Journal on Data Semantics manuscript No.
(will be inserted by the editor)

Data Semantics on the Web

Heiner Stuckenschmidt

Received: date / Accepted: date

Abstract Data Semantics is a wide area that continuously faces new challenges arising from the invention of new information formats and novel applications. An area that is particularly challenging with respect to identifying, representing and using data semantics is the Web. This paper attempts to characterize the nature and challenges of Data Semantics on the Web as an interesting research area to be covered by the Journal on Data Semantics.

1 Introduction

Data Semantics [Sheth(1996)] is a topic that has been investigated in computer science for more than thirty years. It is typically associated with a formal definition of the intended interpretation of the data often in terms of logic or algebraic formalisms [Veltman(1984)]. Over the time, the goals of defining data semantics as well as the ideal of having a clear formal representation of semantics has not change, what has changed and is constantly changing, however, are ways of capturing and using the semantic of data as well as the formalisms used to represent it. These changes are triggered by new kinds of applications that require new types of data (e.g. geo-information or social tagging) and with advances in the start of the art of data management (e.g. distribution and parallel processing) that has come with new problem with respect to data semantics.

One invention that brought significant changes to the field of data semantics is the Web. The Web fundamentally changed the way data is managed if compared to traditional systems [Fielding and Taylor(2002)].

Heiner Stuckenschmidt
University of Mannheim, B6, 26 68159 Mannheim, Germany
Tel.: +49-621-181 2530
Fax: +49-621-181 2682
E-mail: heiner@informatik.uni-mannheim.de

In traditional systems, the basic idea is to keep the system in a consistent state or to move it from one consistent state into another. On the web, many traditional assumptions of data semantics are not valid any more. This makes it hard to even define the notion of a consistent state. As a consequence, data semantics on the web require new methods and principles to be developed. As an answer to this need, a rather active research filed has evolved on such principles and method within the broader area of semantic web technologies. The goal of this research is roughly to transfer traditional ideas of data semantics to the area of web data taking into account the specific challenges and needs of a data infrastructure like the web[Shadbolt et al(2006)Shadbolt, Berners-Lee, and Hall].

In this paper, I try to define the research area of data semantics on the web by giving an overview over the challenges and ideas in that part of semantic web research concerned with data semantics in a wider sense. We start by discussing the nature of data on the web focussing on the kind of data typically found on the web and specific challenges we have to face in the area of Data Semantics. I argue that the web does not come with challenges that are unique, but that we are facing a unique combination of challenges that establishes Data Semantics on the Web as a research area in its own rights. After this general discussion of the challenges I discuss three basic aspects of Data Semantics on the Web, namely the extraction of semantics from Web data, the representation of the semantic Information on the Web and the Use of such Semantic representations for processing data on the Web. I will try to stay away from specific solutions and applications as much as possible and rather focus on general principles and lines of work. I will conclude with a personal view on important research directions in the field of Data Semantics on the Web that need further attention in future research in this area.

2 The Nature of Web Data

Web Data comes in a variety of forms that have emerged along with the development of the web over the past twenty years. While each of these different types of data have their own characteristics and requires different kinds of processing methods and infrastructures, they also have a lot of commonalities in terms of challenges for data processing.

2.1 Types of Data on the Web

During the past 20 years the web has evolved from a document management system used internally at CERN into a global information medium that becomes more important for all parts of the society including science, business, politics and social relations. This development has taken place in a number of phases that can roughly be associated with different kinds of data shared over the Web. We briefly recall these developments and the different kinds of Web data.

Documents and Web Pages Initially, the Web was created as a hypertext system for sharing research results in terms of manually created HTML pages and documents that are linked to these pages [Berners-Lee et al(1994)Berners-Lee, Cailliau, Luotonen, Nielsen, and Secret]. This phase is sometimes referred to a Web 1.0. While today, web pages can contain many different kinds of multimedia information, the dominant kind of information on web pages is still in the form of natural language texts. Thus, the semantics of web data is to some extend always connected to natural language semantics and being able to process natural language resources is a basic requirement for semantic processing of web data.

