the area of the Semantic Web vision: the exploitation of the Web as a platform for data and information integration. To translate this initial success into a world-scale reality, a number of research challenges need to be addressed: the performance gap between relational and RDF data management has to be closed, coherence and quality of data published on the Web have to be improved, provenance and trust on the Linked Data Web must be established and generally the entrance barrier for data publishers and users has to be lowered. In this vision statement we discuss these challenges and argue, that research approaches tackling these challenges should be integrated into a mutual refinement cycle. We also present two crucial use-cases for the widespread adoption of Linked Data.

Keywords: Linked Data, Semantic Web

One of the biggest challenges in the area of intelligent information management is the exploitation of the Web as a platform for data and information integration as well as for search and querying. Just as we publish unstructured textual information on the Web as HTML pages and search such information by using keyword-based search engines, we should be able to easily publish structured information, reliably interlink this information with other data published on the Web and search the resulting data space by using expressive querying. The Linked Data paradigm has evolved as a powerful enabler for the transition of the current document-oriented Web into a Web of interlinked Data and, ultimately, into the Semantic Web. The term Linked Data here refers to a set of best practices [4] for publishing and connecting structured data on the Web. These best practices have been adopted by an increasing number of data providers over the past three years, leading to the creation of a global data space that contains many billions of assertions – the Web of Linked Data. However, in order to sustainably establish the Web of Data and to maximize the value of published data on the Web, we are facing four fundamental challenges: (1) we have to improve the performance of very large-scale RDF Data Management, (2) we have to increase and ease the interlinking and fusion of information, (3) algorithms and tools have to be developed for improving the structure, semantic richness and quality of Linked Data, (4) adaptive user interfaces and interaction paradigms have to deployed for authoring and maintaining Linked Data.

In the remainder of this vision statement we elaborate on these challenges, possible approaches for solv-

*Corresponding author. E-mail: auer@informatik.uni-leipzig.de.

ing them and present two crucial use-cases for Linked Data.[1]

1. Improving the performance of large-scale RDF data management

Experience demonstrates that an RDF database can be an order of magnitude less efficient than a relational representation running on the same engine [5]. This lack of efficiency is perceived as the main obstacle for a large-scale deployment of semantic technologies in corporate applications or for expressive Data Web search. For RDF to be the lingua franca of data integration, which is its birthright, its use must not bring significant performance penalty over the much less flexible best practices prevalent today. The main reason for the difference in performance between triple stores and relational databases is the presence of fine-grained optimised index structures in relational systems in contrast to more flexible and extensible structures in triple stores. The performance gap between relational and RDF data management can be mitigated by developing adaptive automatic data indexing technologies that create and exploit indexing structures as and when needed, entirely based on received query workload.

The performance of knowledge stores can, for example, be significantly increased by applying query subsumption and view maintenance approaches to the RDF data model. Query subsumption can be based on analysing the graph patterns of cached SPARQL queries in order to obtain information on (a) whether the previously cached query result can be reused for answering a subsequent query and (b) which updates of the underlying knowledge base will change the results of cached queries and thus have to trigger invalidation. As queries are executed, intermediate results can be persisted and labeled for reuse. When a subsequent query is executed, it can reuse those results if it either subsumes the previous query or is subsumed by it. In the latter case the query can be performed on the persisted previous results and in the first case join operations between persisted and base data can be performed. This builds query shortcuts across the data, essentially making materialized views on demand. These are either invalidated or brought up to data as data changes and discarded if no longer needed. The cacheable operations are joins and inferences such as `owl:sameAs`, transitive property traversal, class and property hierarchies etc. Intermediate materializations may also cache aggregates. First steps in this direction were for example performed in [6,14]. In addition, caching and view materialization techniques should be able to handle implicit information commonly found in ontologies. In order for Linked Data to be successful in the Web at large and within enterprises in particular, such new RDF indexing technology must ultimately find its way in RDF processing systems.

2. Increase and ease the interlinking and fusion of information

While the sum of data published as Linked Data amounts already to billions of triples and grows steadily, the number of links between them is several orders of magnitude smaller and by far more difficult to maintain (cf. [12]).

The task of interlinking and supplementing the knowledge bases with information from external data sets, knowledge bases and ontologies can draw from previous work within different research communities: Interlinking has a long history in database research and occurs in the literature under a dozen of terms [19] such as Deduplication [9], Entity Identification [13], Record Linkage [7] and many more. Encountered problems are generally caused by data heterogeneity. The processes of data cleaning [16] and data scrubbing [22] are common terms for resolving such identity resolution problems. Elmagarmid et al. (2007) [8] distinguish between structural and lexical heterogeneity and focus their survey on the latter. According to Elmagarmid et al., a stage of data preparation is a necessary prerequisite to efficient record linkage and consists of a parsing, a data transformation and a standardization step. As a new challenge, RDF and OWL, alongside with the Linked Data paradigm and commonly published vocabularies, provide the means necessary to skip the data preparation step as they have already proliferated a shared structural representation of data as well as a common access mechanism. With the availability of large open data sets and links between them, the generation of benchmarks for interlinking becomes feasible and can add an edge to research in this area. The need for adaptive methods also arises in order to cope with changing data. Thus, auto-

[1] The interested reader may also want to have a look at a related article [15] in this issue, which poses similar challenges on dealing with Linked Data from a slightly different perspective.

mated reinforced approaches have to be developed that adapt themselves over time. (Both research directions are mentioned in Elmagarmid et al.) Further reading can be found in a survey paper on Ontology Matching [18]. Although there has been extensive work on the topics of interlinking and ontology matching, the new situation creates new requirements and challenges and calls for an adaptation of existing methods.

The availability of large open data sets and accessibility via Linked Data pose the following requirements, currently insufficiently covered by research. Likewise, numerous specifics related to combined instance and schema matching in RDF and OWL w.r.t. timeliness are hardly addressed:

- ETL (Extraction, Transformation, Loading) of legacy data under the aspect of linking.
- Lack of benchmarks for instance and schema mapping as well as an evaluation framework and metrics.
- As knowledge bases evolve, links and mappings have to evolve likewise. This poses special requirements on scalability and maintenance.
- Web data sources often mix terms from different RDF vocabularies and OWL ontologies. This aspect is not covered by previous work on database schema and ontology matching, which builds upon the assumption of the existence of a single schema or ontology.
- Database schemata impose harder restrictions on the instance structure compared to Semantic Web data where the open world assumption applies and information about instances is not assumed to be complete.
- The standardization of the representation format (RDF) allows the creation of links based on the availability of third-party knowledge bases from the LOD cloud such as DBpedia (similar to the star-like pattern in a mediation-based EAI).

A promising approach, which can respond to some of these requirements is to integrate schema mapping and data interlinking algorithms into a mutual refinement cycle, where results on either side (schema and data) help to improve mapping and interlinking on the other side. Both unsupervised and supervised machine learning techniques should be investigated for this task, where the latter enable knowledge base maintainers to produce high quality mappings. Further research is needed in the area of data fusion, i.e. the process of integrating multiple data items, representing the same real-world object into a single, consistent, and clean representation. The main challenge in data fusion is the reliable resolution of data conflicts, i.e. choosing a value in situations where multiple sources provide different values for the same property of an object.

The usefulness of a knowledge base increases with more (correct) links to other knowledge bases (network effect), since this allows applications to combine information from several knowledge bases. The advantage of de-referenceable URIs (URLs) as identifiers is two-fold. Contrary to a database id, URLs are unique identifiers on the Web, which have a defined semantics and provenance. Furthermore, the available data identified by URLs are easily accessible via content-negotiation and standard retrieval mechanisms (i.e. the HTTP protocol). In this situation, linking brings immediate advantages, because it defines relations, e.g. equality, of Web identities and allows the convenient aggregation of data by following these links. The emerging Web of Data now faces several challenges and problems. a) How to find all links between two knowledge bases (high recall)? b) How to verify the correctness of found links (high precision)? This issue is most important in case of `owl:sameAs` links, which entail strict logical equivalence and therefore need to be very precise. c) How to maintain such a link structure with evolving knowledge bases? To solve those challenges, a constant evaluation of links between knowledge bases is necessary. Ideally, scalable machine learning techniques should be applied to generate links based on manually provided and maintained test sets. Although approaches and tools in this direction are developed, they still require higher usability and scalability to have a wider impact in the Web of Data.

After links are found and verified the next challenge is the fusing of data with respect to completeness, conciseness and consistency.

3. Improving the structure, semantic richness and quality of Linked Data

Many data sets on the current Data Web lack structure as well as rich knowledge representation and contain defects as well as inconsistencies. Methods for learning of ontology class definitions from instance data can facilitate the easy incremental and self-organizing creation and maintenance of semantically rich knowledge bases. Particularly, such methods can

be employed for enriching and repairing knowledge bases on the Data Web.

The enrichment steps can be done by learning axioms, e.g. equivalence and inclusion axioms, whose left-hand side is an existing named class in the knowledge base. The task of finding such axioms can be phrased as a positive-only supervised machine learning problem, where the positive examples are the existing instances of the named class on the left hand side and the background knowledge is the knowledge base to be considered. The advantage of those methods is that they ensure that the learned schema axioms fit the instance data. The techniques should also be robust in terms of wrong class assertions in the knowledge base. We argue that those learning methods should be semi-automatic, i.e. a knowledge engineer makes the final decision whether to add one of the suggested axioms to the knowledge base. Once an axiom is added, it can be used by inference methods, for example, to populate the knowledge base with inferred facts, or spot and repair inconsistencies. One challenge is to be able to apply such machine learning methods to very large knowledge bases. Machine learning methods usually rely heavily upon reasoning techniques and currently it is not possible to reason efficiently over large knowledge bases which consist of more than 100 million RDF triples[2]. Therefore, extraction methods should be used to extract a relevant fragment of a knowledge base with respect to given individuals, which is sufficiently small to reason over it, while still containing sufficient information with respect to those instances to apply class learning algorithms. Experiments that employ such extraction methods against the DBpedia knowledge base have already shown promising results [10].

Another method for increasing the quality of Linked Data is semi-automatic repair. In particular large knowledge bases are often prone to modelling errors and problems, because their size makes it difficult to maintain them in a coherent way. These modelling problems can cause an inconsistent knowledge base, unsatisfiable classes, unexpected reasoning results, or reasoning performance drawbacks. We need algorithms, which detect such problems, order them by severity, and suggests possible methods for resolving them to the knowledge engineer. By considering only certain parts of a (large) knowledge base, those algorithms can be able to find problems in a relevant fragment, even if the overall knowledge base is not consistent. Repair approaches can be also used in combination with knowledge base enrichment algorithms: After learning a formal description of a class, problems in the knowledge base may be spotted. Those problems can then be repaired through the knowledge engineer by giving him or her possible suggestions for resolving them.

Two challenges have to be addressed in order to develop enrichment and repair methods as described above: Firstly, existing machine learning algorithms have to be extended from basic Description Logics such as ALC to expressive ones such as *SROIQ(D)* serving as the basis of *OWL 2*. Secondly, the algorithms have to be optimized for processing very large-scale knowledge bases, which usually cannot be loaded in standard OWL reasoners. In addition, we have to pursue the development of tools and algorithms for user friendly knowledge base maintenance and repair, which allow to detect and fix inconsistencies and modelling errors.

4. Adaptive user interfaces and interaction paradigms

All the different Data Web aspects heavily rely on end-user interaction: We have to empower users to formulate expressive queries for exploiting the rich structure of Linked Data. They have to be engaged in authoring and maintaining knowledge derived from heterogeneous and dispersed sources on the Data Web. For interlinking and fusing, the classification of instance data obtained from the Data Web as well as for structure and quality improvements, end users have to be enabled to effortlessly give feedback on the automatically obtained suggestions. Last but not least, user interaction has to preserve privacy, ensure provenance and, particularly in corporate environments, be regulated using access control.

The adaptive nature of the information structures in the Data Web is particularly challenging for the provisioning of easy-to-use, yet comprehensive user interfaces. On the Data Web, users are not constrained by a rigid data model, but can use any representation of information adhering to the flexible RDF data model. This can include information represented according to heterogeneous, interconnected vocabularies defined and published on the Data Web, as well as newly defined attributes and classification structures.

[2] Amongst others, the LarkC project (http://www.larkc.eu/) works on (incomplete) OWL reasoning.

Hence, the user interface components on the Web of Data should support the reuse of existing vocabularies and ontologies as well as the ad hoc definition of new ones and their refinement and evolution in concordance with the data over time. Different information structures should be seamlessly combinable in a provenance preserving way in a single visualization or authoring environment, even if the information to be visualized or authored is obtained or stored in various Linked Data sources ([3,21] are approaches going in that direction). The authoring tools should hide technicalities of the RDF, RDFS or OWL data models from end users, thus realizing WYSIWIG for the authoring of knowledge bases. Based on the information structures found, the most suitable authoring widgets should be automatically combined.

In particular for ordinary users of the Internet, Linked Data is not yet sufficiently visible and (re-) usable. Once information is published as Linked Data, authors hardly receive feedback on its use and the opportunity of realizing a network effect of mutually referring data sources is currently unused. On the social web, technologies such as *Refback*, *Trackback* or *Pingback* enabled the timely notification of authors once their posts were referenced. In fact, we consider these technologies as crucial for the success of the social web and the establishment of a network effect within the blogosphere. In order to establish a similar network effect for the Data Web, we should investigate how such notification services can be applied to the Web of Data. The Semantic Pingback method as described in [20], for example, can serve here as a technical foundation, but much more work is required to integrate such notification services in particular with adequate user interfaces into the fragmented landscape of ontology editors, triple stores and semantic wikis.

5. Making the Web a 'washing machine' for Linked Data

The four challenges presented in the previous sections should be tackled not in isolation, but by investigating methods which facilitate a mutual fertilization of approaches developed to solve these challenges. Examples for such mutual fertilization between approaches include:

- The detection of mappings on the schema level, for example, will directly affect instance level matching and vice versa.

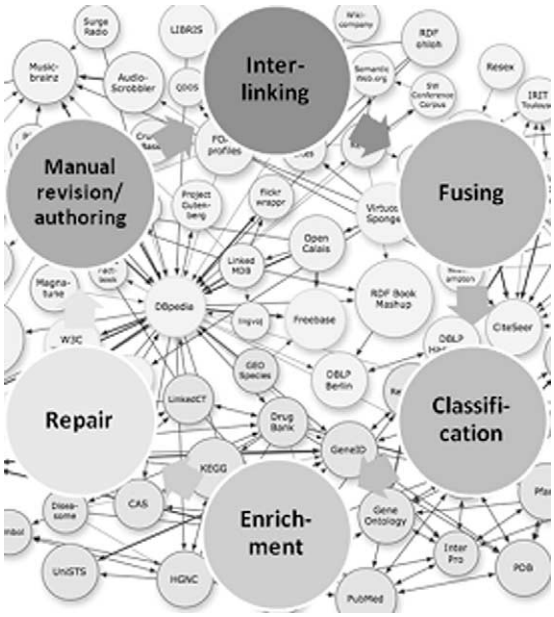

Fig. 1. Linked Data improvement cycle.

- Ontology schema mismatches between knowledge bases can be compensated for by learning which concepts of one are equivalent to which concepts of the other knowledge base.
- Feedback and input from end users (e.g. regarding instance or schema level mappings) can be taken as training input (i.e. as positive or negative examples) for machine learning techniques in order to perform inductive reasoning on larger knowledge bases, whose results can again be assessed by end users for iterative refinement.
- Semantically enriched knowledge bases improve the detection of inconsistencies and modelling problems, which in turn results in benefits for interlinking, fusion, and classification.
- The querying performance of the RDF data management directly affects all other components and the nature of queries issued by the components affects the RDF data management.

As a result of such interdependence, we should pursue the establishment of an improvement cycle for knowledge bases on the Data Web – i.e. make the Web a Linked Data washing machine. The improvement of a knowledge base with regard to one aspect (e.g. a new alignment with another interlinking hub) triggers a number of possible further improvements (e.g. additional instance matches).

The challenge is to develop techniques, which allow to exploit these mutual fertilizations in the distributed medium Web of Data. One possibility is that, various algorithms make use of shared vocabularies for publishing results of mapping, merging, repair or enrichment steps. After one service published his new findings in one of these commonly understood vocabularies, notification mechanisms (such as Semantic Pingback [20]) can notify relevant other services (which subscribed to updates for this particular data domain) or the original data publisher, that new improvement suggestions are available. Given a proper management of provenance information, improvement suggestions can later (after acceptance by the publisher) become part of the original dataset.

