Mind the Web

Valentin Zacharias[1], Andreas Abecker[1], Denny Vrandecic[2], Imen Borgi[1],
Simone Braun[1], and Andreas Schmidt[1]

[1] FZI Research Center for Information Technologies at the University of Karlsruhe
{zach|abecker|borgi|braun|aschmidt}@fzi.de
http://www.fzi.de/ipe
[2] Institut AIFB, Universitaet Karlsruhe (TH)
dvr@aifb.uka.de
http://www.aifb.uni-karlsruhe.de/

Abstract. This paper argues that a significant part of today's Semantic Web research is still dominated by ideas from centralized databases. Furthermore, the main thread of reasoning research focusses on approaches that can never scale to anything similar to the Web. Starting from these negative observations we argue that emergent semantics and ontology maturing are more suitable approaches for dealing with ontologies on the Web. Similarly, a few approaches for more Semantic Web appropriate reasoning exist, but are in dire need of realistic use cases.

1   Introduction

A main element of the Semantic Web (SW) vision is the idea of having data on the Web described such that it can be used by machines for more than just retrieval and visualization; in particular, to enable computer agents to automatically interact in solving sophisticated tasks for their users. Following [1], the SW vision is about interlinking the world's knowledge into a single global system.

The Semantic Web differs from previous attempts at building knowledge based systems (KBS) in the combination of universality, decentrality and sheer size. Universality means that the SW, like the current Web, knows no domain, transcends language and culture, and contains data about everything. From this follow three obvious observations that will be the basis of this paper:

(1) Web scale is not just *a bit larger*: Extrapolating conservatively from January 2005 [2] and assuming that each web page could be represented by a meager 20 assertions we arrive at a size estimate for the public web of more than 1 trillion assertions. Assuming that each human will be described by 10.000 triples we can expect the Semantic Web to grow to 100 trillion triples [3]. Both of these estimates are many dimensions larger than what today's knowledge representation and reasoning (KR&R) techniques were developed for.

(2) Ontologies are always changing: Ontologies are changed because an error is uncovered, new information becomes available, because the domain of an ontology is changing or because the view on the domain has shifted. On the

Web, without centralized control and in the face of the ever accelerating pace of the accumulation of humankind's knowledge, this change will be continuous and measured in milliseconds.

(3) There is no right ontology: A conceptualization underlying an ontology is always an abstract, simplified view of the world created for some purpose; and as there are uncountably many different views and purposes, there are innumerable ontologies. For instance Google Base—a service allowing the sharing of simple structured data—saw 100.000 unique schemata only one year after its opening [4]. Based on this observation we can assume a future Semantic Web to have millions of ontologies—or even billions, should the vision of the Social Semantic Desktop become reality.

Starting from these observations we argue that large parts of the SW research (including much of our own work) fail to address what is specific about the Semantic Web. Similar critiques have been expressed earlier, e.g. in [5, 6, 3, 7]. However, in addition to setting a different focus, we go beyond these works by highlighting some directions not yet in the SW mainstream that do address some of the critical issues. We structure this discussion paper around two big challenges facing SW research: (1) emergence and change on the Semantic Web and (2) reasoning. In the concluding remarks we touch on some equally valid points that we could not elaborate further due to space constraints.

2   Emergence and Change on the Semantic Web

The main thread within the Semantic Web community that deals with changing ontologies is based on the ideas of schema evolution and versioning of databases [8, 9]. As such it has been mainly concerned with the question of how to keep one ontology consistent and how to propagate changes to dependent elements, requiring centralized control over the ontology as well as over all uses of it (e.g., [10]). Over time, more aspects of the core SW properties have been added as complications, like the acceptance that multiple version of an ontology are bound to exist [8] or that there might not be a central repository of changes, but changes need to be detected [11]. However, while this line of research has led to great advances, its roots in a world of centralized control also lead to the ignorance of other interesting research questions:

For instance, how a shared conceptualization can emerge from many independently developed conceptualizations; How a conceptualization of a domain can move from informal to a highly formal ontology; or, how the knowledge can be harnessed that multiple ontologies evolved from a single source.