Databases and the Deep Web Although text is still the most visible form of data on the Web, a significant amount of data on the web today is structured data from databases that have been linked to web pages. The resulting information space available through the web is often referred to as the 'Deep Web' or the 'Hidden Web' as the data is typically not explicitly represented on web pages and is thus not indexed by conventional search engines, but has to be surfaced by posting queries to database interfaces on web pages [He et al(2007)He, Patel, Zhang, and Chang]. While the semantics of relational data is well understood, the problem of data on the deep web is the need to describe the data indirectly via the available interface.

Semi-Structured Data A more direct way of publishing non textual information on the Web, semi-structured data representation languages based on the Extensible Modelling Language XML have been developed. These languages allow the explicit publication of structured data in form of documents on the web. The obvious advantage over the Deep Web is the possibility to directly index and access the content of the data sets in a document. A prominent example is the resource description format RDF that has become the foundation for a large space publication of free data on the web. A specific feature of semi-structured data is the ability to explicitly link data to descriptions of its semantics in terms of conceptual models. This ability, often referred to as Web 3.0, can be seen as a cornerstone in the definition and processing of data semantics on the Web [Decker et al(2000)Decker, Melnik, van Harmelen, Fensel, Klein, Broekstra, Erdmann, and Horrocks].

Social Media A rather recent development is the spread of so-called social media. While all other kinds of data discussed so far are created by data providers, social media provide data consumers with the possibility to add contents to web pages [Gruber(2008)]. This User-created contents has a number of specifics that need to be taken into account. In particular, social contents is characterized by a very low level of regularity that is even below the regularity of natural language. Further, user-created content is highly diverse and subjective asking for methods able to control these aspects. In contrast to other forms of data, user created contents is often very timely and therefore provides an important indicator for trends and buzzes.

2.2 Challenges in Web data Management

While the nature of the different types of data differs, certain challenges for managing data (semantics) on the Web come from the Web itself and are therefore similar for the different data types. These challenges that are described in the following also define the research area of web data management by setting it aside from classical data management. There is no doubt that similar challenges occur in different areas of data management, the combination of all of the following challenges, however, is rather unique.

Heterogeneity Heterogeneity of data is a fundamental problem on the web. As we have seen above, heterogeneity starts with the problem of having many different kinds of data representing using a variety of representations including free text, XML-languages and relational data. However, heterogeneity of data on the Web is not limited to data formats and representations, but also occurs at the level of conceptual models and terminology used to describe data items, often referred to as semantic heterogeneity [Kashyap and Sheth(1998)]. While semantic heterogeneity is already a major problem when staying inside the relational data model, it becomes a real challenge on the web, where semantic heterogeneity has to be addressed across different types of data.

Distribution One reason for the heterogeneity of Web data on the semantic level is the inherently distributed architecture of the Web that lacks central control mechanisms. Instead authors and users create data completely independent from each other which leads to different representations of the same real world objects. Distribution and the resulting independence of data sources is also a challenge for data management as data processing has to be coordinated across different sources and the consistency of results have to ensured across different locations. This significantly increases the effort that has to be put into data management on the Web.

Change While managing change is a problem in data management in general, it is especially difficult on the web because, as mentioned above, there is no central point of control. Different web sites changes completely independent of each other and there are no mechanisms for propagating or even announcing changes to other sources referring to the changes data. The original approach of coping with change on the web is not to care about it and just ensure that the system remains stable. While this approach has proven to be very effective, it is problematic from the point of data semantics, especially when the meaning of data in one source depends on the interpretation of another [Noy and Klein(2004)].

Scale Probably the most significant challenge of semantic web data management is scale. The web is the largest freely available information resource that ever existed and it is constantly growing. The following statistics underline this aspect.

- In 2005 the size of the web was estimated with at least 11.5 billion pages [Gulli and Signorini(2005)] end of 2011 Google indexes more than 50 billion webpages[1]
- in 2004 the number of deep web sites has been estimated as around 300.000 in the entire Web [He et al(2007)He, Patel, Zhang, and Chang]
- in September 2011 the linked open data cloud was estimated to contain about 30 billion statements in 500 million RDF documents[2]
- datacatalogs.org [3] lists more than 200 data catalogs each pointing to thousands of datasets all over the world.