6. Complementing SOA with Linked Data

Competitive advantage increasingly depends on business agility, i.e. becoming the "real-time enterprise." This entirely depends on dealing with a constant flood of information, both internal and external. Linked Data is a natural addition to the existing document and web service or SOA based intranets and extranets found in large corporations. Enterprise information integration needs grow continuously. Mergers and acquisitions further drive diversity of IT infrastructure and the consequent need for integration. Simultaneously, enterprise data warehouse sizes have been more than doubling annually for the past several years, effectively outstripping Moore's law. The rapid development in the quantity and quality of structured data on the Internet creates additional opportunities and challenges. The main issues of integration are the use of different identifiers for the same thing and diversity in units of measure. Classifications, application of Linked Data principles for consistent use of identifiers in information interchange and making schema semantics explicit and discoverable, thus effectively rendering data self-describing, offer great promise with no disruption to infrastructure. An important distinction for the adoption of the Data Web paradigm in corporate scenarios is that the reuse of identifiers and vocabularies is not the same thing as making all data public. Corporate data intranets based on Linked Data technologies can help to reduce the data integration costs significantly and entail substantial benefits [11]. Linking internal corporate data with external references from the Data Web will allow a corporation to significantly increase the value of its corporate knowledge with relatively low effort. Examples include integration of product, customer/supplier, materials, regulatory, market research, financial statistics and other information between internal and external sources. The key to this is resolving the disparity of identifiers and associating explicit semantics to relational or XML schemes.

For example, the information integration with the Data Web will allow the sales database of a company to be enhanced with semantic descriptions of customers, products and locations by linking the internal database values with RDF descriptions from the LOD cloud (e.g. from the DBpedia, WikiCompany or Geonames data sets). The linking can be used to correct, aggregate and merge information. In addition to this, the reasoning and semantic mining tools can infer and generate new knowledge that only experts can provide. For example, using machine learning algorithms and the background knowledge from the Data Web, groups of customers can be better classified and described semantically, potentially leading to better targeting and market intelligence.

7. Publishing public and governmental data on the Web

Besides employing the Linked Data paradigm in corporate environments another scenario with a great application potential is the publishing of public and governmental data (cf. [1,17]). Quite some governmental and administrative information in Europe, for example, is already publicly available in structured form. Unfortunately, this data is scattered around, uses a variety of different incompatible data formats, identifiers and schemata.

The adaptation and deployment of Linked Data technologies in this area will increase the ability of the public to find, download, and creatively use data sets that are generated and held by various governmental branches and institutions, be it supra-national (e.g. European) or national ones as well as regional governments and public administrations. In particular, for the case of Europe this will be very challenging due to the large organizational and linguistic diversity. This decentralization and diversity renders centralized and strictly top–down approaches less suitable and thus European governments and public administrations represent an ideal application scenario for Linked Data

technologies. This scenario has been recognized[3], but should be explored much further.

In order to realize this application scenario, a number of different aspects have to be considered and different technologies should be deployed. For example, a portal as well as a network of decentralized registries should provide descriptions of the data sets (i.e. metadata), information about how to access the data sets, facilities for sharing views and reports, as well as tools that leverage government data sets. Based on research approaches tackling the above mentioned challenges, tools and services should be deployed to classify and interlink data sets automatically, to assess their information quality and to suggest enrichments and repairs to the published data sets. Public participation and collaboration will be paramount in this application scenario. The public has to be engaged to provide additional downloadable data sets, to build applications, conduct analyses, and perform research. The published information can improve based on feedback, comments, and recommendations. As a result Linked Data has the potential to improve access to governmental data and expand creative use of those data beyond the walls of government by encouraging innovative ideas (e.g., mashups and web applications). The Linked Data paradigm can help to make governments more transparent and thus strengthen democracy and promote efficiency and effectiveness in governments.

8. Conclusions

While the past few years have been very succesful for the Linked Data initiative and the Web of Data, there has also been well-founded criticism [12]. As a consequence, we pointed out a number of challenges, which need to be solved in order to exploit the Web of Linked Data as medium for information integration and consumption. The four challenges center around the topics of query performance, data interlinking, data quality and user interaction. In some cases we provided future research directions to overcome these issues. We believe that the success in these research areas over the next few years is crucial for the Web of Data and its adoption by end users and enterprises.

[3]The following are pointers to published data sets:
http://www4.wiwiss.fu-berlin.de/eurostat/,
http://riese.joanneum.at/,
http://www.rdfabout.com/demo/census/,
http://www.govtrack.us/

Acknowledgement

We would like to thank the reviewers and the members of the LOD2 project consortium and in particular Orri Erling for the fruitful discussions, which contributed to this article.

References

[1] Harith Alani, David Dupplaw, John Sheridan, Kieron O'Hara, John Darlington, Nigel Shadbolt, and Carol Tullo. Unlocking the potential of public sector information with semantic web technology. In *Proceedings of ISWC/ASWC2007, Busan, South Korea*, volume 4825 of *LNCS*, pages 701–714. Springer, 2007.

[2] Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors. *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30–June 3, 2010, Proceedings, Part II*, volume 6089 of *Lecture Notes in Computer Science*. Springer, 2010.

[3] Sören Auer, Raphael Doehring, and Sebastian Dietzold. Less – template-based syndication and presentation of linked data. In Aroyo et al. [2], pages 211–224.

[4] Tim Berners-Lee. Linked data – design issues. web page, 2006.

[5] Christian Bizer and Andreas Schultz. The berlin sparql benchmark. *Int. J. Semantic Web Inf. Syst.*, 5(2):1–24, 2009.

[6] Roger Castillo, Christian Rothe, and Ulf Leser. Rdfmatview: Indexing rdf data for sparql queries. Technical report, Department for Computer Science, Humboldt-Universität zu Berlin, 2010.

[7] Abhirup Chatterjee and Arie Segev. Data manipulation in heterogeneous databases. *SIGMOD Record*, **20**(4):64–68, 1991.

[8] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, **19**(1):1–16, 2007.

[9] David Geer. Reducing the storage burden via data deduplication. *IEEE Computer*, **41**(12):15–17, 2008.

[10] Sebastian Hellmann, Jens Lehmann, and Sören Auer. Learning of owl class descriptions on very large knowledge bases. *Int. J. Semantic Web Inf. Syst.*, **5**(2):25–48, 2009.

[11] Afraz Jaffri. Linked data for the enterprise – an easy route to the semantic web. web page, March 2010.

[12] Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, and Amit P. Sheth. Linked data is merely more data. Technical Report SS-10-07, Menlo Park, California, 2010.

[13] Ee-Peng Lim, Jaideep Srivastava, Satya Prabhakar, and James Richardson. Entity identification in database integration. In *ICDE*, pages 294–301, 1993.

[14] Michael Martin, Jörg Unbehauen, and Sören Auer. Improving the performance of semantic web applications with sparql query caching. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *ESWC (2)*, volume 6089 of *Lecture Notes in Computer Science*, pages 304–318. Springer, 2010.

[15] Axel Polleres, Aidan Hogan, Andreas Harth, and Stefan Decker. Can we ever catch up with the Web? *Semantic Web – Interoperability, Usability, Applicability*, **1**(1,2):45–52, 2010.

[16] Sunita Sarawagi. Letter from the special issue editor. *IEEE Data Eng. Bull.*, **23**(4):2, 2000.

[17] John Sheridan and Jeni Tennison. Linking uk government data. In Christian Bizer, Tom Heath, Michael Hausenblas, and Tim Berners-Lee, editors, *Proceedings of the WWW2010 Workshop on Linked Data on the Web (LDOW 2010)*, 2010.

[18] Pavel Shvaiko and Jérôme Euzenat. A survey of schema-based matching approaches. *J. Data Semantics IV*, **3730**:146–171, 2005.

[19] Andreas Thor. *Automatische Mapping-Verarbeitung von Web-Daten*. Dissertation, Institut für Informatik, Universität Leipzig, 2007.

[20] Sebastian Tramp, Philipp Frischmuth, Timofey Ermilov, and Sören Auer. Weaving a social data web with semantic pingback. In *Proceedings of the EKAW 2010, Knowledge Engineering and Knowledge Management by the Masses; 11th October–15th October 2010, Lisbon, Portugal*, 2010.

[21] Sebastian Tramp, Norman Heino, Sören Auer, and Philipp Frischmuth. Making the semantic data web easily writeable with rdfauthor. In Aroyo et al. [2], pages 436–440.

[22] Jennifer Widom. Research problems in data warehousing. In *CIKM*, pages 25–30. ACM, 1995.

Semantic Web 1 (2010) 61–68
DOI 10.3233/SW-2010-0020
IOS Press

Towards a pattern science for the Semantic Web

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Pascal Hitzler, Wright State University, USA; Benjamin Adams, University of California, Santa Barbara, USA

Aldo Gangemi * and Valentina Presutti
Semantic Technology Lab, Institute for Cognitive Sciences and Technology, CNR, Via Nomentana 56, 00161, Rome, Italy
E-mail: {aldo.gangemi,valentina.presutti}@cnr.it

Abstract. With the web of data, the semantic web can be an empirical science. Two problems have to be dealt with. The *knowledge soup* problem is about semantic heterogeneity, and can be considered a difficult technical issue, which needs appropriate transformation and inferential pipelines that can help making sense of the different knowledge contexts. The *knowledge boundary* problem is at the core of empirical investigation over the semantic web: what are the meaningful units that constitute the research objects for the semantic web? This question touches many aspects of semantic web studies: data, schemata, representation and reasoning, interaction, linguistic grounding, etc.

Keywords: Knowledge patterns, cognitive science, frames, contexts, linked data

1. Introduction

Linked data [5] are creating the web of data, which on its turn can be considered the bootstrapping of the semantic web. For the first time in the history of knowledge engineering, we have a large set of realistic data, created by large communities of practice, on which experiments can be performed, so that the semantic web can be founded as an empirical science, as a branch of web science [6].

An empirical science needs clear *research objects*, e.g. cells, proteins, or membranes are types of research objects in different branches of biology. Such research objects, which can typically change and evolve (within different time scales), need to be shared by a community working on them. The community should also develop a language that is at least partly shared by its members and that is appropriate to describe those research objects. Based on these basic resources, a science develops procedures for making *patterns* emerge out of the research objects.

Until few years ago, the research objects of the semantic web used to be extracted from mostly small or toy examples, which had not the coverage and form that one can expect from data emerging from the use of the web by people and organizations. That coverage and form now exist, and are sometimes wild, but this is probably what an empirical science should deal with. Currently the web of data, including datasets such as DBpedia, geographical and biological data, social network data, bibliographical, musical, and multimedia data, etc., as well as the data emerging from the use of RDFa, Microformats, etc., has eventually provided an empirical basis to the semantic web, and indirectly to knowledge engineering.

There are two main problems (presented in Sections 1.1 and 1.2) for the identification, selection and construction of patterns from the empirical research objects of the semantic web. In Section 2 we provide some scenarios that exemplify the knowledge boundary problem, and in Section 3 we suggest a practical

*Corresponding author.

approach for making an empirical pattern science over the semantic web. Finally, we come back to the relation between the soup and the boundary problems, and discuss conclusion.

1.1. The knowledge soup problem

Several authors (e.g. [20]) have pointed out that the research objects in the web of data are just data, and one should ask what is really "semantic" in them, besides the usage of certain knowledge-oriented syntaxes like RDF. This is a variation of a problem spotted by AI scientists years ago (e.g. [24]), called the *knowledge soup* problem (Section 1.1): since people maintain and encode heterogeneous knowledge, how can formal knowledge be derived from the soup of triplified data (this is also related to the "reengineering bottleneck", see Hoekstra in this special issue)?

The web of data is a knowledge soup because of the heterogeneous semantics of its datasets: real world facts (e.g. geo data), conceptual structures (e.g. thesauri, schemes), lexical and linguistic data (e.g. wordnets, triples inferred from NLP algorithms), social data about data (e.g. provenance and trust data), etc. The authors of [20] envisages a situation where those datasets are formally represented, e.g. through the semantics that would derive from an alignment to DOLCE[1]. In that way (by means of appropriate reengineering practices) we can imagine that conceptual structures are represented as classes or properties, real world facts as individuals or assertional axioms, lexical data into annotations, etc. They also propose to remove the inconsistencies that can derive from the alignment.

1.2. The knowledge boundary problem

The second problem we have singled out is the *knowledge boundary* problem: how to establish the *boundary* of a set of triples that makes them meaningful, i.e. relevant in context, so that they constitute a knowledge pattern? and how the very different types of data (e.g. natural language processing data, RDFa, database tables, etc.) that are used by semantic web techniques contribute to carve out that boundary?

Patterns in general can be defined as *invariances across observed data or objects*. The patterns in the semantic web emerge from data, but we need to distinguish the symbolic patterns of mathematical pattern science [19], as studied in data mining, complex systems, etc., from the knowledge patterns of a semantic web pattern science. Knowledge patterns are not only symbolic patterns: they also have an interpretation, be it formal, or cognitive. Such interpretation consists in the *meaning* of the pattern, e.g. a fact reported in news, a soccer event in a picture, an aggressive attitude in a sentence, a subtle plan revealed by the analysis of a set of documents. In practice, a meaningful pattern to be discovered in the web of data has to be *relevant in a certain context*; we need a notion of *boundary* for sets of triples that matter. How many semantic applications address relevance in context explicitly? How many of them succeed in achieving it, solving problems that matter to anyone?

2. Some scenarios and their requirements

Being meaningful is usually associated with relevance in context, i.e. having a clear boundary in order to matter to someone.

In this section, we exemplify this concept and highlight the importance to have a semantic web able to recognize, handle, and exploit such boundaries.

For example, consider an application that leverages the web of data to provide information on some topic. Tools of this sort, such as Sig.ma[2] already exist. The information that Sig.ma or similar tools can collect, *mesh up*, and serve is much easier to consume than the typical results of a traditional search engine such as Google. This is due to the form of information handled: data as opposed to documents. Nevertheless, in order to establish which of the retrieved data are relevant in the context that matters to the user is still an issue. For example, consider the situation of searching information about a person e.g. Aldo Gangemi, who teaches as tutor in a school for researchers. Figure 1 shows the difference between all data that are collected from the web of data, and those that are relevant for the specific purpose. Figure 1(a) shows the data collected about "Aldo Gangemi": it can be noticed that such data contain information about Aldo's favorite music, civil status, political views, hobbies, etc., which are not relevant for his role of tutor at the school. Among all such information only those depicted in Fig. 1(b) would solicit some interest in this context. Notice that the selection of relevant information impacts at both the level

[1] http://www.ontologydesignpatterns.org/ont/dul/DUL.owl

[2] http://sig.ma/

(a) Data about Aldo Gangemi collected and meshed up from the web of data.

(b) Relevant data about Aldo Gangemi as tutor, collected and meshed up from the web of data.

Fig. 1. The difference between data about Aldo Gangemi collected and meshed up from the web of data, and the selection on such data of what is relevant for a specific context i.e. looking for information about the tutor of a school for researchers.

of properties and the level of property values e.g. in this context it could be relevant to provide the information that Enrico Motta, who is a researcher, is in Aldo's contact network, while it is less important to mention that the network includes Aldo's mother. How could relevance boundaries for this context be identified? Another clarifying example is the situation of a query aimed at identifying a more complex entity than a person. Consider the situation of a talk involving a discussion about two politicians belonging to different periods. In this case the application should be able to extract and interpret relevan knowledge by exploring the relationships that hold between the two politicians. One such application is RelFinder[3], a tool based on the web of data that extracts and visualizes relationships between objects in datasets, and to make these relationships interactively explorable. Figure 2 depicts a fragment of the result produced by RelFinder for "Benito Mussolini" and "Silvio Berlusconi". Despite the extent of the data about the two politicians available on the web of data, the similarity between them that RelFinder discovers is shallow and somehow irrel-

[3]http://relfinder.semanticweb.org/

Fig. 2. Relations between Benito Mussolini and Silvio Berlusconi extracted from the web of data by the RelFinder application.

evant to most people. For example, it emerges that the two men have been both Prime Minister of Italy but the relevant knowledge that would matter to the user in this context would be the possible similarity of approaches that they followed, and the attitude that they have shown during their respective careers. Hence, the issue is still how to establish what are the boundaries that identify the relevant knowledge for a specific context. It is even clearer to understand the issue, and its importance for the semantic web, if we consider that a key goal of the semantic web is to drastically decrease the cognitive load of humans when performing some task by delegating it to smart software agents. Consider the situation of a user that plans to arrange a holiday trip to San Francisco. In order to solve this task the user considers personal preferences e.g. on hotels, personal past traveling experiences, working calendar, available budget, etc., and compares them to what can be extracted from the information gathered from the web. A common task like this requires a significant amount of time, and cognitive effort. Additionally, the capability of identifying relevant contextual information in order to produce e.g. trip arrangement proposals, is key for enabling automatic support that goes beyond information or data retrieval.