Emergent Semantics And Folksonomies: Under the labels of *Emergent Semantics* and *Folksonomies*, research threads have come up that represent a different approach: Putting the focus firmly on the distributed nature of the SW. There is the view of shared understanding arising bottom-up, through interaction, construction and communication [12–14]. Emergent semantics is seen as an agreement within a specific context of common interpretations which is con-

structed in an incremental process of negotiations and local interactions between self-organizing, autonomous agents.

A framework for large-scale distributed systems proposed by [15] shows how to obtain global agreements from such pair-wise and local interactions through so called "semantic gossiping". Agents rely on their own ontology and provide mappings for adjacent agents, but they can also learn new translations by routing queries whereas the information quality is measured along feedback cycles. Thus, connections are established on demand and through the spreading of simple mappings global consensus can be obtained gradually .

A related phenomenon are folksonomies—informal social organization systems emerging in Web2.0 social resource-sharing systems. Studies have shown that in folksonomies common vocabularies evolve during usage (e.g., [16–18]). For instance, [19] have revealed a stabilization of tag distributions into power law distributions. Based on the tripartite model of semantic-social networks developed by [20], they have also shown a simple methodology to create an ontology from this common vocabulary. [21] and [22] have developed other methods for the detection of emergent semantics in folksonomies.

While active and exciting, the fields of emergent semantics and folksonomies are not yet integrated into the SW mainstream. On the one hand, these methods still very much deal with very lightweight formalisms, and on the other hand the semantic structures produced are statistic by nature and hence lose information when transferred into rigid, non-fuzzy, logic formalisms.

Ontology Maturing: The *ontology maturing* [23] approach considers not only the technical side of ontology evolution, but also its interaction with social aspects. It can be characterized by the following assumptions: (1) Ontology building is a constructivist learning process, it is not just about eliciting knowledge and formalizing it according to a particular formalism. Rather this construction process itself is a learning process in which the involved individuals deepen their understanding of the real world and of an (appropriate) vocabulary to describe it. (2) Ontologies continuously evolve in work (i.e., usage) processes. Ontology building is not supposed to be a one-time activity of an expert committee, but rather a sustainable process of continuous evolution. Concepts in the ontology undergo a process of continuous evolution where ideas and understanding emerge implicitly in daily work and mature only gradually through the interaction with others. The notion of maturity can be characterized along three different dimensions: the level of social agreement (corresponding to the understanding of ontology as a shared understanding), the level of formality (corresponding to the "formal conceptualization") and the level appropriateness (how well an ontology is appropriate for the task at hand). An important conclusion from such a maturing perspective is that we have to consider, at every instant in time, the coexistence of various levels of formality within one ontology and various levels of social agreement among ontologies. The focus of Semantic Web technology then, must be not only on static snapshots, but also on the support of the transitions. In the future we plan to extend this line of research by studying maturing processes in wiki (e.g [24]) and in particular semantic wiki systems. We will also

continue to develop and evaluate the SOBOLEO [25] and ImageNotion [26] tools as support for ontology maturing.

3 Reasoning over the Semantic Web

Formally well-founded, powerful reasoning over expressive background knowledge and metadata on the Web is one of the key elements of the Semantic Web definition which distinguishes the area from conventional Web research or federated databases. Considering this, there are surprisingly few survey papers on the topic: A well-structured survey of reasoning use cases in the Semantic Web with the derived implications for required reasoning tasks, logical formalisms, software architecture settings, expected size and characteristics of ontologies and metadata still has to be written. Since such application-specific considerations have a huge impact on reasoning approach and feasibility, we constrain ourselves in this short paper on a few general, abstract remarks which are valid for the vast amount of SW reasoning research—well-aware of the fact that for each statement there may be a counter-example in some specific project.

First let us state that contemporary RDFS triple stores (such as Sesame, YARS, JENA, Kowari, AllegroGraph) can manage and query some 10^8 triples in the sub-second area (cp. [27, 28]). Even slight extensions towards more powerful modeling and reasoning capabilities, however, lead to serious performance decreases, even for the most developed tools. For instance, the OWLIM [29] OWL + DLP reasoner (on top of Sesame) only manages to process some 10^7 explicit statements with an logarithmic to linear complexity in the number of statements by using in-memory reasoning (!) and an extremely expensive update procedure for the fully materialized knowledge base (KB).