While the web itself provides a stable infrastructure for dealing with scalability issues, the size of the web makes it very hard to go into data semantics as traditional approaches to semantics (schema information, metadata), only work on a very limited part of the whole web. Further, processing semantics typically is a computationally hard problem that works best on small data sets.

3 Creating Data Semantics

In closed systems, the intended meaning of data is defined by its intended use that is determined by the systems' developers and users and often reflected in the specific schema or the data structures used for representing it. On the web, this is only partially the case. While the intended meaning of data is of course also determined by the intended use, the universal availability of the data via the Web infrastructure encourages the use of data for applications different from the one it was originally intended for. In order to do this in a meaningful way, the intended meaning of the data has to be understood to correctly relate it to the new application. Thus getting hold of the intended meaning, the Semantics of Data on the Web is essential. We can distinguish different general approaches to the problem of understanding the intended meaning of Data on the web.

3.1 Semantics from Models

A viable way of dealing with the problem arising from the attempt to use data for a different purpose than originally intended is to make the intended meaning of the data explicit by publishing the conceptual model in terms

[1] http://www.worldwidewebsize.com/
[2] http://en.wikipedia.org/wiki/Linked_data
[3] http://datacatalogs.org/

of an ontology and linking the data to it using metadata [Hendler(2001)]. Meanwhile this approach is well supported by language standards such as RDF and OWL and can be seen as a cornerstone of data semantics on the web [Horrocks et al(2003)Horrocks, Patel-Schneider, and van Harmelen]. A closer look at this idea reveals, that publishing the ontology along with a data set does not really solve the problem as long as every data set comes with its own ontology. In this case the problem of possible misinterpretations is just lifted from the data to the conceptual level. In order to be able to really interpret the data, what is needed is either a jointly used ontology that is shared between the data source and the potential user of the data. In various domains, such ontologies have been developed that cane be used to assign an agreed interpretation to data in that domain (e.g. [Lindberg et al(1993)Lindberg, Humphreys, and McCray, Ashburner et al(2000)Ashburner, Ball, Blake, Botstein, Butler, Cherry, Davis, Dolinski, Dwight, and Eppig]). Further, some so-called top-level ontologies have been developed that provide a common interpretation at a very high level and can be used to harmonize domain ontologies (e.g. [Niles and Pease(2001), Matuszek et al(2006)Matuszek, Cabral, Witbrock, and Deoliveira]). The other possible solution is the use of semantic mappings either between the ontologies of data provider and consumer or directly between different datasets [Euzenat et al(2011)Euzenat, Meilicke, Stuckenschmidt, Shvaiko, and Trojahn]. Meanwhile it is commonly accepted that semantic mappings between models or data are an important mechanism in the description of data semantics, especially on the Web. In particular, the use of same-as links to indicate different representations of the same real-world object has recently gained a lot of attention in the context of linked data [Halpin et al(2010)Halpin, J. Hayes, McCusker, McGuinness, and Thompson1].

3.2 Semantics from Data

The downside of the model-based approach to data semantics is the effort that has to be spend on creating and maintaining the semantic models, especially if they are built manually. This has lead to various attempts to directly extract intended interpretations from the data and either use it to construct semantic models automatically or to directly support the semantically correct processing of the data. As the initial Web mostly consisted of textual information, a lot of attention has been put on the extraction of semantic representations from text leading to a number of different approaches [Buitelaar et al(2005)Buitelaar, Cimiano, and Magnini]. From a technical point of view, we can distinguish between statistical approaches to text semantics and information extraction approaches try to identify semantics based on grammatical patterns in the text. While statistical approaches usually determine the semantics of textual data in terms of word distributions and co-occurrences, information extraction analyses the inherent semantics of language constructs. With the increasing availability of at least partially structured information on the Web, researchers have started to investigate the use of relational learning approaches to extract conceptual knowledge from partially structured and somewhat heterogeneous information [Lehmann et al(2011)Lehmann, Auer, Bhmann, and Tramp, Vlker and Niepert(2011)]. Both, linguistic analysis and learning from data also plays a significant role in the creation of semantic mappings between models and data sets. Ontology Matching and Data linking are two areas of research concerned with the creation of such mappings make extensive use of linguistic knowledge using both statistical and structural methods and from learning from data. A prominent example of a learning approach to sdata matching is [Doan et al(2003)Doan, Domingos, and Halevy].