The current semantic web and its frontline web of data branch are far from enabling support for tasks like this, although it enables the interpretation and consumption of data in the form of knowledge. We need a new paradigm, which enables a move from representing knowledge in general to shaping knowledge to be represented for a certain task, i.e. meaningfully.

3. Frames and knowledge patterns

A research paradigm for the semantic web should include at least two key foundational, interleaving aspects: (i) a *unit of meaning* for the semantic web, more complex and effective than scattered classes or properties (binary relations), and (ii) a multidimensional context model.

We believe that the unit of meaning of the semantic web should be the *knowledge pattern*, a special name for *frames* [4,21] in the semantic web, and that a formal context model should address the distinction between four dimensions i.e. descriptive, informational, situational, and social, besides the formal dimension.

3.1. Frames

A frame is a pattern whose intuition goes back to the notion of *schema* by [2][4]. Many-flavoured varieties of schemata or frames have been proposed later by [3,11,12,17,21,23], etc. The intended meaning of a frame across the different theories can be summarized as "a structure that is used to organize our knowledge, as well as for interpreting, processing or anticipating information".

Different theories highlight the static vs. dynamic behavior of frames, their adaptability, or even people's ability to create them on the fly. Some older studies cast doubts on their role in e.g. scene recognition [7], but besides the huge positive literature in linguistics and cognitive science, recent biological evidence seems to be quite convincing for considering them more than a philosophically-fascinating hypothesis. For example, [1] found a clear congruence between activations of visually presented actions and of actions described by literal phrases. These results suggest a key role of mirror neuron areas in the re-enactment of sensory-motor representations during conceptual processing of actions invoked by linguistic stimuli. These findings support *embodied semantics* hypotheses, in which frames are the core unit of meaning, as event-oriented, embodied structures that abstract basic sensory-motor competences acquired by cognitive agents (see [14] for an overview of embodied semantics applied to ontologies).

Marvin Minsky [21] introduced frames into computer science, claiming that "*there would be large advantages in having mechanisms that could use these same structures both for thinking and for communicating*". This means that the cognitive notion of a frame should have *counterparts* into the computational world. [21] exemplified counterparts in the form of modeling, programming, and interaction schemata. That was very successful, since the frame metaphor has been used as a formal schema in frame logics (frames involving closed world assumption) and description logics (concepts); as a design structure in object oriented design (classes); as an interaction design pattern in human-computer interaction (templates), etc. Therefore, a cognitive frame is an embodied structure that can be *partly represented* with constructs from languages with different formal semantics: as a polymorphic intensional relation, as a F-Logic frame, as an OWL class, etc.

However, the cognitive origin and motivation of Minsky's proposal were abandoned during the evolution of knowledge representation and engineering, because other scientific problems, such as computational complexity and formal semantic foundations, overruled the original agenda, which demanded a lot on the *design* side rather than on representation and reasoning. We believe it is time to resurrect that agenda.

Based on the above considerations, we suggest the usage of frames as the primary research objects over the semantic web, as opposed to simple concepts or binary relations, and we call them *knowledge patterns*. Such notion must be pragmatic, in order to provide a meaning unit that acts as a "hub" between requirements for semantic applications, reusable ontologies, data to be queried, patterns in indexed texts, interaction patterns for semantic data, etc. The benefits of knowledge patterns can be several: easier ontology design, advanced exploration of data, extraction, as well as lenses over data, query patterns, rich linguistic grounding of ontologies for hybrid applications, etc.[5]

In the next section, we briefly present a model of the different aspects (data, language, interaction, etc.) of a knowledge pattern as revealed in the complex semiotic activity of cognitive agents.

3.2. A research framework for knowledge patterns

Over the semantic web, agents and reasoners should be able to discover and/or recognize knowledge pat-

[4]The genealogy of Bartlett's schemata goes back to Kant's distinction between objects of perception and their interpretation by an agent, and the need to postulate background schemata that enable interpretation over pure perception.

[5]Some of them have been experimentally demonstrated [8].

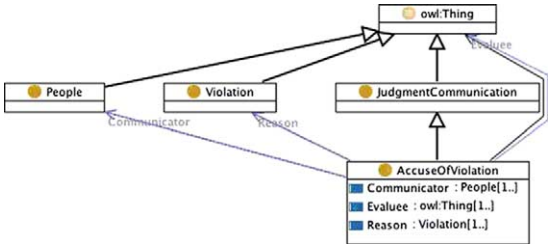

Fig. 3. An OWL content pattern learnt from a text corpus, based on a FrameNet frame and machine learning techniques. It models the relevant entities participating in an *accuse of violation*, a special case of the *judgement communication* frame in FrameNet [4], learnt as described in [9].

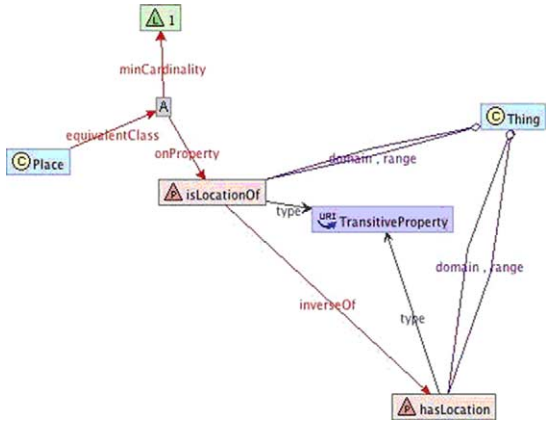

Fig. 4. An OWL content pattern from http://www.ontologydesignpatterns.org (ODP), manually defined. It models places, their characteristics, and the relationships between them e.g. the transitive property `hasLocation`.

terns (KP), and reason on them. Examples of KP representations that are already reused within the semantic web are mentioned in Section 3.3. A KP derived from a FrameNet [4] frame is depicted in Fig. 3[6], while a native semantic web KP (a content pattern[7]) is depicted in Fig. 4. Currently, technology and languages allow us to find occurrences of KPs that we already know. For example, we can use SPARQL for querying the linked data cloud in order to find occurrences of the KP depicted in Fig. 4. However, how to discover new meaningful KPs is still an open issue. How to decide a boundary within the giant graph of LOD? What topology should be adopted? How to hybridize different data? While some of these questions are left

[6]A comprehensive discussion on how to transform linguistic frames to knowledge patterns, and how to enrich them is [9].

[7]See http://www.ontologydesignpatterns.org and [22] for more details on content and other ontology patterns.

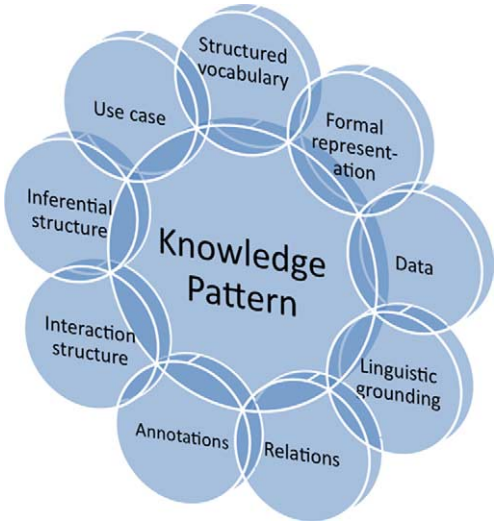

Fig. 5. The façades of a knowledge pattern.

to more extensive work for the joy of the semantic web community, the following proposal provides a unified model for describing the different aspects of an empirical research for KPs over the semantic web.

A KP can be modeled as a polymorphic relation that takes arguments from a number of *façades*, as depicted in Fig. 5. A façade represents a type of knowledge that can be associated with a frame, and can be used to motivate, test, discover, and use it. We use an elementary KP that models "persons with a role in a research group" for exemplifying some typical façades of a KP.

- *Vocabulary*: a set of terms that can be structured with informal relations, for example, for a KP about researchers, the following set of terms could be activated: {Person, Role, ResearchInterest, ResearchGroup, Name}
- *Formal representation*: axioms that provide a formal semantics to the vocabulary. For example (in OWL):
  $\texttt{Person} \sqsubseteq \exists \texttt{hasRole.Role} \wedge \exists \texttt{hasTopic.ResearchInterest} \wedge \exists \texttt{memberOf.ResearchGroup} \wedge =1\texttt{hasName.Name}$
- *Inferential structure*: rules that can be applied to infer new knowledge from the formal representation of the KP, e.g.:
  $\texttt{similarInterest(?p1 ?p2)} \leftarrow \texttt{hasTopic(?p1 ?i), hasTopic(?p2 ?i)}$
- *Use case*: requirements addressed by the KP. They can be expressed in various forms e.g. including one or more competency questions [15];

in our lead example a competency question could be: *What PhD students from a research group have a certain research interest?*
- *Data* that can be used to populate an ontology whose schema is a formal representation for the KP
- *Linguistic grounding*: textual data that express the meaning of the KP, e.g.: *The AST group has developed significantly in the last year. Professor João spawned AST interests from theoretical work on strong AI to applications by making an agreement with the WQP software engineering lab*
- *Interaction structure*: mappings between elements in the formal representation of a KP, and interface or interaction primitives, e.g.: `Person` ≈ `Static container`, `Role` ≈ `Container features (color, size)`, `Research group` ≈ `Drop down list`, `Research interest` ≈ `Static container`, `Relation` ≈ `Link` | `Containment`
- *Relations* to other KPs, e.g.: `hasComponent {personrole, collection, topic}`
- *Annotations*: provenance data, comments, tags, and other informal descriptions not yet covered by any existing façade.

In an empirical research perspective, the different façades provide research objects, which can be RDF triples in the formal and data façades, texts to be indexed or parsed, data in different formats that can be reengineered according to specific needs, etc. Each of them is the result of a particular approach for representing our knowledge. For example, evidence of a known KP, or the discovery of a new one, can be the result of machine learning techniques applied to large corpora (cf. [9]), of defining RDF named graphs, of reengineering existing data models or structures, or of the harvesting of formalized ontologies.

The KP framework presented here is an invitation to start empirical work for a knowledge pattern science that has cognitive objects as primary research objects (see also Raubal and Adams in this special issue), and that provides a unifying framework to all diverse approach for representing knowledge. Although it is reasonable to envision an implementation of the KP framework, semantic web research can also take advantage from it as a research model, and individual KPs can be used for annotating results of data integration, pattern discovery or detection, etc., and eventually to make them converge for specific projects or applications.

3.3. Partial realizations of knowledge patterns

Existing projects that (partially) realize the KP research model include: the Component Library [10] encoded in the KM language, which realizes the vocabulary as well as formal façades; the FrameNet project [4], which realizes the vocabulary and linguistic grounding façades; Microformats, which realize the data and vocabulary façades; the Ontology Design Patterns project, which includes several types of design patterns: content patterns (realizing the use case, vocabulary and formal façades) [22], logical patterns (realizing types of formal façades), reengineering patterns (targeting good practices in mapping vocabulary to formal façades, or formal to formal), etc. Ontology design patterns also highlight some similarities between frames and "design patterns" as employed in software engineering [16].

Besides projects that explicitly address frames or patterns, we remark that a lot of semantic web research and applications already implicitly address the typical data involved in some KP façades. The difference is that they lack of explicit use of KPs as research objects, therefore the patterns that eventually emerge in e.g. representation and reasoning, data reengineering, linked data, etc. do not necessarily address cognitively meaningful structures. Moreover, those heterogeneous patterns are usually disconnected from each other. In such situations, KPs can play the role of "attractors" for the diverse patterns: linguistic, data, axiom schemata, reasoning-oriented, etc. by providing a unified research model that supports integration between those diverse, although complementary results.

3.4. Knowledge patterns and multidimensional contexts in the soup

As mentioned in Section 1.1, data on the web is affected by the knowledge soup problem, and needs to be cleaned up when making advanced reasoning on it. The solution proposed by [20] is to align it to foundational ontologies, and to perform inconsistency debugging. While data cleaning through alignment to foundational ontologies can be feasible and desirable to some extent, the removal of inconsistencies can be hardly sustainable. It would require the removal of triples that express possible relevant knowledge, hence distorting the original intentions of data curators. Rather than removing them, we might want to live with inconsistencies by isolating them when a consistent reasoning pipeline is needed.

The choice of a foundational ontology is key in this process as it heavily impacts on the effects of reasoning. In order to make such choice effective two key aspects have to be addressed.

Task. The coverage and axiomatic complexity of the foundational ontology should be tailored to the task at hand, and alternative solutions may co-exist. E.g. reasoning on sequence-related axioms (temporal, spatial, scheduling) is different from reasoning on participation-related axioms (events, situations). Using all of them at the same time is not mandatory. Using a unique ontology for all of them is not mandatory either. Knowledge patterns such as ontology design patterns [22] are a blessing for this criterion, since they can be used to plug and play with manageable subsets of axioms and aligned data, which are related to relevant use cases.

Context. The data in the soup implicitly assume different domains of interpretation, or "contexts": e.g. WordNet datasets live in an informational (linguistic) domain of discourse, FOAF profiles live in a social context, most DBpedia and geographic datasets come from a situational domain (they are bare "facts"), a lot of DBpedia and biological data are about conceptual entities, etc. To make sense of these data, formal semantics alone is not enough. We need something that supports multidimensional interpretations of data linked across different, sometimes logically incompatible contexts: descriptive (or conceptual) context, informational context, situational context, social context, and formal context.

Based on the above reasons, some of the types and axioms of the foundational ontology should cover the knowledge contexts that make up the soup, typically individuals, facts, concepts, information objects, and social metadata.

The research on KPs exemplifies the typical contexts that are mixed up in the soup. The vocabulary, linguistic grounding, formal representation, and definition-oriented annotation façades of a KP contain data that exemplify the *descriptive context* of knowledge; the linguistic grounding façade exemplifies the *informational context*; the interaction façade and provenance annotations exemplify the *social context*; the data façade exemplifies the *situational context*; the formal and inferential façades exemplify the *formal context*.

A foundational ontology that formally represents such different contexts is "Constructive Description and Situation" (c.DnS) [13,18]. It leaves it open the choice about controversial distinctions such as objects and events vs. three-dimensional entities, qualities vs. values, etc., which are typically fixed by most upper-level and foundational ontologies. The basic machinery of c.DnS involves a strict separation of the domains for the different context types, and includes appropriate relations between contexts of the same or different type. Several content ontology design patterns encode parts of c.DnS for its use within the KP paradigm. For space reasons, we redirect the reader to the cited literature for further details.

4. Conclusion

In this paper we argue that the semantic web can be an empirical science based on a new paradigm built upon two foundational aspects: (i) the *knowledge pattern* as a unit of meaning for the semantic web, and (ii) a multidimensional context model able to capture its descriptive, informational, situational, social, and formal characters. The current situation is that only triples and named graphs on one side, and large or complicated ontology schemata on the other side, are de facto research objects: in neither case they are close to the way knowledge is contextually relevant for people. We have suggested *knowledge patterns* (KPs) as a primary research object to focus on. KPs both reflect the intuition of frames on the semantic web and provide the structure needed for representing the different context dimensions.

Acknowledgements

This work has been part-funded by the EC under grant agreement FP7-ICT-2007-3/ No. 231527 (IKS - Interactive Knowledge Stack). We would like to thank Giovanni Pezzulo, Eva Blomqvist, Alfio Gliozzo and the STLab[8] for the nice and useful discussions on knowledge patterns.

References

[1] Lisa Aziz-Zadeh, Stephen M. Wilson, Giacomo Rizzolatti, and Marco Iacoboni. Congruent Embodied Representations for Visually Presented Actions and Linguistic Phrases Describing Actions. *Current Biology*, **16**(18):1818–23, 2006.

[8] http://stlab.istc.cnr.it

[2] Frederic C. Bartlett. *Remembering: An Experimental and Social Study*. Cambridge University Press, Cambridge, 1932.

[3] Lawrence W. Barsalou. Perceptual Symbol Systems. *Behavioral and Brain Sciences*, **22**:577–600, 1999.

[4] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada, 1998. Association for Computational Linguistics.

[5] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*, **5**(3):1–22, 2009.

[6] Tim Berners-Lee, Wendy Hall, James A. Hendler, Nigel Shadbolt, and Daniel J. Weitzner. Web Science. *Science*, 313, August 2006.

[7] Irving Biederman, Robert J. Mezzanotte, and Jesse C. Rabinowitz. Scene Perception: Detecting and Judging Objects Undergoing Relational Violations. *Cognitive Psychology*, **14**(2):143–177, 1982.

[8] Eva Blomqvist, Valentina Presutti, Aldo Gangemi, and Enrico Daga. Experimenting with eXtreme Design. In Philipp Cimiano and Sofia Pinto, editors, *Proc. of the Conference on Knowledge Engineering and Knowledge Management (EKAW2010)*, Galway, Ireland, 2010. Springer.