It becomes much worse if we consider tools able to deal with the expressive power of arbitrary rules or more powerful Description Logics dialects. Even high-end implementations grounded in 20 years of AI and Deductive Database research are obviously completely unable to become faster at a speed at least somewhat "comparable" to the growth rate of the Semantic Web's data and knowledge base. During the last 15 years, an enormous amount of resources has been invested in better understanding, theoretically analyzing and refining KR&R approaches which can, *by their very nature* never face the challenges of SW. Already in 2002, van Harmelen pointed out a number of essential KR&R problems of the Semantic Web that "conventional" KBS technology would not be able to address [5]. Until 2006, the situation had not really been improved, but the topics of scalability, ontology evolution, and mapping had at least been noticed by the SW community [6]. Still in 2006, Pellet—which happens to be today's most advanced ABox reasoner for OWL-DL—is in the area of taking seconds for answering queries over the 100.000-instances LUBM benchmark [30]. Similar results hold true for Kaon-2, also on KBs of moderate size [31]. Following [32], most OWL reasoners do not even provide for secondary storage mechanisms which shows that such developments cannot be taken serious from a real-world oriented software engineering point of view. Now, in 2007, Fensel and van Harme-

len diagnose the situation as more or less hopeless [3]—if we insist on traditional, correct and complete, logic-based reasoning.

In the last few years, there is some, yet under-developed, but maybe promising, work in boosting reasoning performance:

– Under labels such as *sub-ontology extraction, KB partitioning, ontology winnowing, or summary ABoxes*, several authors try to reduce KBs to only those fragments really relevant for the specific reasoning task at hand ([32–35]).
– Obviously, much can be gained in practical applications by employing *special-purpose reasoners*, e.g., for temporal or geospatial reasoning ([36, 37]). Other domain and application specific special-purpose reasoners can be imagined.
– Several techniques well-known from deductive databases can be applied, such as *incremental reasoning* and *deductive closure materialization* [38, 39] or *query reorderings and optimizations* [40, 41].

Though such approaches may increase reasoning efficiency by up to one order of magnitude, they can hardly keep the pace of SW growth. Hence we have to think about *fundamentally different* approaches to reasoning. One idea, probably indeed the only imaginable approach to defy content growth while keeping the traditional notion of correct and complete inferences, is **massive parallelization**: For instance, based on the theory of *distributed description logics*, the DRAGO system [42] shows how tableau computations can be distributed to many machines for speeding up inferences. In other pioneering work, [43] implemented consequence finding in propositional logic, distributed over 1000 peers. Grid-based distributed inference has also been suggested by some authors. It also sounds promising to examine the extent to which SPARQL query answering can be parallelized.

So far, we kept the idea of deductive, correct and complete reasoning which is definitely not analogue to the human way of thinking and seems to be the very reason for much of the scalability problems in the Semantic Web—not to mention the fact that it assumes a consistency and completeness of the underlying KB which is by far not given in the Web. Hence, other approaches which give up these assumptions, seem very promising:

Approximate reasoning [44, 45] has recently been considered as a necessity to deal with sloppy knowledge and has been suggested for SW problems such as query answering or ontology mapping. [46] propose a trade-off between efficiency and precision by relaxing a query and approximating its answer in a question answering task. Basically, this is similar to the idea of anytime algorithms that trade resource consumption against, e.g., completeness of reasoning. A very recent work applies approximate reasoning to ontology mapping by using Google distance [47]. [48] gained significant performance improvements by disregarding non-Horn aspects of OWL DL reasoning, thus coming to an unsound, but complete inference procedure. Initial approximate reasoning ideas have been developed for several ABox reasoning tasks such as instance retrieval or conjunctive query answering [49]. All these works can be considered in an initial stadium and are performed only in very few research groups. Whereas complete and sound reasoning procedures are hardly massively parallelizable on

such standard infrastructures like MapReduce, it can be imagined that a certain fragment of description logics, probably under specific structural constraints, can be answered approximatively by a parallel system.

Logical next steps in researching reasoning procedures for the SW could be to look beyond the description logics paradigm and examine, in the SW context, the whole bunch of Expert System, Theorem Proving and Logic Programming work under reasoning paradigms such as *possibilistic logic, defeasible reasoning, plausible inferences, or argumentative inference.* While all our arguments, up to now, were targeting efficiency aspects, the latter approaches are also suitable to address aspects such as incomplete, uncertain and inconsistent KBs—a natural phenomenon in the WWW.