3.3 Semantics from Users

A new trend in data semantics is the involvement of the data user in the process of making semantics explicit. Efforts in this direction can broadly be distinguished into such approaches, where the user is explicitly asked to provide semantic descriptions and such where the user is unaware of the fact that she is contributing to the establishment of data semantics. The first type of approaches makes use of web 2.0 technologies for handling user generated contents. In contrast to classical knowledge acquisition scenarios, it is not assumed that the users are experts in the particular domain. The idea is more that asking many users the same question will ultimately generate the right result because a majority of users will give the correct interpretation of the piece of data. The most prominent approach of this type is tagging, where users attach simple descriptions (tags) to information objects. The extraction of deeper semantics, i.e. concept hierarchies from these tags is an area of ongoing research [Specia and Motta(2007)]. The problem with explicit user participation in the creation of semantic descriptions is motivation. The User is typically only willing to contribute if he sees a benefit for himself or for his peer group. This can be achieved by offering value added service that improve with user input or the users have to be motivated in a different way. the most direct way of providing such motivation is to pay the user for his contribution. This way is taken by crowdsourcing platforms such as the Amazon mechanical Turk, that has already successfully been used to let users create semantic descriptions [Eckert et al(2010)Eckert, Niepert, Niemann, Buckner, Allen, and Stuckenschmidt]. The other way is to motivate the user by making it fun to participate. This idea is implemented in so-called games with a purpose that disguise the creation of semantics in an online game that users enjoy playing. Such games are already half way between the two types of user participation as the player of a game is not necessarily aware of his contribution to information semantics [Siorpaes and Hepp(2008)]. An even more indirect way is to just observe user behavior and try to derive information semantics from this behavior [Berendt et al(2011)Berendt, Hollink, Hollink, Luczak-Rsch, Mller, and Vallet].

Search engines, for instance can analyze which of the search results are actually selected by a user or search sessions of one user can be analyzed to find related search terms.

4 Representing Data Semantics

As we have seen, the intended interpretation of Data on the web can be explored in different ways, but it is widely agreed that models of this intended meaning play a central role in intelligent information processing. With the introduction of OWL as a W3C standard for representing ontologies, the representation of data semantics is often associated with it. the spectrum of models for representing Data semantics, however, is actually broader than that. Instead of discussing specific languages that have been proposed for this purpose, we rather discuss different principles underlying these languages that provide different ways of approaching the problem of representing semantics. This of course includes logic as a classical way of representing meaning, but the characteristics of Web Data discussed in section 2 ask for more than a purely logical treatment of data semantics. In particular, successful approaches have to deal with uncertainty and linguistic/pragmatic semantics of data on the Web.

4.1 Logics

The traditional way of defining Data Semantics in in terms of algebraic structures and formal logic using the model- theoretic semantics of the respective formalism as a mathematical framework for defining and analyzing the semantics. In databases, Datalog has become the most important model for talking about the semantics of relational data [Ceri et al(1989)Ceri, Gottlob, and Tanca], in the field of knowledge representation, description logics have been invented as a specific family of logics for talking about conceptual knowledge [Baader et al(2003)Baader, Calvanese, McGuinness, Nardi, and Patel-Schneider]. Although have started to investigate connections between these two families of logics (e.g. [Motik and Rosati(2010)]), there are still two different areas of research with limited interaction. Both formalisms also play a central role in Data Semantics on the Web as the basis for languages like OWL and RDF. However, it has been recognized by different researchers that the special characteristics of Web Data often requires an interpretation that goes beyond the abilities of classical logic. In particular, the notion of consistency that is quite central to logic-based formalisms for capturing data semantics has to be re-investigated as inconsistency is rather a rule than an exception on the Web. Consequently, various extensions to classical logics have been proposed to capture the semantics of heterogeneous and inconsistent data on the Web [Huang et al(2005)Huang, van Harmelen, and ten Teije, Flouris et al(2006)Flouris, Huang, and Jeff Z. Pan]. Another line of research is concerned with the development of methods for using classical models for representing smaller consistent subsets of data on the web and managing these local models on a global level (e.g. [Bouquet et al(2004)Bouquet, Giunchiglia, van Harmelen, Serafini, and Stuckenschmidt]). While the representation and processing of conceptual data models have been in the focus of early investigations, the attention has now shifted towards logical models that support more practical use of data via semantic query processing [Calvanese et al(2007)Calvanese, De Giacomo, Lembo, and und Riccardo Rosati, Lutz et al(2009)Lutz, Toman, and Wolter]. As a result, methods from database theory are becoming more important and are successively adapted to the special needs of Web Data.