[9] Bonaventura Coppola, Aldo Gangemi, Alfio Gliozzo, Davide Picca, and Valentina Presutti. Frame Detection over the Semantic Web. In *Proceedings of the Fifth European Semantic Web Conference*. Springer, 2009.

[10] Peter Clark and Bruce Porter. Building Concept Representations from Reusable Components. In *Proceedings of AAAI'97*, page 369–376. AAAI press, 1997.

[11] Peter Clark, John Thompson, and Bruce Porter. Knowledge Patterns. In Anthony G. Cohn, Fausto Giunchiglia, and Bart Selman, editors, *KR2000: Principles of Knowledge Representation and Reasoning*, pages 591–600, San Francisco, 2000. Morgan Kaufmann.

[12] Charles J. Fillmore. The Case for Case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–210. Holt, Rinehart, and Winston, New York, 1968.

[13] Aldo Gangemi. Norms and Plans as Unification Criteria for Social Collectives. *Journal of Autonomous Agents and Multi-Agent Systems*, **16**(3):70–112, 2008.

[14] Aldo Gangemi. *What's in a Schema?* Cambridge University Press, Cambridge, UK, 2010.

[15] Michael Gruninger and Mark S. Fox. The Role of Competency Questions in Enterprise Engineering. In *Proceedings of the IFIP WG5.7 Workshop on Benchmarking – Theory and Practice*, 1994.

[16] Erich Gamma, Richard Helm, Ralph E. Johnson, and John Vlissides. *Design Patterns: Elements of Reusable Object-Oriented Software*. Addison-Wesley, Reading, MA, 1995.

[17] Vittorio Gallese and Thomas Metzinger. Motor Ontology: the Representational Reality of Goals, Actions and Selves. *Philosophical Psychology*, **16**(3):365–388, 2003.

[18] Aldo Gangemi and Peter Mika. Understanding the Semantic Web through Descriptions and Situations. In *CoopIS/DOA/ODBASE*, pages 689–706, 2003.

[19] Ulf Grenander. *Elements of Pattern Theory*. Johns Hopkins University Press, Baltimore, 1996.

[20] Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, and Amit P. Sheth. Linked Data is Merely more Data. In *Proceedings of the AAAI Symposium Linked Data Meets Artificial Intelligence*, pages 82–86, 2010.

[21] Marvin Minsky. A Framework for Representing Knowledge. In P. Winston, editor, *The Psychology of Computer Vision*. McGraw-Hill, 1975.

[22] Valentina Presutti and Aldo Gangemi. Content Ontology Design Patterns as Practical Building Blocks for Web Ontologies. In *ER'08: Proceedings of the 27th International Conference on Conceptual Modeling*, pages 128–141, Berlin, Heidelberg, 2008. Springer-Verlag.

[23] Jean Piaget. *Six psychological studies*. Random House, New York, 1967.

[24] John F. Sowa. The Challenge of Knowledge Soup. In J. Ramadas and S. Chunawala, editors, *Research Trends in Science, Technology and Mathematics Education*, Mumbai, 2006. Homi Bhabha Centre for Science Education.

Semantic Web 1 (2010) 149–153
DOI 10.3233/SW-2010-0021
IOS Press

Ontology use for semantic e-Science

Editors: Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited Reviews: Michel Dumontier, Carleton University, Canada; Manfred Hauswirth, DERI, National University of Ireland, Ireland

Boyan Brodaric[a,*] and Mark Gahegan[b]
[a] *Geological Survey of Canada, 234B – 615 Booth St., Ottawa, Ontario, Canada, K1A 0E9*
[b] *Centre for eResearch and School of Environment, University of Auckland, Human Sciences Building, 10 Symonds St., Auckland, New Zealand*

Abstract. Ontologies are being widely used in online science activities, or e-Science, most notably in roles related to managing and integrating data resources and workflows. We suggest this use has focused on enabling e-Science infrastructures to operate more efficiently, but has had less emphasis on scientific knowledge innovation. A greater focus on online innovation can be achieved through more explicit representation of scientific artifacts such as theories and models, and more online tools to enable scientists to directly generate and test such representations. This should lead to routine use of ontologies by scientists, and foster new and potentially different scientific results to help usher in next generation e-Science.

Keywords: Ontology use, e-Science, semantic web

1. Introduction

Before the onset of scientific computing, the data, methods and theory used for science were often kept close together, in the head and notebook of the researcher. Development of computational infrastructure over the last 50 years has allowed first data and next methods to move far from their scientific creators. Many research communities are now congregating around online infrastructures that contain shared repositories of primarily data and methods. Such infrastructures are being used for the discovery, retrieval, and integration of online scientific resources, mainly scientific databases, and increasingly also to capture and describe scientific instruments, software, workflows, and experiments. These infrastructures and associated activities collectively comprise e-Science [6]. The number of e-Science initiatives is vast. Some examples are:

- The Geosciences Network: *GEON* (www.geongrid.org) [10]
- Cancer Biomedical Informatics Grid: *caBIG* (https://cabig.nci.nih.gov/) [13]
- Global Ocean Observing System: *GOOS* (http://www.ioc-goos.org/) [2]

These and similar efforts are realizing important scientific benefits, which we claim can be largely attributed to three factors: improvements in resource quantity, improvements in representation, and improvements in communication:

(1) **Improvements in resource *quantity*** are realized by leveraging and integrating greater numbers of relevant online resources. New results ensue when more and bigger online assets are brought to bear on a problem, for example, such as when distributed computing is running remote applications, often automated and in parallel, over networks of massive databases or sensors. Data that is often expensive to capture or create is then more likely to see secondary use. The same goes for methods and other e-resources.

* Corresponding author. E-mail: brodaric@nrcan.gc.ca.

(2) **Improvements in *representation*** are realized by recording more complete and complex expressions of scientific knowledge as well as related research activities. More reliable results then accrue from far better levels of repeatability and explanation, because online environments host machine-processable representations of many (ideally all) aspects of scientific investigation, and these can be accessed by greater numbers of scientists.

(3) **Improvements in *communication*** are realized by facilitating deeper and more frequent online collaboration between scientists. The enhanced connectivity of online science environments then increases the exchange of ideas.

Ontologies are already playing a pivotal role in these areas. For example, in virtual observatories [8] ontologies are: (1) being used to annotate the structure and content of scientific databases and workflows to make them interoperable, (2) helping guide the structure and content of scientific workflow provenance to illuminate scientific reasoning [14], and more generally they are (3) facilitating scientific discourse by providing content and context for online dialog in virtual communities.

However, these improvements are mainly impacting the online use of scientific data and methods, while the surrounding knowledge, the theory, assumptions, reasoning and other context, have largely been left behind. This is highlighted by the position of ontologies in the infrastructures where they are frequently shuffled to the background. Indeed, ontologies are rarely used directly by scientists, despite the potential for them to help represent knowledge that might otherwise seem to be absent. Instead, they are more often directly used by computers to enable automated components of the infrastructure to work properly. This raises outstanding questions about how effectively ontologies are being used to innovate knowledge from their background position in the infrastructure.

We suggest that ontologies are underutilized in the development of new scientific knowledge in each of the three aspects above. This is largely due to the fact that—for the most part—ontologies are being treated as engineering artifacts required to execute tasks more efficiently, rather than knowledge artifacts that, for example, help to describe some gap in scientific theory or flaw in the reasoning. Indeed, we claim e-Science ontology use is at present largely motivated by operational efficiency, with downstream impacts on scientific knowledge development minimized at present, and significantly below their potential. A contrasting vision prioritizes knowledge innovation in which scientists use ontologies both to express hypotheses, theories and models, and also to generate and test them [4,18]. In this aspirational vision, scientists use ontologies directly as part of routine scientific investigation because the e-Science environments are designed to facilitate this. Such direct scientist interaction with the ontology-enabled knowledge, i.e. 'in-silico' semantic science, should then help revitalize online scientific methodology by helping generate richer insights, and improving our ability to repeat, report, and validate scientific findings.

2. Resource quantity

The focus on operational efficiency is best exemplified by the quantity aspect (described in 1 above), in which significantly more online resources can be marshaled and then applied to some task. This usually involves ontology-enabled semantic interoperability to connect greater volumes of data, software, instruments, and computing resources. The associated ontologies typically consist of application ontologies that describe particular resources, or a slightly more general domain ontology that spans the application ontologies and serves as a unifying conceptualization for the system [16,21]. However, neither of these ontology types typically encapsulates broad domain knowledge, as each tends to include only those concepts needed to enable the interoperability of specific resources. These ontologies are seldom even seen by scientists and they mainly remain part of black-box components that allow the system to automatically handle greater volumes of resources than could perhaps be handled manually. Even when the ontologies are seen by scientists, for example in query interfaces used to search distributed databases, the focus is on efficient retrieval of resources. Knowledge innovation is thus tied to insights gleaned by scientists from greater and faster resource retrieval and integration, rather than deeper understanding of the resources. Arguably, this often does not involve the application of new online scientific methods, but rather the mirroring of manual methods within the online environment, such that conventional lines of reasoning are carried out online by scientists. While this is certainly leading to new scientific results, there remains the real possibility that dramatic new insights might be achieved with complementary lines of investigation that involve increased use of machine techniques

related to learning, analogical and abductive reasoning, data mining, and so on, that are starting to be applied to scientific discovery [5,13].

The contrasting vision would thus leverage ontologies and more automated methods to facilitate the proposal of new hypotheses as well as offer mechanisms to test their validity. The role of the scientist is not diminished but the system plays a greater and more direct role in knowledge innovation as some tasks are automated, like resource comparison and evaluation, and as new avenues of investigation are recommended to the scientist. Ontologies also play a greater role, because as authoritative representations of domain knowledge they become key expressions of the inputs and outputs of the research, which causes them to be consulted and updated regularly. The ontologies then constitute a far richer knowledge repository for a domain, and consist of theories, models, methods, and other artifacts of scientific work. This is a significant advance on present ontology contents, which primarily contain scientific categories such as 'granite', 'mass', 'temperature', and 'melanoma'. It is also a significant shift in ontology use as ontologies would be deployed directly by scientists, as well as machines, in all stages of knowledge discovery.

3. Representation

The representation aspect is probably best exemplified by the role of ontologies in online scientific provenance [11,19], where ontologies are used to represent many aspects of scientific investigation. Scientific provenance refers to the historical context surrounding some scientific activity or result, and typically involves a description of the methods and applications used, the processes and reasoning steps carried out by a scientist for some purpose, and the old as well as new states of knowledge and data [20]. It is most widely encountered in established scientific workflow environments, such as myExperiment [7], which orchestrate scientific processing and from which the provenance elements can be readily obtained. While traditional best practices would necessarily have such process information recorded manually, online environments allow this to happen transparently by recording each operation as it occurs, and also recording it more finely so that each step can be captured, repeated and questioned.

Ontologies are widely used in provenance systems. They serve as common conceptualizations in the query interface for viewing and querying provenance, and for semantic interoperability across various data and provenance stores [11]. They are also used to annotate metadata associated with components in a scientific workflow [10,11], and underpin trust systems that evaluate the quality and reliability of a scientific resource [2]. However, as with the quantity factor, such ontology use primarily has an efficiency imperative: more often than not the ontologies are used to describe low-level system resources such as a web service interface or a specific data product, rather than scientific objectives such as the hypothesis being tested or the reasoning used. The ontologies are thus primarily used to make the provenance system work, but how this affects knowledge generation is left to the scientist to determine. Our vision of provenance extends these notions to include ontologies of scientific method and reasoning, such that online processing steps can be understood in terms of scientific objectives, for instance to verify a result or evaluate a hypothesis. A particular workflow could thus be described in terms of system operations as well as scientific reasoning steps, so that scientists could interact with the workflow in terms of scientific goals as well as system mechanics. This necessarily involves a conceptualization of the general science knowledge cycle as well as effective interfaces and functions to operate over it.

4. Communication

The communication factor is best exemplified by online scientific collaboratories in which scientists utilize multi-media and social networking resources to work together on common tasks [17]. The general intent is scientific progress through increased scientific interaction, with a particular focus on augmented and clearer online discourse. Tools to represent and search scientific discourses are usually coupled to literature repositories or other resources, which provide subject matter for the discourse. Ontologies are used to represent concepts inherent in the discourse, including discourse concepts and scientific domain concepts, and these are often realized as annotations to papers in the literature repositories. The emphasis, though, is on the nature of the rhetoric surrounding some knowledge [3,14] and on the validity of a given line of reasoning typically within a descriptive logic, with far less focus on the representation and evolution of higher-order scientific concepts such as theories and models. Again, this can be largely viewed as

a gain in efficiency in that scientific statements indexed against rhetorical or basic domain concepts can be more readily found, likely in semantically annotated repositories [15], and more scientists are able to collaborate more often. It can also be viewed as a marginal gain in knowledge interpretation as the knowledge is parsed into relatively simple structures, which nevertheless are critically evaluated such that inconsistencies in the reasoning are identified and conceptual gaps are highlighted. Critical evaluation might include the proposal of new hypotheses, but discourse systems on their own do not enable those hypotheses to be tested in a scientific sense, against data using established methods; at least not without being coupled to additional resources such as workflows, databases, instruments, and so forth. Our vision would see that coupling take place, such that dynamic hypothesis generation and testing could occur on deeper knowledge structures during online scientific discourse, where it could be tracked as well as evaluated for trust and eventual re-use.

5. Challenges

The vision of a scientific semantic web, in which ontologies drive science knowledge discovery, comes with many significant challenges related to capturing, designing, and using ontologies:

(1) **Ontology capture**: although some domains such as biomedical are routinely evolving ontologies, the vast bulk of science knowledge exists in growing literature repositories from which ontologies are absent and must be captured. At present, existing automated and semi-automated techniques for ontology extraction are limited to the capture of relatively simple science concepts and shallow structures explicit in the text, such as domain terms and large rhetorical blocks. A serious challenge is the capture of complex concepts and deep structures often implicit in the text, such as theories and lines of reasoning, and automation of this capture to deal with the large volume of source material. However, it is likely that techniques to capture knowledge as it develops within workflows will be more effective than those geared towards extraction from texts produced after some experiment has been completed, because the former contains more sources of context and more opportunities for direct interaction with the researchers. The design, management, and interoperability of science knowledge repositories is a related concern.

(2) **Ontology design**: challenges for ontology design include the development of guidelines, design patterns, and formal methods for the construction and evaluation of ontologies within and across science domains. At present, general ontology engineering approaches are being successfully adapted to help domain ontology construction, but largely without recourse to general knowledge elements common to science. The main hurdle to overcome involves tuning these established techniques specifically to science domains, taking into account commonalities and differences. This further requires careful work to build general ontologies of science, which should lead to more consistent and coherent ontologies within domains, and facilitate connectivity across domains by providing a unifying upper-level of generic concepts such as theory, data, model, induction, method, experiment, and so on. Significant challenges also abound concerning how scientists might collaborate on the development of these elements, such as theories that are shared, overlapping, or in conflict, and how online resources can aid in the resolution of knowledge disputes.

(3) **Ontology use**: perhaps the holy grail of semantic e-Science is the quest for online (semi-) automated knowledge discovery. This requires a combination of human and computer methods to analyze and compare ontology-driven knowledge elements, such as theories and models, and to propose and test knowledge gaps. Coordinating deductive, inductive, and abductive reasoning in workflows operating over distributed online resources is an important part of this challenge. The development of user-friendly query and browsing interfaces attuned to a large-scale science knowledge framework is another significant challenge that must be overcome to ensure the framework will be usable.

6. Conclusions

Our conception of semantic e-Science amalgamates the enhanced visions discussed above. It includes semantic repositories of knowledge in which ontologies are a base representation for scientific concepts, theories, models, methods, and other science knowledge elements. These are coupled to workflow operations driven by scientific objectives and methods, and to scientific provenance described in terms of scientific reasoning steps. Finally, scientific collaboratories enable community discourse to

occur over any of the previously mentioned components, to evaluate them for quality, trust, veracity and re-usability. In such an online environment scientists would focus on knowledge innovation, in as transparent a way as possible, harnessing both efficiency and innovation objectives for next generation science.

Acknowledgements

We gratefully acknowledge of the support of the Geological Survey of Canada and the Ministry of Research, Science and Technology, New Zealand, and kindly thank the editors and reviewers for their suggestions which led to improvements in the manuscript: Krzyszof Janowicz, Pascal Hitzler, Michel Dumontier, and Manfred Hauswirth.

References

[1] Alverson, K. (2008). Filling the gaps in GOOS. *The Journal of Ocean Technology*, **3**(3):19–23.

[2] Artz, D., Gil, Y. (2007). A Survey of Trust in Computer Science and the Semantic Web. *Journal of Web Semantics*, **5**(2):58–71.