Combining retrieval and reasoning: [50] use statistically quantified ontologies together with a careful application of the k-top query algorithms to define a relatively scalable semantic search engine. In [3], an approach is sketched that uses retrieval techniques to find a small subset of statements that is then used for reasoning. Finally, [51] shows (in a setting with unstructured data) how redundancy in the data set can be used to reduce the need for reasoning.

We have seen that existing reasoning fails and will continue to fail to address the problem of reasoning with KBs as large and diverse as the Web. We have pointed to some approaches that seem to have the potential to scale to such settings. These heuristic and statistic approaches, however, will continue to rely on the formal study of semantics as benchmark, in order to define *what* should be concluded. Even in this role as benchmark, many logics currently under discussion seem insufficient. For example, contradictions may not always be something to be removed but may identify a difference in opinion that we wish to preserve—either to further examine it or to give different results tailored to user groups [52]. Even then, however, the research into this topics risks being of purely academic interest unless concrete use cases are identified first.

4 Concluding Remarks

In this short position paper, we could not touch on other corollaries of the core Semantic Web properties of universality, decentrality and size. Probably most important among these is the issue of trust—one of the most neglected issues in Semantic Web research. We refer the interested reader to [7], where this issue has been examined in some detail. Another core issue is the mapping and alignment of millions of ontologies—here the interested reader could start with the very recent approaches PowerMap [53] and PayGo [4], still preliminary work that could actually tackle this challenge.

References

1. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American **284**(5) (2001) 34–43

2. Gulli, A., Signorini, A.: The Indexable Web is More Than 11.5 Billion Pages. In: Proc. WWW-2005. (2005)
3. Fensel, D., van Harmelen, F.: Unifying Reasoning and Search to Web Scale. IEEE Internet Computing **11**(2) (March/April 2007) 94–96
4. Madhavan, J., et al.: Web-scale Data Integration: You Can Only Afford to Pay As You Go. In: Proc. 3rd Biennial Conf. on Innovative Data Systems Research. (2007)
5. van Harmelen, F.: How the Semantic Web will change KR: Challenges and Opportunities for a New Research Agenda. The Knowledge Engineering Review **17**(1) (2002) 93–96
6. van Harmelen, F.: Semantic Web Research anno 2006: Main Streams, Popular Fallacies, Current Status and Future Challenges. In Klusch, M., et al., eds.: Proc. CIA-2006. Number 4149 in LNAI, Springer (2006) 1–7
7. Kalfoglou, Y., Alani, H., Schorlemmer, M., Walton, C.: The Emergent Semantic Web and Overlooked Issues. In: Proc. ISWC'04. (2004) 576–591
8. Noy, N., Klein, M.: Ontology Evolution: Not the Same as Schema Evolution. Knowl. Inf. Syst. **6**(4) (2004) 428–440
9. Roddick, J.: A Survey of Schema Versioning Issues for Database Systems. Information and Software Technology **37**(7) (1995) 383–393
10. Stojanovic, L., Maedche, A., Motik, B., Stojanovic, N.: User-Driven Ontology Evolution Management. In: Proc. EKAW-02. (2002) 285–300
11. Klein, M., Noy, N.: A Component-Based Framework for Ontology Evolution. In: IJCAI-03 Workshop on Ontologies and Distributed Systems. (2003)
12. Aberer, K., et al.: Emergent Semantics Principles and Issues. In Lee, Y.J., Li, J., Whang, K.Y., Lee, D., eds.: Proc. DASFAA'04, Springer (2004) 25–38
13. Cudr-Mauroux, P., et al.: Viewpoints on Emergent Semantics. J. Data Semantics VI **4090** (2006) 1–27
14. Staab, S., Santini, S., Nack, F., Steels, L., Maedche, A.: Emergent Semantics. IEEE Intelligent Systems **17**(1) (2002) 78–86
15. Aberer, K., Cudré-Mauroux, P., Hauswirth, M.: The Chatty Web: Emergent Semantics Through Gossiping. In: Proc. WWW-03, ACM Press (2003) 197–206
16. Sen, S., et al.: Tagging, Communities, Vocabulary, Evolution. In: Proc. CSCW'06, ACM Press (2006) 181–190
17. Golder, S., Huberman, B.: The Structure of Collaborative Tagging Systems. J. Information Sciences **32**(2) (2006) 198–208
18. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Position Paper, Tagging, Taxonomy, Flickr, Article, ToRead. In: Collaborative Web Tagging Workshop. (2006)
19. Halpin, H., Robu, V., Shepard, H.: The Dynamics and Semantics of Collaborative Tagging. In: Proc. SAAW'06. (2006)
20. Mika, P.: Ontologies Are Us: A Unified Model of Social Networks and Semantics. In Gil, Y., et al., eds.: Proc. ISWC-05. Volume 3729 of LNCS., Springer (2005) 522–536
21. Hotho, A., Jschke, R., Schmitz, C., Stumme, G.: Emergent Semantics in BibSonomy. In: Proc. Workshop on Applications of Semantic Technologies, Informatik 2006. (October 2006)
22. Zhang, L., Wu, X., Yu, Y.: Emergent Semantics from Folksonomies: A Quantitative Study. In Spaccapietra, S., Aberer, K., Cudr-Mauroux, P., eds.: J. Data Semantics VI. Volume 4090 of LNCS., Springer (2006) 168–186
23. Braun, S., Schmidt, A., Walter, A., Nagypal, G., Zacharias, V.: Ontology Maturing: a Collaborative Web 2.0 Approach to Ontology Engineering. In: WWW-07 Workshop on Social and Collaborative Construction of Structured Knowledge. (2007)