4.2 Lexical and Distributional Semantics

In parallel to logic-based approaches to data semantics, a completely different approach to semantics has been developed in the area of language processing. As an answer to the infeasibility of capturing the semantics of natural language, less formal ways of defining and exploring the meaning of terms have been investigated in the area of lexical semantics [Saint-Dizier and Viegas(1995)]. Here, the meaning of terms is described in terms of their relations to other terms. While this idea is similar to that of logic-based ontologies, the relations used are not formally defined, but rather describe the use of a term in textual resources. Using these relations terms can be defined and disambiguated via the related terms [Thomas K. Landauer and Laham(1998)]. An even more light-weight approach to describe the semantics of terms in natural language is via co-occurrence with other terms. While this approach also referred to as the vector space model of semantics [Turney and Pantel(2010)], as the meaning of a document or a term is represented using a term vector, is a purely statistical one, it can also be seen as a simplified version of lexical semantics that only uses a single relation. Due to its simplicity and scalability, lexical and especially distributional semantics has become quite popular in information search and retrieval. The popularity of the distributional model has also been caused by the ability to generate semantic models from large corpora using machine learning techniques. While there are also approaches for learning lexical and logical models of information semantics, the automatic generation of such richer models from data is still an unsolved problem as mentioned above. This means that lexical and distributional models of semantics provide complementary approach to data semantics on the web that comes into play when the creation of rich semantic models or their use is not feasible or too costly (e.g. [Novacek et al(2011)Novacek, Handschuh, and Decker]).

4.3 Models of Uncertainty

The characteristics of Web data described above inevitably leads to uncertainty about both the truth and the intended meaning of statements on the web. Uncertainty about the truth of a statement often arises from doubt in the quality of the source the statement originates from or from statements that have been learned from a data sample or generated using heuristic knowledge. Uncertainty about the intended meaning often originates from vagueness in statements like often found in natural language. When building semantic models about this kind of data, the inherent uncertainty also has to be reflected in the model. While this is not possible in standard languages like OWL or RDF, the extension of logical models of information semantics with representations of uncertainty about truth or intended meaning is an active field of research [Lukasiewicz and Straccia(2008)]. A prominent line of research is concerned with the extension of description logics with concepts from fuzzy theory and related methods [Stoilos et al(2006)Stoilos, Simou, Stamou, and Kollias]. The resulting models allow the representation of vagueness in the subsumption relation as well as in typical properties of concepts. Another line of research is concerned with combining logic-based representations of semantics with probability theory (e.g. [Cali et al(2009)Cali, Lukasiewicz, Predoiu, and Stuckenschmidt]). Here the focus is much more on data and querying than on purely terminological knowledge, often inspired by work in information retrieval [Fuhr(2000)] and probabilistic databases [Dalvi et al(2009)Dalvi, Re, and Suciu].

5 Using Data Semantics

The specific characteristics of Web Data often requires the explicit use of semantic models the data processing. This distinguishes semantic data models on the Web from traditional settings where semantic models like conceptual schemas were primarily used for the design and the documentation of Data and did not play a central role in the actual process of using the data afterwards. Foregoing an explicit use of semantic models is possible if the data is centralized and there is an agreement on the intended meaning and use of the data. This is not the case on the Web as argued above. Consequently, semantic models play an important role in data processing on the web. In particular, there are two basic operations that have been shown to benefit from an explicit use of semantic models, namely the search for and the integration of distributed data sources.