[3] Blake, C. (2010). Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles. *Journal of Biomedical Informatics*, **43**(2):173–89.

[4] Brodaric, B., Reitsma, F., Qiang, Y. (2008). SKIing with DOLCE: toward an e-Science Knowledge Infrastructure. In: Eschenbach, C., Gruninger, M., (Eds.) Formal Ontology in Information Systems, *Proceedings of the Fifth International Conference (FOIS08)*, IOS Press, pp. 208–219.

[5] Colton, S., Steel, G. (1999). Artificial Intelligence and Scientific Creativity. *Artificial Intelligence and the Study of Behaviour Quarterly*, Vol. 102, 1999.

[6] De Roure, D., Gil, Y., Hendler, J. (2004). E-Science. *IEEE Intelligent Systems*, **19**(1):24–25.

[7] De Roure, D., Goble, C., Stevens, R. (2009). The design and realisation of the myExperiment Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, **25**(5):561–567.

[8] Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J.L., Middleton, D. (2009). Ontology-supported scientific data frameworks: The Virtual Solar-Terrestrial Observatory experience. *Computers and Geosciences*, **35**(4): 724–738.

[9] Gahegan, M., Luo, J., Weaver S.D., Pike, W., Banchuen, T. (2009). Connecting GEON: Making sense of the myriad resources, researchers and concepts that comprise a geoscience cyberinfrastructure. *Computers and Geosciences*, **35**(4): 836–854.

[10] Gil, Y. (2009). From Data to Knowledge to Discoveries: Scientific Workflows and Artificial Intelligence. *Scientific Programming*, **17**(3).

[11] Golbeck, J., Hendler, J. (2008). A semantic web approach to the provenance challenge. *Concurrency Computation Practice and Experience*, **20**(5):431–439.

[12] Hede, K. (2010). In silico research: pushing it into the mainstream. *Journal of the National Cancer Institute*, **102**(4): 217–219.

[13] Langley, P. (2000). The computational support of scientific discovery. *International Journal of Human-Computer Studies*, **53**(3):393–410.

[14] Li, G., Uren, V., Motta, E., Shum, S.B., Domingue, J. (2002). ClaiMaker: Weaving a Semantic Web of Research Papers. Lecture Notes In Computer Science, Vol. 2342, *Proceedings of the First International Semantic Web Conference on The Semantic Web*, pp. 436–441.

[15] Novacek, V., Groza, T., Handschuh, S. Decker, S. (2010). CORAAL – Dive into Publications, Bathe in the Knowledge. *Journal of Web Semantics*, **8**(2–3):176–181.

[16] Noy, N.F. (2004) Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, **33**(4):65–70.

[17] Olson, G.M., Zimmerman, A., Bos, N. (Eds.) (2008). Scientific Collaboration on the Internet. MIT Press, Cambridge, MA, pp. 432.

[18] Poole, D., Smyth, C., Sharma, R. (2009). Ontology Design for Scientific Theories that Make Probabilistic Predictions. *IEEE Intelligent Systems*, **24**(1):27–36.

[19] Sahoo, S.S., Sheth, A., Henson, C. (2008) Semantic provenance for eScience: Managing the deluge of scientific data. *IEEE Internet Computing*, **12**(4):46–54.

[20] Simmhan, Y.L., Plale, B., Gannon, D. A Survey of Data Provenance in e-Science. *SIGMOD Record*, **34**(3):31–36.

[21] Wache, H., Vogele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., Hubner, S. (2001). Ontology-based integration of information – a survey of existing approaches. In: *Proceedings of the IJCAI'01: 17th International Joint Conferences on Artificial Intelligence*. Seattle, WA, pp. 108–117.

Semantic Web 1 (2010) 17–24
DOI 10.3233/SW-2010-0022
IOS Press

Model-Assisted Software Development: Using a 'semantic bus' to automate steps in the software development process

Editors: Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited Reviews: Axel Polleres, DERI, National University of Ireland, Ireland; Krzysztof Janowicz, Pennsylvania State University, USA

Kunal Verma* and Alex Kass
Accenture Technology Labs, 50 W. San Fernando Street, Suite 1200, San Jose, CA, USA

Abstract. The early phases of the software-development lifecycle (SDLC) for enterprise-scale systems – in particular, requirements elicitation, functional design, and technical design – are difficult to automate because they involve the application of several different kinds of domain knowledge. In this paper, we will provide a vision of how creating semantic models of domain knowledge used in each phase, and defining semantic representation, through which tools in the various phases can communicate knowledge across phases, can help provide more automation both *within* and *across* these phases. We refer to the collection of semantic models needed to support this automation as *the semantic bus for software development*. We refer to the semi-automated process that we envision making use of this bus to support the SDLC as, *Model-Assisted Software Development (MASD),* which is a variation on the Model-Driven Development idea. We will describe tooling we have built which realizes part of this vision, and will outline a roadmap of potential research opportunities in this space.

Keywords: Software engineering, semantic bus, model-assisted software development

1. Introduction

The effective execution of each phase of a complex process often involves applying a type of knowledge that is specific to that phase of the process, as well as accessing and building on the knowledge created in other phases. The effective automation of complex processes involving multiple teams of people generally requires a loosely coupled set of tools supporting each phase of the process. Keys to success include supporting the application of phase-specific knowledge by each tool and communication of knowledge *across* phases and tools.

The early phases of enterprise-scale software development – from specification through functional and technical design – represent an important and complicated example of this challenge. These activities typically involve experts in multiple business domains, as well as expert functional and technical architects. Each group of experts contributes their respective types of knowledge to the process: For instance, in the elicitation sessions within the requirements phase, expert stakeholders' knowledge of business objectives and processes is leveraged to identify key requirements that the solution must fulfill. During functional design, architects leverage their knowledge of functional and software frameworks to create functional design artifacts. In technical design, another set of architects leverages their knowledge of architectural patterns to choose relevant technical services and create technical design artifacts.

The complex, knowledge driven nature of the work in the early phases of the software-development life-

*Corresponding author. E-mail: k.verma@accenture.com.

cycle makes semantic technologies very relevant to their automation – both to represent the knowledge that gets applied in each phase, and to support knowledge transfer between phases. However, there are no widely-deployed commercial tools that support the use of semantic models for the specification and design of enterprise systems. We believe that this is one reason that these up-front phases remain the least automated portions of the software-development lifecycle.

Filling this gap is important because it could help address the well-documented problems associated with these early software-engineering phases: Defects in requirements and design are common and expensive, often leading to extensive rework or delays in software deployment [17]. Defects can arise from failure to leverage applicable knowledge *within a phase*, or from a failure to reason *across phases* to keep the deliverables of various phases in synch with each other; both of these issues could be mitigated by tools that support automated application of appropriate knowledge models to the development and review of each activity's deliverables.

In this vision statement, we will describe how semantic web technologies, combined with intelligent tools that leverage knowledge encapsulated in the semantic models, can be used to support the early phases of the SDLC. We refer to the collection of semantic models needed to support this automation as *the semantic bus for software development*. We refer to the semi-automated process that we envision making use of this bus to support the SDLC as *Model-Assisted Software Development (MASD).* As in the well-known vision often referred to as *Model-Driven Development* (MDD), the MASD process we envision relies on models that can be transformed to support progression between phases. However, as we shall describe, the role of the models, and the nature of the automation we envision is rather different than in traditional MDD.

Of course, our vision has not yet been fully realized in software. However, to make our discussion as concrete as possible, we will briefly describe some tooling which has been built to support aspects of the requirements and design activities, which begins to realize parts of the vision. Our hope is that over time, not only will we be able to implement more of the pieces, but that a rich eco-system of tools from other developers will emerge to support each phase of the SDLC, all of which can be loosely coupled through the envisioned semantic bus, allowing development teams to mix and match freely to suit their specific needs.

2. Sketching the semantic bus

We see a need for at least two kinds of semantic models, or ontologies, making up the semantic bus.

1. Artifact inter-communication models: A set of linked models for representing the artifacts produced by the various phases of software engineering. These models are intended to be independent of any business domain or underlying technology domain. At the schema level, these models contain descriptions of the component structure of a phase's artifacts, and the relationships between components in one phase and corresponding components in other phases. At the instance level, they contain domain-specific data from the actual artifacts. This part of the bus will be used to transfer data across phases – for instance from requirements to functional design. Some work in this space has already begun in the context of the Open Services for LifeCycle Collaboration (OSLC) [7], where an IBM-led community is creating RDF based representations for artifacts in different stages of software lifecycle such as requirements and testing. The goal of this initiative is not automated generation of artifacts from one stage to the other, rather it is maintaining traceability links across tools. However, having a standardized ontology (or set of linked ontologies) for representing artifacts in different stages of the software lifecycle will be an important enabler of the semantic bus.

2. Domain-specific phase-enablement models: These ontologies model facts about specific business and technology domains relevant to the systems being developed. They will be used by a set of tools for tasks requiring domain specific reasoning, such as gap analysis (during requirements phase) or selection of relevant technical services (during design). One example would be an ontology that describes the typical decision points in the Apache Axis Framework [1]. Another would be an ontology that provides a model of commonly used requirements and capabilities in the banking domain.

3. Using the semantic bus in model assisted software development

In this paper we are focused on three specific activities that come early in the software-development lifecycle: 1) requirements development, 2) functional design, and 3) technical design.

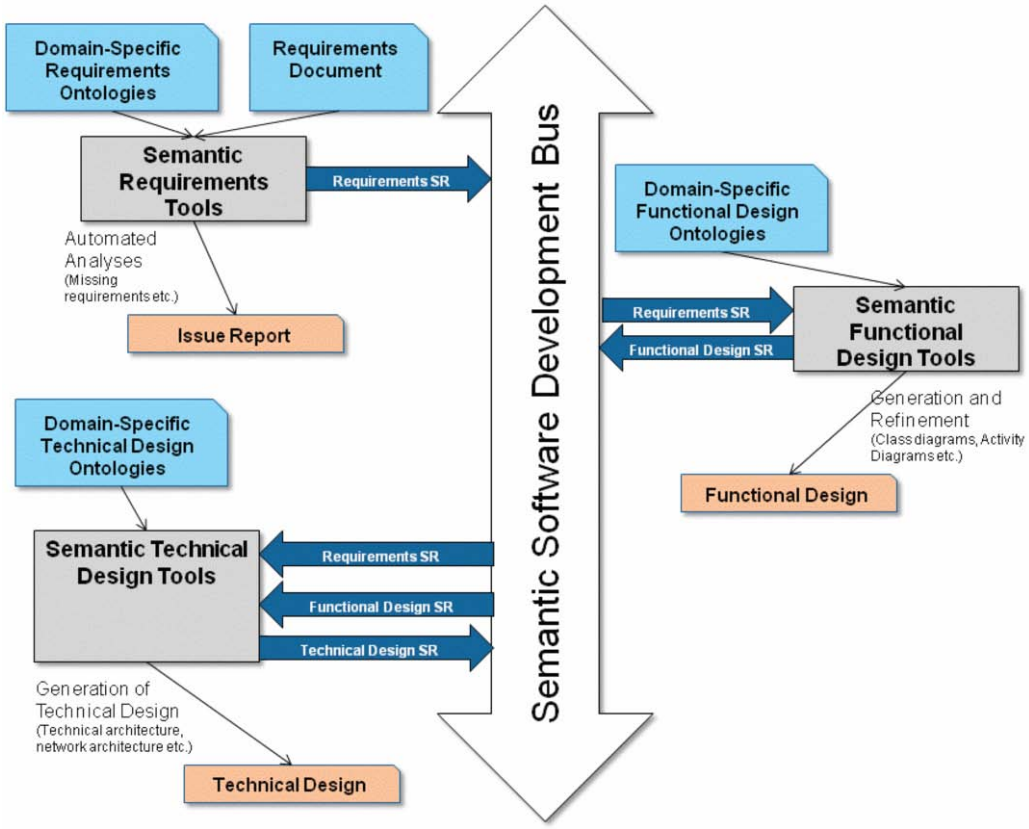

Fig. 1. Semantic bus for software development.

Figure 1 illustrates the flow of information from the requirements document to functional and technical design. As the figure suggests, we expect all the tools to leverage semantic models in two ways – 1) use the domain specific ontologies to help users perform the specific task and 2) leverage the semantic bus for software development to get data input from previous stages or make their data output available to other tools.

One may ask how our vision is different from the much-discussed vision of Model Driven Development (MDD) [3] or its popular instantiation by OMG-Model Driven Architecture [8]. MDD, where downstream artifacts are automatically generated from models, has been a long-term aspiration of the software engineering community. We obviously agree that models will play an important role in the future of software development. The question really is *what* role? How will models be used in a more automated software development process, and what will that process look like? On some of these questions our vision is a bit different from the most common existing approaches to MDD.

For example, a common activity pattern in the MDD vision is as follows:

- **Step 1:** A human, skilled in one space (such as requirements) uses MDD tool to create a formal model of the deliverable in that space.
- **Step 2:** A system automatically performs a transform on parts of that model to generate artifacts in the downstream space (such as design).
- **Step 3:** Another human, expert in the downstream space, then modifies the automatically-generated artifact.

One characteristic of the MDD vision is that the artifacts created at each stage have to be more formal in order to enable automated transformation. An advantage, at least in theory, is that the transformation from the upstream space to the downstream space (in Step 2) is fully automated. However, as others have noted (for e.g., [2] and [3]), there are some significant

challenges which make this conception of MDD very difficult to apply in practice.

1. The human-compatibility problem: Experience has shown that practitioners in an area such as requirements often find thinking in terms of a formal modeling notation very unnatural. Business Analysts, for example, are used to writing sentences, not creating models. Furthermore, the stakeholders who must sign off on these requirements are used to reading sentences, not model diagrams. Furthermore, while more technical practitioners may not find working with small models challenging, even for them it is very challenging to work with very large models at the scale that will be required for enterprise systems.

2. The knowledge representation problem: the amount of knowledge needed to perform Step 2 well on real-world artifacts is much greater than what any existing system has. As a result, the amount of modification required in Step 3 is often fairly high.

3. The update problem: MDD looks good for initial generation of artifacts. However, unless Step 3 is eliminated, the challenge of how to handle updates to upstream artifacts (such as requirements) arises: An architect who spends time understanding the output of Step 2, and then modifies it, will need to repeat that modification step every time the output of Step 1 changes even slightly, since the changes made in Step 3 will be erased when Step 2 is run again. The effort of having to repeat Step 3 for small changes in Step 1 can wipe out the gains from automating Step 2.

To address these concerns, we envision a variation on the MDD theme, which we call Model-Assisted Software Development. One key distinguishing feature of MASD is that the artifacts that the human participants are asked to create and understand are the more traditional human-readable descriptions. In our proposed approach, the formal models designed for automated consumption are automatically created from the human-created documents. For instance, a business analyst is asked to create a well-structured, *natural language* requirements specification. The structured model of the requirements is *automatically* generated by a system capable of analyzing the requirements text. The model then lives alongside the human-generated document, and is automatically kept in synch. The text document is used by human analysts and stakeholders, while the model is available to support automated reasoning, and generation of downstream transformations. So MASD sidesteps the human-compatibility problem by relegating the model to a behind-the-scenes role: in MASD, models are treated as an internal representation for systems to manipulate as much as possible, and for practitioners as little as possible. A second distinguishing characteristic is that we envision more of a *semi*-automated transformation process resulting in higher-quality downstream artifacts. By employing more sophisticated knowledge models in the systems performing activities like Step 2 above, *and* involving humans (with their much larger knowledge models) we seek to minimize Step 3. Human-supplied knowledge helps us side-step the knowledge-representation problem, and minimizing Step 3 reduces that pain of the update problem. In other words, one way to think about MASD vs. MDD is this: The MASD vision retreats from the theoretical ideal of complete automation embraced by MDD in favor of a more modest level of automation which can actually be achieved even with the complex artifacts required for real-world enterprise software development.

4. Semantics in requirements engineering

Requirements Engineering is the first phase of most software projects. This phase involves business analysts eliciting requirements from stakeholders and documenting them. The business analysts use their domain-specific knowledge to ask relevant questions and guide discussions. Once the requirements are documented, the practitioners use their knowledge to detect issues in requirements documents, such as conflicting or missing requirements. Practitioners also often need to manually perform impact analyses to determine the cost of changes requested by the stakeholders during the course of the projects.

Much of the work in this field (for e.g., [15]) makes the assumption that users will create formal models, instead of natural language requirements specifications, which are still the norm. We believe that the focus should be on generating formal models from natural language text, since it is likely to be the mode of writing requirements specifications for years to come. We have developed a tool called the Requirements Analysis Tool (RAT) [16] that converts a textual requirements document into a semantic RDF graph, which can be queried and reasoned upon. Currently, RAT helps users detect missing non-functional requirements with the help of non-functional requirements ontology. In addition, it automatically generates interaction diagram with the help of some rules. However a number of issues such

as conflict detection and impact analysis still remain open issues.