24. Braun, S., Schmidt, A.: Wikis as a technology fostering knowledge maturing: What we can learn from wikipedia. In: 7th International Conference on Knowledge Management (IKNOW '07), Special Track on Integrating Working and Learning in Business (IWL). (2007)
25. Zacharias, V., Braun, S.: Soboleo - social bookmarking and lightweight ontology engineering. In: Workshop on Social and Collaborative Construction of Structured Knowledge (CKC), 16th International World Wide Web Conference (WWW 2007). (May 2007)
26. Walter, A., Nagypal, G.: Imagenotion - methodology, tool support and evaluation. In: Proceedings of the ODBASE 2007 27. - 29. November 2007, Algave, Portugal. (2007)
27. Portwin, K., Parvatikar, P.: Building and Managing a Massive Triple Store: An Experience Report. In: XTech 2006: Building Web2.0. (2006)
28. Wood, D.: Scaling the Kowari Metastore. In et al., M.D., ed.: WISE 2005 Workshops. Volume 3807 of LNCS., Springer (2005) 172–181
29. Kiryakov, A., Ognyanov, D., Manov, D.: OWLIM – A Pragmatic Semantic Repository for OWL. In Dean, M., et al., eds.: WISE 2005 Workshops. Volume 3807 of LNCS., Springer (2005) 172–181
30. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A Practical OWL-DL reasoner. J. Web Semantics (2007) To appear.
31. Motik, B., Sattler, U.: A Comparison of Reasoning Techniques for Querying Large Description Logic ABoxes. In Hermann, M., Voronkov, A., eds.: Proc. LPAR 2006. Volume 4246 of LNCS., Springer (2006) 227–241
32. Fokoue, A., Kershenbaum, A., Ma, L., Schonberg, E., Srinivas, K.: The Summary Abox: Cutting Ontologies Down to Size. In: Proc. ISWC-06. Volume 4273 of LNCS., Springer (2006) 343–356
33. Alani, H., Harris, S., O'Neil, B.: Winnowing Ontologies Based on Application Use. In Sure, Y., Domingue, J., eds.: Proc. ESWC-2006, Springer (2006) 185–199
34. Bhatt, M., Flahive, A., Wouters, C., Rahayu, J., Taniar, D., Dillon, T.: A Distributed Approach to Sub-Ontology Extraction. In: Proc. AINA-2004, IEEE Computer Society (2004) 636–641
35. Guo, Y., Heflin, J.: A Scalable Approach for Partitioning OWL Knowledge Bases. In: 2nd Int. Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2006), Springer (2006) 636–641
36. Ohlbach, H.J.: Computational Treatment of Temporal Notions: The CTTN-System. In Fages, F., Soliman, S., eds.: Principles and Practice of Semantic Web Reasoning (PPSWR-2005). Volume 3703 of LNCS., Springer (2005) 133–144
37. Bry, F., Lorenz, B., Ohlbach, H., Rosner, M.: A Geospatial World Model for the Semantic Web. In Fages, F., Soliman, S., eds.: Principles and Practice of Semantic Web Reasoning (PPSWR-2005). Volume 3703 of LNCS., Springer (2005) 145–159
38. B. Parsia, C. Halaschek-Wiener, E.S.: Towards Incremental Reasoning Through Updates in OWL DL. In: Proc. WWW-2006. (2006)
39. Stuckenschmidt, H., Broekstra, J.: Time - Space Trade-Offs in Scaling up RDF Schema Reasoning. In Dean, M., et al., eds.: WISE 2005 Workshops. Volume 3807 of LNCS., Springer (2005) 172–181
40. Ünalir, M., Özacar, T., Öztürk, Ö.: Reordering Query and Rule Patterns for Query Answering in a Rete-Based Inference Engine. In et al., M.D., ed.: WISE 2005 Workshops. Volume 3807 of LNCS., Springer (2005) 255–265
41. Ruckhaus, E., Ruiz, E., Vidal, M.: Query Optimization in the Semantic Web. In: Int. Workshop on Applications of Logic Programming in the Semantic Web and Semantic Web Services, Seattle. (2006)