There has been some work on trying to detect missing requirements with the help of domain models. Kaeya and Saiki [1] proposed manually mapping requirements to elements in a domain-specific ontology and they have a measure of completeness based on the number of ontological elements that do not have any requirements mapped to them. We have recently developed a tool called the Process Model Requirements Gap Analyzer (ProcGap) [17] that uses natural language processing technologies automatically maps requirements to process models, such as "*Order to Cash,*" and helps users see the gaps and similarities. ProcGap, for instance, will flag any elements of the standard process model that do not seem to be covered by any project requirement, since these may represent missing requirements.

The early work described above, on leveraging semantic models in requirements analysis, gives some hint of semantic models can achieve in the requirements space, but we believe that there are a number of very important unsolved problems in this area. There is much work to be done in developing domain-specific ontologies. Currently, most software providers provide specifications of their software in textual form. The ability to extract formal specifications from textual specifications will be extremely valuable. In addition, research is needed on what type of reasoning is suitable for various kinds of analyses needed in requirements engineering, such as conflict detection and impact analysis. Finally, we have done some initial work on creating a requirements ontology that can be used for creating downstream artifacts, but much more work needs to be done creating a comprehensive requirements ontology that could serve as the basis for the semantic bus for software engineering.

5. Semantics in functional design

The next stage after requirements analysis is functional design. This is another knowledge-intensive stage, in which the functional architect must create a functional design based on their knowledge of software frameworks/tools and their understanding of the requirements. Usually, the functional design is represented as a number of artifacts, ranging from informal figures to more formal UML class or activity diagrams.

We believe that there are two main issues in this phase. First, while there is some tool support in this phase (for e.g., Rational Software Architect helps users generate UML diagrams), there is not much support for the knowledge-intensive decisions that need to be made by the functional architects. For example, if a functional architect realizes that a solution must be service oriented, she must make decisions on which framework to use (for e.g. Apache Axis) and once the framework is decided she must make decisions on the granularity of services and which parts of the framework to be used. Domain-specific ontologies, along with reasoning engines, should be able to assist functional architects with these kinds of decision-making. Second, some knowledge is often lost in the transition between the requirements and functional design phases. This can have various causes, ranging from lack of time, to misinterpretation of some requirements by the functional architect.

A number of researchers have proposed using natural language processing (NLP) to automatically generate design artifacts out of requirement documents and use case specifications. Examples include UCDA [6], LIDA [9], work by Ilieva et al. [4] and Gelhausen and Tichy [12]. We also explored an early prototype called Functional Design Creation Tool (FDCT) for generating a first cut at some functional-design artifacts from requirements based on heuristics [11].

However, none of these approaches (including ours) leverage domain-specific ontologies to generate the design. While these approaches are able to provide a model based on the requirements, they are typically not sufficient to model non-trivial systems, since they do not capture relationships of generated model of existing software libraries or systems. For example, here are some questions that these approaches do not answer:

1. Which modules of SAP do the generated classes interact with?
2. Which classes of frameworks such as Spring or Apache Axis should be should be used to implement some of the generated classes?

We believe that there is a clear opportunity to help functional architects generate functional designs with the help of domain specific ontologies and associated reasoning engines. As with requirements, there is some previous work in this space, but there are a number of open issues and unanswered questions.

6. Semantics in technical design

Technical design involves a number of activities, such as deciding the type of architecture (for example, three tier architecture vs. cloud-based architecture) and the types of infrastructural services that are needed such as encryption and logging. Also activities such as choosing appropriate hardware based on the architectural decisions, infrastructural services needed and performance criteria specified using non-functional requirements. It is performed on the basis of inputs from requirements and functional design. The technical architects also use their domain-specific knowledge to come up with such architectures.

Though this is a very important area of the project, there is practically no work on building knowledge-based tools to support technical design. In a preliminary work [10], we explored how description logics and subsumption-based reasoning can help users select relevant services and hardware based on vendor recommendation. This is an extremely rich area in which to explore the use of semantic technologies.

7. A deeper look at the semantic bus with the help of an end-to-end scenario

In this section, we will discuss an end-to-end example to illustrate how the semantic bus can be used to transfer information from one phase of software engineering to another. Consider the following steps from a use case:

UC-4-1: Project Manager navigates to employee page in PRMS.
UC-4-2: Project Manager searches for employee record by specifying employee id / employee full name.
UC-4-3: Project Manager modifies the employee record.
UC-4-4: PRMS sends updated employee record to the Employee Repository.
UC-4-5: PRMS sends notification of change to the Resource Manager.

The business analyst captures the use case in restricted natural language supported by the requirements analysis tool. In addition, the business analyst creates a glossary that defines the different terms used in the requirements and use cases. The glossary (shown in Fig. 2) captures different types of information about the entities, such as the type of the entity (for e.g., person, system, data attribute, etc.) and whether an entity is part of another entity (for e.g., Employee ID is an attribute for Employee record).

Term	Explanation	Type	Part of
Active projects	Projects that are active	PassiveEntity	
Administrator	Admin of project resource management system.	Person	
Assign Resource Module	Module of PRMS that provides capability for resource assignment.	System	Project Resource Management System
Backup Master Employee Repository	Backup for Master Employee Repository.	System	
budget	Budget for a project	DataAttribute	
Client	Client for the selected project.	Person	
contractor	External contractors who are available for staffing.	Person	
Credit Check System	Vendor system to check credit of employees.	System	
Credit verification report	Report from credit verifier	PassiveEntity	
Deployment Resource Manager	Deployment resource manager of project.	Person	
Employee	Employee of Accenture.	Person	
employee details	Employment and personal details of employee.	PassiveEntity	
employee full name	Full name of employee	DataAttribute	employee record
employee id	Unique ID for the employee	DataAttribute	employee record
employee page	Employee web page	System	Project Resource

Fig. 2. Entity glossary in the requirements analysis tool.

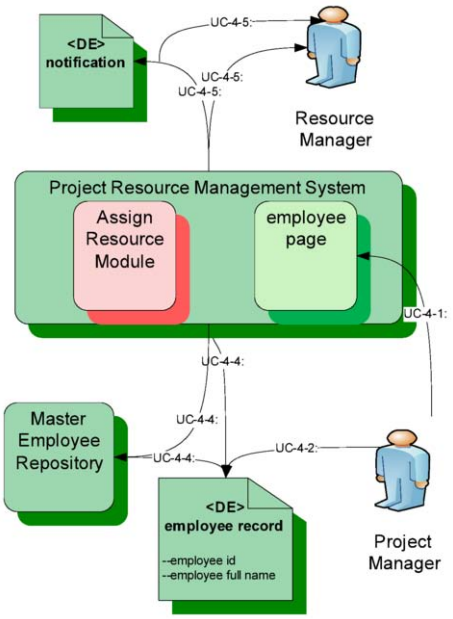

Fig. 3. Automatically generated visualization of use case text using requirements analysis tool.

The requirements analysis tool then uses a combination of lexical and semantic techniques to analyze the use case text, extract structured content for each use case step and populate the core requirements ontology defined in [16] to create a semantic RDF graph. Thus, the information is transformed from being simple text to semantic graph that can reasoned upon and transformed to create different kinds of reports and models. We show an interaction diagram generated by the requirements analysis tool in Fig. 3. This diagram is generated by applying heuristics to figure out which requirements/use case steps represent interactions between the systems and users.

A number of other tools such as UCDA [6], LIDA [9] and FDCT [11] can also be used to generate downstream artifacts such as class diagrams. In addition tools such as ProcGap [17] can be used to compare how the requirements and use cases compare to reference processes and capabilities in that domain. However, none of these tools provide any support using domain-knowledge to create artifacts. While there are tools such as Skyway Builder [12], which can generate Java code based on a model of the Spring framework, there are no such tools for earlier phases of software engineering.

To illustrate our point about tools that leverage domain-specific ontologies, let us assume the above use case had to be implemented as a Web service using the Apache Axis Framework at the front end and using a batch-style architecture in the back-end. A tool that would guide users in including relevant classes using a semantic model of the Apache Axis Framework would be very helpful. Similarly, a tool that helps users select relevant infrastructure services for the batch-style architecture would also be helpful. Currently, the aspects of functional and technical design, that involve domain knowledge, are done manually with no automated support.

8. Conclusions

In this paper, we have presented a vision for a suite of tools to enable the early phases of the software engineering process. Even though these phases are extremely important, and highly knowledge intensive, there is very little knowledge-based tool support for practitioners. One reason for is that the software-engineering community has tended to focus on tool support for the most tangible phases, such as coding, rather than the early phases which involve messy, less structured artifacts, and require significant amounts of domain knowledge.

This is not the first paper to talk about automating aspects of software development. Interested readers can read a nice summary article [14] that discussed a number of previous vision papers. We believe that there are a number factors that have recently emerged to enable our vision of the semantic bus – 1) the field of natural language processing has evolved significantly enough that converting natural language text to formal models is getting more feasible; 2) OSLC) [7], which uses RDF to represent software-engineering artifacts is gathering momentum and is currently supported by a number of commercial tools; and 3) the Requirements Analysis Tool, which has been deployed at over 400 projects within our organization, has been used by many of those projects to generate high-level design from natural-language requirements.

These are encouraging signs, but, much work still needs to be done, especially around tooling that leverages domain and phase-specific ontologies. We believe that this represents a rich area in which Semantic Web researchers can leverage their skills to build tools that will have a large impact on the state of the art in software engineering.

Acknowledgements

We would also like to acknowledge our collaborators Rey Vasquez, Santonu Sarkar, Vibhu Sharma and Edy Liongosari, who helped us in shaping this vision paper.

References

[1] Apache Axis Framework, http://ws.apache.org/axis/.
[2] A.E. Bell, Death by UML Fever, *ACM Queue Magazine* **2**, No. 1, March 2004.
[3] B. Hailpern and P. Tarr, Model-driven development, The good, the bad and the ugly, *IBM Systems Journal*, **45**(3), 2006.
[4] M. Ilieva and O. Ormandjieva, Automatic transition of natural language software requirements specification into formal presentation, in *Lecture Notes in Computer Science*. Springer-Verlag, 2005, pp. 392–397.
[5] H. Kaiya and M. Saeki., Using domain ontology as domain knowledge for requirements elicitation, in *Proceedings of the IEEE International Requirements Engineering Conference (RE)*, pp. 186–195, 2006.
[6] D. Liu, K. Subramaniam, A. Eberlein, and B. H. Far, Natural language requirements analysis and class model generationusing UCDA, in *Lecture Notes in Computer Science*. Springer-Verlag, pp. 295–304, 2004.
[7] Open Services for Lifecycle Collaboration (OSLC), http://open-services.net/html/Home.html.
[8] Object Management Group, Model Driven Architecture, http://www.omg.org/mda/.
[9] S.P. Overmyer, B. Lavoie, and O. Rambow, Conceptual modeling through linguistic analysis using LIDA, in *Proceedings of the 23rd International Conference on Software Engineering*, 2001, pp. 401–410.
[10] S. Sarkar and K. Verma, Accelerating technical design of business applications: a knowledge-based approach, in *Proceedings of the 3rd India Software Engineering Conference. (ISEC)* 2010.
[11] V.S. Sharma, S. Sarkar, K. Verma, A. Panayappan, and A. Kass, Extracting High-Level Functional Design from Software Requirements, *17th Asia-Pacific Software Engineering Conference*, 2010.
[12] Skyway Software, http://www.skywaysoftware.com/.
[13] Tom Gelhausen, Walter F. Tichy: Thematic Role Based Generation of UML Models from Real World Requirements. *IEEE International Conference on Semantic Computing (ICSC)*, pp. 282–289, 2007.

[14] Michael Uschold, Ontology-Driven Information Systems: Past, Present and Future. *Proceeding on the 5th Conference on Formal Ontology in Information Systems (FOIS)*, pp 3–18, 2008.

[15] A. van Lamsweerde, E. Letier, and R. Darimont, Managing Conflicts in Goal-Driven Requirements Engineering. *IEEE Transactions on Software Engineering*, **24**(11), 1998.

[16] K. Verma, and A. Kass, Requirements Analysis Tool: A Tool for Automatically Analyzing Software Requirements Documents, *7th International Semantic Web Conference*, 2008.

[17] K. Verma, A. Kass, and R. Vasquez, Aligning Requirements Documents to Industry-Specific Process Models, Technical report, 2010.

[18] K. Wiegers 2003, *Software Requirements*, Microsoft Press.

Semantic Web 1 (2010) 89–96
DOI 10.3233/SW-2010-0023
IOS Press

Semantic search on the Web

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Krzysztof Janowicz, Pennsylvania State University, USA; Axel Polleres, DERI Galway, Ireland

Bettina Fazzinga [a,*] and Thomas Lukasiewicz [b]
[a] *Dipartimento di Elettronica, Informatica e Sistemistica, Università della Calabria, Italy*
[b] *Computing Laboratory, University of Oxford, UK*
E-mail: thomas.lukasiewicz@comlab.ox.ac.uk

Abstract. Web search is a key technology of the Web, since it is the primary way to access content on the Web. Current standard Web search is essentially based on a combination of textual keyword search with an importance ranking of the documents depending on the link structure of the Web. For this reason, it has many limitations, and there are a plethora of research activities towards more intelligent forms of search on the Web, called *semantic search on the Web*, or also *Semantic Web search*. In this paper, we give a brief overview of existing such approaches, including own ones, and sketch some possible future directions of research.

Keywords: Semantic search on the Web, Semantic Web search, Web search, Semantic Web, ontologies

1. Introduction

Web search is a key technology of the Web, which is essentially based on a combination of textual keyword search with an importance ranking of the documents depending on the link structure of the Web. For this reason, it has many limitations, and there are a plethora of research activities towards more intelligent Web search, called *semantic search on the Web*, or also *Semantic Web search*, which is currently one of the hottest research topics in both the Semantic Web and Web search (see [18] and [1], respectively).

There is no unique definition of the notion of semantic search on the Web. However, the most common use is the one as an improved form of search on the Web, where meaning and structure are extracted from both the user's Web search queries and different forms of Web content, and exploited during the Web search process. Such semantic search is often achieved by using Semantic Web technology for interpreting Web search queries and resources relative to one or more underlying ontologies, describing some background domain knowledge, in particular, by connecting the Web resources to semantic annotations, or by extracting semantic knowledge from Web resources. Such a search usually also aims at allowing for more complex Web search queries whose evaluation involves reasoning over the Web. Another common use of the notion of semantic search on the Web is the one as search in the large datasets of the Semantic Web as a future substitute of the current Web. This second use is closely related to the first one, since the above semantic annotation of Web resources, or alternatively the extraction of semantic knowledge from Web resources, actually corresponds to producing a knowledge base, which may be encoded using Semantic Web technology. That is, the latter semantic search on the Web can essentially be considered as a subproblem of the former one.

Another closely related use is the one as natural language search on the Web, where search queries are formulated in (written or even spoken) natural language. Many approaches try to translate such queries into formal queries in a structured query language, which are generally available in the above semantic search in the context of the Semantic Web. The answers to such natural language queries may be Web resources as usual, or they may also be structured or natural language results, towards more informative results, e.g., by show-

*Corresponding author. E-mail: bfazzinga@deis.unical.it.

ing structured information extracted from the resulting Web pages, and by additionally connecting the search result with Wikipedia articles. This is another meaning of semantic search, which is actually already a very simple form of question answering.

Frequently, the notion of semantic search also covers some other (often less) semantic ideas and concepts. For example, faceted search allows for exploring results according to a collection of predefined categories, called facets. Closely related is clustered search, where such facets are not predefined. A further example is the suggestion of related searches, such as the completion and correction of Web search queries, which are well-known from standard Web search engines. Another example is full-text similarity search, where blocks of text ranging from phrases to full documents, rather than few keywords, are submitted. Closely related is ontological similarity search (e.g., [19]), based on the similarity of ontological entities.

In this paper, we discuss especially the two initial interpretations of the notion of semantic search on the Web, which both refer to the context of the Semantic Web, as well as their generalizations towards natural language search on the Web. The rest of this paper is organized as follows. In Section 2, we describe some representative approaches to semantic search on the Web. Section 3 sketches our own such approach. In Section 4, we conclude and describe our vision for the future of semantic search on the Web.

2. Overview of existing approaches

State-of-the-art approaches to semantic search on the Web can be classified as follows:

1. approaches based on structured query languages, such as [6,12,17,20,25,26,28];
2. approaches for *naive* users, where no familiarity with ad-hoc query languages is required. In turn, these approaches can be divided into:
   - keyword-based approaches, such as [2,4,14, 15,21,29,30,32,34], where queries consist of lists of keywords;
   - natural-language-based approaches, such as [5,8,11,13,22,23], where users can express queries by means of the natural language.

In the following, we give an overview of the main approaches belonging to the above categories.

2.1. Approaches based on structured languages

SHOE [17] is one of the first attempt to semantically query the Web. SHOE provides the following: a tool for annotating Web pages, allowing users to add SHOE markup to a page by selecting ontologies, classes, and properties from a list; a Web crawler, which searches for Web pages with SHOE markup and stores the information in a knowledge base (KB); an inference engine, which provides new markups by means of inference rules (basically, Horn clauses); several query tools, which allow users to pose structured queries against an ontology. One of the query tools allows users to draw a graph in which nodes represent constant or variable instances and arcs represent relations. To answer the query, the system retrieves subgraphs matching on the user graph. The SHOE search tool allows user to pose queries by first choosing an ontology from a drop-down list and next choosing classes and properties from another list. Finally, the system builds a conjunctive query, issues the query to the KB, and presents the results in a tabular form.

Subsequent approaches are [6,12], which mainly focus on RDF. Swoogle [12] is a crawler-based system for discovering, indexing, and querying RDF documents. Swoogle mainly provides a search for Semantic Web documents and terms (i.e., the URIs of classes and properties). It allows users to specify queries containing conditions on the document-level metadata (i.e., queries asking for documents having `.rdf` as the file extension), and it also allows users to search for Semantic Web documents using RDF/XML as the syntax language. Retrieved documents are ranked according to a ranking algorithm measuring the documents' importance on the Semantic Web.

The Corese system presented in [6] is an ontology-based search engine for the Semantic Web, which retrieves Web resources annotated in RDF(S) by using a query language based on RDF(S). Corese is able to *approximately* search the Semantic Web. Approximation is provided by employing inference rules and by computing the semantic distance of classes or properties in the ontology hierarchies. Specifically, Corese retrieves Web resources whose annotations are specializations of the query, and it also retrieves those resources whose annotations refer to concepts and relations that are hierarchically *close enough* to those of the query. Another approach dealing with approximation is presented in [26]. The aim of this approach is approximately querying RDF datasets with SPARQL [33]. To this end, a SPARQL query is encoded as a set of triple

constraints with variables, and an approximate answer is a substitution of the variables with data that may not satisfy all the constraints. The proposed strategy refines the accuracy of the answers progressively, so that the algorithm searching for the answers can be stopped at any time producing a meaningful result.

More recent approaches based on structured languages are [20,25,28]. In particular, [28] introduces ONTOSEARCH2, which is a search and query engine for ontologies on the Semantic Web. It stores a copy of the ontologies in a tractable description logic and allows SPARQL queries to be evaluated on both the structures and instances of ontologies. The Coraal system [25] is a knowledge-based search engine for biomedical literature. Coraal uses NLP-based heuristics to process texts and build RDF triples from them. These RDF triples are integrated with existing domain knowledge and all the collected information can be queried by the user by means of a specific query language. The NAGA semantic search engine [20] provides a graph-based query language to query the underlying KB represented as a graph. The KB is built automatically by a tool for knowledge extraction from Web sources, which extends the approach proposed in [27]. The nodes and edges in the knowledge graph represent entities and relationships between entities, respectively. The NAGA query language extends SPARQL, allowing complex graph queries with regular expressions over relationships on edge labels. Answers to a query are subgraphs of the knowledge graph matching the query graph and are ranked using a specific scoring model for weighted labeled graphs.

2.2. Keyword-based approaches

Two preliminary approaches to the problem of Semantic Web search are proposed in [2,14]. In particular, [2] focuses on issues dealing with ontology search, presenting the (ontology) search engine OntoSelect. This allows users to search for ontologies by specifying the ontology title or the topic of interest. In the latter case, users can specify an URL of a Web document containing information dealing with the topic. Then, a linguistically/statistically derived set of relevant keywords is extracted automatically and used in the search. Whereas [14] focuses on augmenting the results of traditional keyword search with data retrieved from the Semantic Web. Query processing can be summarized as follows: when a user query is issued, query terms (keywords) are mapped to Semantic Web nodes: in the case of multiple matching, some heuristics (for instance, taking into account the user profile, etc.) are employed to find the right one. Once nodes matching the search terms are found, the approach uses some heuristics to choose what part of the Semantic Web graph around these nodes, has to be returned as a result (i.e., the first N triples, where N is some threshold). Moreover, [14] proposes an approach to improve traditional keyword search by disambiguating the meaning of the terms in the query. To this end, an additional link next to each search result is added, so that, if the user clicks on this link, only Web documents having a content *semantically similar* to the document reachable from that link are shown.

More recent approaches for naive users based on keyword search are [4,21,29]. SemSearch [21] provides a Google-like query interface allowing users to specify queries without requiring any knowledge about ontologies or specific languages. User queries consist of two or more keywords, whose semantic meaning is taken into account to reformulate the queries themselves according to a formal query language syntax. Keywords are assigned a semantic meaning by matching them against a collection of classes, properties, and instances in semantic data repositories. Since each keyword can match a class, a property, or an instance, several combinations of semantic matchings of the keywords are considered. For instance, it can be the case that every keyword matches a class, or that the first keyword matches a class, while the second matches a property, and so on. All the combinations of matchings are taken into account in the reformulation process, and each combination leads to a distinguished formal query, obtained from a pre-determined set of query templates. After the reformulation, formal queries are exactly evaluated, and this yields results that are semantically related to all the user keywords.

In [29], a similar approach to [21], keyword queries are translated into conjunctive queries to be evaluated against an underlying KB. Here, the structure of the formal queries that are eventually evaluated does not conform to pre-determined templates. Formal queries are built exploiting a graph-based technique to find the connections between the entities in the user queries. Specifically, query translation consists of the following three steps. First, the keywords in the user query are mapped onto ontology elements. Then, relations among these ontology elements are examined, and subgraphs of the KB are extracted. Each subgraph represents a set of relations connecting all the considered elements, thus the set of these subgraphs represents all the possible relationships among user keywords that

could not be explicitly specified by the user. Hence, these subgraphs correspond to the different queries that the user may be interested in. Finally, formal queries are generated by translating the subgraphs according to a proper language, and evaluated against the KB.

Falcons [4] is a keyword-based search engine for the Semantic Web, allowing concept and object search. Concept search is carried out by searching the classes and properties that match the query terms in the ontology selected by the user, and, furthermore, recommending other ontologies on the basis of a combination of the TF-IDF technique and the popularity of ontologies. Object search is performed in a similar way: besides returning the objects that match the query terms, the system also recommends other types of objects that the user is likely to be interested in.

SWSE [15] and Sig.Ma [30] are two recent tools allowing users to locate RDF entities via keyword search. Specifically, the result of a keyword search in SWSE is a list of entities matching the keyword along with a small description and a concept name, such as Person, Professor, etc. If the user clicks on "Person", then the results are filtered and only a list of "Person" entities is shown. The information about the entity is aggregated from multiple sources and is presented in a homogeneous view. The core of SWSE is YARS2 [16], a distributed architecture for indexing and querying RDF datasets. YARS2 collects pieces of information and aggregates them either by exploiting the URI of the entities (in the case that a unique identifier is used in the different sources), or by exploiting other object consolidation techniques. Furthermore, SWSE provides a SPARQL endpoint that allows expert users to pose complex queries. Similarly to SWSE, Sig.Ma [30] integrates results from several sources providing the user with an aggregate view of information, along with the sources. The disambiguation phase is similar to that of SWSE, but in this case user clicks are used to eliminate irrelevant sources. Sig.Ma also allows users to specify a list of properties besides the entities. User keywords are translated into a set of interrogations: some of which are submitted to Yahoo Boss [31] to retrieve Web pages, while the others are submitted to Sindice [9], a Semantic Web data index, to collect RDF entities and properties. Finally, all the retrieved information is integrated by exploiting some heuristics, based on the use of URIs and of label consolidation techniques.

A very recent approach aiming at helping the user to build *semantic queries* from keyword queries is the QUICK system [34]. A semantic query is a query to be evaluated on a domain-specific ontology. QUICK, whose approach is similar to that of [21], starts with a keyword query formulated by the user, and translates it into several semantic queries, each obtained by assigning an ontology concept (property, entity, etc. from a selected ontology) to each keyword. Then, the user is called for choosing the most appropriate semantic query among those generated by the system. If no semantic query among all those generated by QUICK is considered as appropriate by the user, then the user herself can guide the system towards the generation of an appropriate one by providing further specifications (e.g., indicating if a given keyword has to be intended as a property or an entity, etc.).

Among the keyword-based search engines for the Semantic Web, it is important to include YahooSearchMonkey [32], which is a framework aiming at improving the quality of the results of Yahoo! search. It allows publishers to specify how and what information about the Web page that they are willing to publish has to be displayed on the page of the results of Yahoo! search. Publishers can give these specifications in the form of microformats, eRDF, or RDFa metadata, which will be automatically extracted during the crawling process and will provide the search engine with a lot of information about the most relevant content of the Web page. This way, users will be able to see all the searched information, grouped and well organized, directly on the Web page of the results of Yahoo! search, without clicking on the target Web page.

2.3. Natural-language-based approaches

Some of the most known approaches focusing on natural language queries are [5,11]. In [5], the ORAKEL system is presented, where, before being evaluated, queries are first translated into a logical form, and then reformulated according to a target language, i.e., the language of the underlying KB. The translation from the logical form to the target language is described declaratively by a Prolog program. The overall approach is independent from the specific target language, since changing the ontology language only requires a declarative description of the transformation as a Prolog program, but no further change to the underlying system. The system relies on a specific kind of user, called lexicon engineer, who specifies how natural language expressions can be mapped onto predicates in the KB, i.e., how verbs, adjectives, and relational nouns can be mapped onto corresponding relations specified in the domain ontology.

The system presented in [11] supports (i) Semantic Web search over ontologies and (ii) semantic search over non-Semantic-Web documents. As regards the first kind of search, answers to a natural language query are retrieved by exploiting a previous system, called PowerAqua [23], which works in the following way: first, the user query is translated from natural language into a structured format, called *linguistic triple*; second, the terms of the linguistic triple are mapped to semantically relevant ontology entities. Finally, the ontological entities that best represent the user query are selected and returned. PowerAqua extends the AquaLog system proposed in [22], which works in the presence of a single ontology only, to the case of multiple ontologies. The second kind of search in [11], namely, the semantic search over non-Semantic-Web documents, is accomplished by extending the system proposed in [3]. Specifically, this relies on a new approach for annotating documents, consisting of the following steps: (i) extracting the textual representation of semantic entities, (ii) searching this textual representation in Web documents, and (iii) generating an annotation linking the semantic entities to each of the documents containing their textual representation. Furthermore, [11] deals with the problem of knowledge incompleteness, by switching to the traditional keyword search when no ontology satisfies the query.

A very recent approach for building SPARQL queries from natural language queries is presented in [8]. The first step in the SPARQL query generation is the transformation of the natural language query into a set of ontology concepts (classes, instances, properties, and literals), which is based on the assignment of a proper ontology concept to each word. If the system is not able to assign a proper ontology concept to a word, then the user is called for selecting the correct one. The user selections are used for training the system in order to improve its performance. The second step is the construction of triples of ontology concepts, which are finally inserted into SELECT and WHERE clauses for generating a SPARQL query. The results of the evaluation of the obtained SPARQL query are shown to the user both in a tabular and in a graphical form.

The most recent approach belonging to the category of natural-language-based approaches is the newest version of Google [13]. Besides being a widely used keyword search engine, Google is now evolving to a natural-language-based search engine. In fact, it has been recently augmented with a new functionality, which provides more precise answers to queries: instead of returning Web page links as query results, Google now tries to build query answers, collecting information from several Web pages. As an example, the simple query "barack obama date of birth" gets the answer "4 August, 1961". Next to the answer, the link *Show sources* is shown, that leads to the Web pages from which the answer has been obtained.

3. The FGGL approach

We now describe our approach to semantic search on the Web presented in [10], which is based on a structured query language that allows to formulate complex ontology-based (conjunctive) search queries.

More specifically, an ontologically enriched Web along with complex ontology-based search on the Web is achieved on top of the existing Web and using existing Web search engines. Intuitively, rather than being interpreted in a keyword-based syntactic fashion, the pieces of data on existing Web pages are connected to (and via) some ontological KB (in a lightweight ontology language) and then interpreted relative to this KB. That is, the pieces of data on Web pages are connected to (and via) a much more precise semantic and contextual meaning. More concretely, Web content is associated with semantic annotations; or, from another perspective, the Web is actually mapped into an ontological KB, which then allows for semantic search on the Web relative to the underlying ontology. In [10], we assume that the semantic annotations and their underlying ontology are explicitly given; in recent work, we also explore the automatic mapping of Web content to an ontological KB using rule-based data extraction techniques. Intuitively, such a KB can be considered as an ontological index over the Web, against which ontological Web search queries can be answered. This allows for answering Web search queries in a much more precise way, taking into account the meaning of Web search queries and pages, and it also allows for more complex ontology-based Web search queries that involve reasoning over the Web, which are also much closer to complex natural language search queries than current Boolean keyword-based search queries.

Query processing in our approach to semantic search on the Web is divided into (i) an offline inference step for pre-compiling the given ontological knowledge using standard ontology reasoning techniques, thus transforming the semantic annotations into so-called completed semantic annotations, which are published as standard Web pages so that they can be searched via standard Web search engines, and

(ii) an online reduction of complex ontology-based Web search queries into (sequences of) standard Web search queries, of which the answers are obtained by standard Web search and then used to construct the answer of the original ontology-based Web search query. This way of processing semantic search queries on the Web is shown to be ontologically correct (and in many cases also complete). The ranking of the search results is based on a ranking on objects, called *ObjectRank*, which generalizes (and can be reduced to) the standard PageRank ranking on Web pages. That is, essential parts of ontological search on the Web are actually reduced to state-of-the-art search engines. As important advantages, this approach can immediately be applied to the whole existing Web, and it can be done with existing Web search technology (and so does not require completely new technologies). Such a line of research aims at adding ontology-based structure and semantics (and thus in a sense also intelligence) to current search engines for the existing Web by combining existing Web pages and queries with ontological knowledge.

The ontological knowledge and annotations that are underlying our semantic search on the Web can be classified according to their contents: (a) the ontological knowledge and annotations may either describe fully general knowledge (such as the knowledge encoded in Wikipedia) for general ontology-based search on the Web, or (b) they may describe some specific knowledge (such as biomedical knowledge) for vertical ontology-based search on the Web. The former results into a general ontology-based interface to the Web similar to Google, while the latter produces different vertical ontology-based interfaces. Here, the ontology-based interface to the Web itself may be based on the full power of a structured query language for more expert users (to whom the underlying ontology should be visible in order to support query formulation) or on predefined simple form-based interfaces (e.g., similar to the ones used in Google's advanced Web search) for less expert users.

In [7], a variant of the above approach is explored, which uses inductive reasoning techniques rather than deductive ones. This adds especially the ability to handle inconsistencies, noise, and incompleteness.

4. Conclusion and vision for the future

We have given a brief overview of approaches to *semantic search on the Web* (also called *Semantic Web search*), which is currently one of the hottest research topics in both the Semantic Web and the Web search community. In semantic search on the Web, the current strong research activities of the former to realize search on the Semantic Web are merged with the current strong research activities of the latter to add semantics to Web queries and content when performing Web search. It is through this integration that the reasoning capabilities envisioned in Semantic Web technologies are coming to Web search and the Web. As we have seen, the formulation of queries and their results in semantic search on the Web is ultimately directed by a third area, namely, the one of question answering systems, which is based on natural language processing.

Although many approaches and systems to semantic search on the Web already exist, the research in this area is still at the very beginning, and many open research problems still persist. Some of the most pressing research issues are maybe (i) how to automatically translate natural language queries into formal ontological queries, and (ii) how to automatically add semantic annotations to Web content, or alternatively how to automatically extract knowledge from Web content.

Another central research issue in semantic search on the Web is (iii) how to create and maintain the underlying ontologies. This may be done either (a) manually by experts, e.g., in a Wikipedia like manner, where different communities may define their own ontologies, or (b) automatically, e.g., by extraction from the Web, eventually coming along with existing pieces of ontological knowledge and annotations (e.g., from existing ontologies or ontology fragments, and/or from existing annotations of Web pages in microformats or RDFa), or (c) semi-automatically by a combination of (a) and (b). Clearly, the larger the degree of automation, the larger is also the potential size of ontologies that can be handled and the smaller are the costs and efforts for generating and maintaining them. So, for the very large scale of the Web, a very high degree of automation is desirable. A closely related important research challenge is (iv) the evolution and updating of and mapping between the ontologies that are underlying semantic search on the Web, where it is similarly desirable to have a very high degree of automation.