42. Serafini, L., Borgida, A., Tamilin, A.: Aspects of Distributed and Modular Ontology Reasoning. In: Proc. IJCAI'05. (2005)

43. Adjiman, P., Chatalic, P., Goasdou, F., Rousset, M.C., Simon, L.: Distributed Reasoning in a Peer-to-Peer Setting: Application to the Semantic Web. J. AI Research **25** (2006) 269–314

44. Schaerf, M., Cadoli, M.: Tractable Reasoning via Approximation. Artificial Intelligence **74**(2) (1995) 249–310

45. Cadoli, M., Schaerf, M.: Approximate Inference in Default Logic and Circumscription. In: Proc. ECAI-92. (1992) 319–323

46. Stuckenschmidt, H., van Harmelen, F.: Approximating Terminological Queries. In Larsen, H., et al., eds.: 4th Int. Conf. on Flexible Query Answering Systems (FQAS)'02. Number 2522 in Advances in Soft Computing, Springer (2002) 329–343

47. Gligorov, R., ten Kate, W., Aleksovski, Z., van Harmelen, F.: Using Google Distance to Weight Approximate Ontology Matches. In: Proc. WWW-07. (2007)

48. Hitzler, P., Vrandecic, D.: Resolution-Based Approximate Reasoning for OWL DL. In Gil, Y., et al., eds.: Proc. ISWC-2005, Springer (2005) 383–397

49. Wache, H., Groot, P., Stuckenschmidt, H.: Scalable Instance Retrieval for the Semantic Web by Approximation. In Dean, M., et al., eds.: WISE 2005 Workshops. Volume 3807 of LNCS., Springer (2005) 172–181

50. Weikum, G., Graupmann, J., Schenkel, R., Theobald, M.: Towards a Statistically Semantic Web. Volume 3288 of LNCS., Springer (2004) 3–17

51. Brill, E., Dumais, S., Banko, M.: An Analysis of the AskMSR Question-Answering System. In: EMNLP'02: Proc. ACL-02 Conf. on Empirical Methods in Natural Language Processing. (2002) 257–264

52. Ma, Y., Hitzler, P., Lin, Z.: Algorithms for paraconsistent reasoning with OWL. In Franconi, E., Kifer, M., May, W., eds.: The Semantic Web: Research and Applications. Proceedings of the 4th European Semantic Web Conference, ESWC2007, Innsbruck, Austria. Volume 4519 of Lecture Notes in Computer Science., Springer (June 2007) 399–413

53. Lopez, V., Sabou, M., Motta, E.: PowerMap: Mapping the Real Semantic Web on the Fly. In: Proc. ISWC-06. Volume 4273 of LNCS., Springer (2006)