A further important issue is (v) how to consider implicit and explicit contextual information to adapt the search results to the needs of the users. For example, the needs and motivations of users may be defined in terms of ontology-based strict and/or soft (weighted) constraints and (conditional) preferences (e.g., similar to [24]), which may then implicitly be expanded into

the semantic search query and/or used in the computation of the ranking on objects and search results.

Performing Web search in the form of returning simple answers to simple questions in natural language is still science fiction, let alone performing Web search in the form of query answering relative to some concrete domain or even general query answering. However, with the current activities towards semantic search on the Web, we are moving one step closer to making such science fiction become true, which ultimately aims at a human-like interface to the knowledge, information, services, and other resources available on the Web.

Acknowledgments

Thomas Lukasiewicz's work was supported by a Yahoo! Research Fellowship and by the European Research Council (ERC) under the European Union's 7th Framework Programme (FP7/2007-2013)/ERC grant no. 246858 — DIADEM.

References

[1] R. A. Baeza-Yates and P. Raghavan. Next generation Web search. In S. Ceri and M. Brambilla, editors, *Search Computing*, *LNCS* 5950, pp. 11–23. Springer, 2010.

[2] P. Buitelaar, T. Eigner, and T. Declerck. OntoSelect: A dynamic ontology library with support for ontology selection. In *Proc. Demo Session at ISWC-2004*, 2004.

[3] P. Castells, M. Fernández, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *IEEE Trans. Knowl. Data Eng.*, **19**(2):261–272, 2007.

[4] G. Cheng, W. Ge, and Y. Qu. Falcons: Searching and browsing entities on the Semantic Web. In *Proc. WWW-2008*, pp. 1101–1102. ACM Press, 2008.

[5] P. Cimiano, P. Haase, J. Heizmann, M. Mantel, and R. Studer. Towards portable natural language interfaces to knowledge bases — The case of the ORAKEL system. *Data Knowl. Eng.*, **65**(2):325–354, 2008.

[6] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker. Querying the Semantic Web with Corese search engine. In *Proc. ECAI-2004*, pp. 705–709. IOS Press, 2004.

[7] C. d'Amato, F. Esposito, N. Fanizzi, B. Fazzinga, G. Gottlob, and T. Lukasiewicz. Inductive reasoning and Semantic Web search. In *Proc. SAC-2010*, pp. 1446–1447. ACM Press, 2010.

[8] D. Damljanovic, M. Agatonovic, and H. Cunningham. Natural language interface to ontologies: Combining syntactic analysis and ontology-based lookup through the user interaction. In *Proc. ESWC-2010*, Part I, *LNCS* 6088, pp. 106–120. Springer, 2010.

[9] R. Delbru, A. Polleres, G. Tummarello, and S. Decker. Context dependent reasoning for semantic documents in Sindice. In *Proc. SSWS-2008*, 2008.

[10] B. Fazzinga, G. Gianforme, G. Gottlob, and T. Lukasiewicz. Semantic Web search based on ontological conjunctive queries. In *Proc. FoIKS-2010*, *LNCS* 5956, pp. 153–172. Springer, 2010.

[11] M. Fernández, V. Lopez, M. Sabou, V. S. Uren, D. Vallet, E. Motta, and P. Castells. Semantic search meets the Web. In *Proc. ICSC-2008*, pp. 253–260. IEEE Computer Society, 2008.

[12] T. W. Finin, L. Ding, R. Pan, A. Joshi, P. Kolari, A. Java, and Y. Peng. Swoogle: Searching for knowledge on the Semantic Web. In *Proc. AAAI-2005*, pp. 1682–1683. AAAI Press / MIT Press, 2005.

[13] Google. http://www.google.com.

[14] R. V. Guha, R. McCool, and E. Miller. Semantic search. In *Proc. WWW-2003*, pp. 700–709. ACM Press, 2003.

[15] A. Harth, A. Hogan, R. Delbru, J. Umbrich, S. O'Riain, and S. Decker. SWSE: Answers before links! In *Proc. Semantic Web Challenge 2007*, *CEUR Workshop Proceedings* 295. CEUR-WS.org, 2007.

[16] A. Harth, J. Umbrich, A. Hogan, and S. Decker. YARS2: A federated repository for querying graph structured data from the Web. In *Proc. ISWC/ASWC-2007*, *LNCS* 4825, pp. 211–224. Springer, 2007.

[17] J. Heflin, J. A. Hendler, and S. Luke. SHOE: A blueprint for the Semantic Web. In D. Fensel, W. Wahlster, and H. Lieberman, editors, *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential*, pp. 29–63. MIT Press, 2003.

[18] J. Hendler. Web 3.0: The dawn of semantic search. *Computer*, **43**(1):77–80, 2010.

[19] K. Janowicz, M. Wilkes, and M. Lutz. Similarity-based information retrieval and its role within spatial data infrastructures. In *Proc. GIScience-2008*, *LNCS* 5266, pp. 151–167. Springer, 2008.

[20] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, and G. Weikum. NAGA: Searching and ranking knowledge. In *Proc. ICDE-2008*, pp. 953–962. IEEE Computer Society, 2008.

[21] Y. Lei, V. S. Uren, and E. Motta. SemSearch: A search engine for the Semantic Web. In *Proc. EKAW-2006*, *LNCS* 4248, pp. 238–245. Springer, 2006.

[22] V. Lopez, M. Pasin, and E. Motta. AquaLog: An ontology-portable question answering system for the Semantic Web. In *Proc. ESWC-2005*, *LNCS* 3532, pp. 546–562. Springer, 2005.

[23] V. Lopez, M. Sabou, and E. Motta. PowerMap: Mapping the real Semantic Web on the fly. In *Proc. ISWC-2006*, *LNCS* 4273, pp. 414–427. Springer, 2006.

[24] T. Lukasiewicz and J. Schellhase. Variable-strength conditional preferences for ranking objects in ontologies. *J. Web Sem.*, **5**(3):180–194, 2007.

[25] V. Novácek, T. Groza, and S. Handschuh. CORAAL — Towards deep exploitation of textual resources in life sciences. In *Proc. AIME-2009*, *LNCS* 5651, pp. 206–215. Springer, 2009.

[26] E. Oren, C. Guéret, and S. Schlobach. Anytime query answering in RDF through evolutionary algorithms. In *Proc. ISWC-2008*, *LNCS* 5318, pp. 98–113. Springer, 2008.

[27] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proc. WWW-2007*, pp. 697–706. ACM Press, 2007.

[28] E. Thomas, J. Z. Pan, and D. H. Sleeman. ONTOSEARCH2: Searching ontologies semantically. In *Proc. OWLED-2007*, *CEUR Workshop Proceedings* 258. CEUR-WS.org, 2007.

[29] T. Tran, P. Cimiano, S. Rudolph, and R. Studer. Ontology-based interpretation of keywords for semantic search. In *Proc. ISWC/ASWC-2007*, *LNCS* 4825, pp. 523–536. Springer, 2007.

[30] G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. Sig.ma: Live views on the Web of data. In *Proc. WWW-2010*, pp. 1301–1304. ACM Press, 2010.

[31] YahooSearchBoss. http://developer.yahoo.com/search/boss/.

[32] Yahoo!SearchMonkey. http://developer.yahoo.com/searchmonkey.

[33] W3C. SPARQL Query Language for RDF, 2008. W3C Recommendation (15 January 2008). Available at http://www.w3.org/TR/rdf-sparql-query/.

[34] G. Zenz, X. Zhou, E. Minack, W. Siberski, and W. Nejdl. From keywords to semantic queries — Incremental query construction on the Semantic Web. *J. Web Sem.*, **7**(3):166–176, 2009.

Semantic Web 1 (2010) 117–119
DOI 10.3233/SW-2010-0024
IOS Press

Towards the ubiquitous Web

Editor(s): Krzysztof Janowicz, Pennsylvania State University, USA; Pascal Hitzler, Wright State University, USA
Solicited review(s): Marta Sabou, MODUL University Vienna, Austria; Boyan Brodaric, Geological Survey of Canada, Canada

Andreas Hotho [a,b,*] and Gerd Stumme [a,c]

[a] *Knowledge & Data Engineering Group, Department of Mathematics and Computer Science, University of Kassel, Wilhelmshöher Allee 73, 34121 Kassel, Germany*
http://www.kde.cs.uni-kassel.de/
[b] *Data Mining and Information Retrieval Group, Department of Mathematics and Computer Science, University of Wuerzburg, Am Hubland 97074 Wuerzburg, Germany*
http://www.is.informatik.uni-wuerzburg.de/
[c] *Research Center L3S, Appelstr. 9a, 30167 Hannover, Germany*
http://www.l3s.de/

Abstract. Today, we observe the amalgamation of the Social Web and the Mobile Web, which will ultimately lead to a Ubiquitous Web. The integration and aggregation of the different kinds of available data, and the extraction of useful knowledge and its representation has become an important challenge for researchers from the Semantic Web, Web 2.0, social network analysis and machine learning communities. We discuss the Ubiquitous Web vision, by addressing the challenge of bridging the gap between Web 2.0 and the Semantic Web, before widening the scope to mobile applications.

Keywords: Ubiquitous Web, Semantic Web, Web 2.0, data mining, machine learning, social network analysis

1. Introduction

Highly popular user-centered applications such as blogs, social tagging systems, and wikis have come to be known as "Web 2.0" or the "Social Web" [6]. At the same time, mobile phones became more and more powerful and are equipped with more and more sensors, giving rise to Mobile Web applications. Today, we observe the amalgamation of these two trends, leading to a Ubiquitous Web, whose applications will support us in many aspects of the daily life at any time and any place. The integration of the different kinds of available data, their integration and aggregation, and finally the extraction of useful knowledge and its representation has become an important challenge for different research communities, since it requires the confluence of previously separated lines of research. Consequently, the last years have seen increasing collaboration of researchers from the Semantic Web, Web 2.0, social network analysis and machine learning communities. Applications that use these research results are achieving economic success. Data now become available that allow researchers to analyze the use, acceptance and evolution of their ideas.

In this position paper, we will discuss the Ubiquitous Web vision in two steps. First, we will address the challenge of bridging the gap between Web 2.0 and the Semantic Web, before widening the scope to mobile applications.

2. Bridging the gap between Web 2.0 and the Semantic Web

A major reason for the immediate success of Web 2.0 systems is their high ease of use. The result is that the "wisdom of the crowd" and the wisdom of the experts are converging. The online encyclopedia

*Corresponding author. E-mail: hotho@informatik.uni-wuerzburg.de.

Wikipedia, for instance, reaches (and in some areas even surpasses) the quality of traditional dictionaries [3]. We anticipate that wikis, resource sharing systems, and blogs are only the first appearances of an emerging family of web cooperation tools. These sites do not only provide content but also generate an abundance of weakly structured metadata. A good example is tagging. Here, users add keywords from an uncontrolled vocabulary, called tags, to a resource. Such metadata are easy to produce, but lack any kind of formal grounding, as used in the Semantic Web.

The Semantic Web complements the bottom–up effort of the Web 2.0 community in a top–down manner. Its central point is a stronger knowledge representation, based on some kind of ontology with a fixed vocabulary and typed relations [5]. Such a structure is typically something users implicitly have in mind when they provide their content in Web 2.0 systems. However, for further use, this structure is hidden in the content and needs to be extracted. In the Semantic Web community, such approaches are known as Ontology Learning [1]. Techniques to analyze network structures or weak knowledge representations, such as those found in the Web 2.0, have also a long tradition in different other disciplines, like social network analysis, machine learning and data mining. These kinds of automatic mechanisms are necessary to extract the hidden information and to reveal the structure in a way that the end user can benefit from it. Using established methods to represent knowledge gained from unstructured content will also be beneficial for the Web 2.0 in that it provides Web 2.0 users with enhanced Semantic Web features to structure their content.

Besides the application of Semantic Web technology, it may also be beneficial to consider more light weight knowledge representations, since not always approaches with strong formal semantics are needed. One example are statistical representations, such as association rules, tag similarity measures, and similarity measures in search engines or recommender systems. A careful analysis of the intended application will decide the way to be followed.

The main research question can be summarized as follows: *How will current and emerging Web 2.0 systems support untrained users in sharing knowledge on the Web within the next years?* The scientific challenge is to develop "minimal-invasive" and scalable techniques for large, web-wide spread user communities for knowledge sharing. While knowledge acquisition and management has a long research history, the new aspect of the Web 2.0 is a) the real large number of users who are willing to share their knowledge but b) who are very selective in participating and will stop their cooperation soon if the barriers are set too high. An important requirement is thus how to build, from the uncoordinated input of many people, where each individual is providing very little and/or unstructured input only, a shared knowledge space which allows for similar benefits as those usually promised for approaches with one central, well-designed, heavyweight ontology. This representation will probably not be presented to the users, as the interaction has to be kept as simple as possible, but will be the basis for the systems' enhanced functionalities.

3. The ubiquitous Web

The emergence of ubiquitous computing [7] has started to create new environments consisting of small, heterogeneous, and distributed devices that foster the social interaction of users in several dimensions. Similarly, the upcoming Social Semantic Web also integrates the user interactions in social networking environments. For instance, nowadays modern smartphones allow everyone to have access to the WWW at every place and at every time. At the same time, these systems are equipped with more and more sensors. Typical sensors in today's smartphones are measuring proximity, ambient light, acceleration, loudness, moisture, geographic north. Furthermore, access to the most prominent Web 2.0 platforms – in particular Facebook, Flickr, Youtube – is frequently pre-installed by the vendor. This example shows that the worlds of WWW, Web 2.0, the Mobile Web, and sensor technology are rapidly amalgamating. Going even one step further, we assume the rapid convergence of the Ubiquitous Web with the Internet of Things (cf. [2]) – more and more, the real world that is surrounding us will have its digital counterpart.

Applications in the Ubiquitous Web will thus rely on a mix of data from sensors, social networks and mobile devices. These data need to be integrated, aggregated, and analyzed by means of Data, Text, and Web Mining techniques to all for semantic and/or statistical representations of knowledge, which will then fuel the ubiquitous applications.

Mining in ubiquitous and social environments is thus an emerging area of research focusing on advanced systems for data mining in such distributed and network-organized environments. It also integrates some related technologies such as activity recognition,

Web 2.0 mining, privacy issues and privacy-preserving mining, predicting user behavior, etc. (cf. [4]).

In typical ubiquitous settings, the mining system can be implemented inside the small devices and sometimes on central servers, for real-time applications, similar to common mining approaches. However, the characteristics of ubiquitous and social mining are in general quite different from current mainstream data mining and machine learning. Unlike in traditional data mining scenarios, data does not emerge from a small number of (heterogeneous) data sources, but potentially from hundreds to millions of different sources. As there is only minimal coordination, these sources can overlap or diverge in any possible way.

Semantic Web technology can bridge the gap between all kinds of information independent of its source and its origin and can be used as a starting point to put everything together. The real world information gathered by sensors will be used by applications running on mobile devices and will be connected with the information of their users from the social web. Semantic Web technology may become the right knowledge representation for connecting these worlds.

4. Conclusion

Today, we see the first steps towards an integration of the Social, Mobile and Semantic Webs. This path allows for exciting challenges for researchers of different communities. New insights provided by machine learning and social network analysis techniques will lead to a new type of knowledge. We envision that research in this area will be of growing interest, as the automatic extraction of knowledge from weakly structured sources contributed by a huge mass of users and the combination with structured knowledge will be an important basis for the Semantic Web. It will lead to a broad range of new applications, which allow for combining knowledge of different types, levels and from different sources to reach their goals. The upcoming Ubiquitous Web is one target application area which will benefit from the newly integrated knowledge.

References

[1] P. Cimiano. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications*. Springer-Verlag, New York, Inc., Secaucus, NJ, USA, 2006.

[2] S. Dodson. Internet of things. *The Guardian*, 2003.

[3] J. Giles. Internet encyclopaedias go head to head. *Nature*, **438**(7070):900–901, 2005.

[4] M. May, B. Berendt, A. Cornuéjols, J. Gama, F. Giannotti, A. Hotho, D. Malerba, E. Menesalvas, K. Morik, R. Pedersen, L. Saitta, Y. Saygin, A. Schuster, and K. Vanhoof. Research challenges in ubiquitous knowledge discovery. In *Next Generation of Data Mining (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series)*. Chapman & Hall/CRC, 1 edition, 2008.

[5] S. Staab and R. Studer, editors. *Handbook on Ontologies*. Springer, Berlin, 2 edition, 2009.

[6] O. Tim. What is web 2.0? design patterns and business models for the next generation of software. http://oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html, 2005.

[7] M. Weiser. Ubiquitous computing. *Computer*, **26**(10):71–72, 1993.