5.1 Semantic Search

Finding information has been a central problem on the Web from its creation on. Meanwhile commercial search engines in particular Google provide excellent support for finding web pages and textual documents based on keyword matching and advanced ranking methods. Traditional web search, however, quickly reaches its limits when the task is not simply to find a document or a web page but to find a specific piece of information (e.g. the number of inhabitants of a country), to collect information about a single unique person (e.g. a person named Michael Schumacher, who is not the famous racing driver) or to find an answer to a complex question (e.g. what causes climate change?). Such advanced scenarios require the use of explicit semantic models at various points in the process [Mangold(2007)]. For instance, semantic models are used to query semi-structured information on the web in order to find specific facts rather than a related document based on structured queries [Hartig et al(2009)Hartig, Bizer, and Freytag]. Semantic models are also used to translate natural language questions into structured queries to enable precise querying [Tablan et al(2008)Tablan, Damljanovic, and Bontcheva]. Further, the use of semantic models to disambiguate query results has a long tradition in information search, not only with respect to distinguishing between different real world objects sharing the same name [Hassell and und I. Budak Arpinar(2006)], but also to distinguish between different possible interpretations of the same word. All kind of semantic models have been used in this context. While logic-based representations of semantics are typically used to support structured querying, lexical and distributional models play a central role in the disambiguation of information. Models of uncertainty have the potential to combine the two aspects in a single step, however, so far their are not used that frequently.

5.2 Data Integration

The nature of Web Data makes data integration a central problem. On the conventional web, the integration is a purely technical one: heterogeneous contents co-exists in different formats and can be accessed through the same infrastructure. The invention of semi-structured data description languages, i.e. XML addresses the data integration problem at the syntactical level: Data is represented using the same format enabling users to process data using the same tools. Models of data semantics comes into play when a syntactic integration is not sufficient, but an integration on a structural and semantic level is needed. It has been widely acknowledged that Ontologies can play a valuable role for semantic data integration by providing a unified structure information from different sources can be linked to and by providing a common interpretation of terminology used in different information sources [Kashyap and Sheth(1998)]. In a similar way, it has been shown that semantic models are important for linking ontologies and schemas to each other: Most semantic matching systems use some form of background knowledge, often in terms of lexical resources or special domain ontologies [Euzenat and Shvaiko(2007)]. There are even approaches that make use of any freely available model on the web. A typical use of semantic models in the context of semantic matching are - as in semantic search - the disambiguation of terms. Beyond that semantic models are used to derive implicit semantic relationships between Data items [Aleksovski et al(2006)Aleksovski, Klein, and Warner ten Kate] and for detecting inconsistencies that arise due to wrong matches [Meilicke et al(2007)Meilicke, Stuckenschmidt, and Tamilin].

6 Research Directions

In this article, I have tried to summarize the main issues connected with research relating to data semantics on the web. Starting with a discussion of the special properties of Web Data that makes it unique I have provided a brief survey of current ideas and principles of generating, representing and using data semantics on the Web. As the topic is a very broad and popular one, a lot of research is being done and will be done on this issue. In this section I want to provide a personal perspective on some lines of work that are very promising topics that either require further investigation to solve existing problems or that have a high potential for creating progress.

6.1 Semantic Processing at Scale

As mentioned in the first section, scale is the dominant characteristic of the Web. While scalability has been achieved for conventional web infrastructure, new forms of data processing on the web, in particular the processing of structured data and semantic models faces significant scaling problems. While there has been progress in the development of scalable infrastructures for storing and querying semi-structured data, they are still far from being 'web-scale' even without taking semantic models into account. It is pretty clear that centralized architectures will never scale to this level, therefore the investigation of decentralized architectures and algorithms remains to be an important issue for web data processing as such. The same holds for semantic models of data. Here there problem of scalability is even more pressing as reasoning with semantic models is known to be computational expensive and show high degree of interaction that makes it hard top distribute. A first step towards scaling up has been achieved by shifting the focus from very expressive languages to comparably simple representations that have a higher chance of scaling, however, there is still not enough work on distributed and parallel approaches for reasoning in these models. Work is needed both on the theory to better understand what kind of models support parallel processing and on practice by building and evaluating systems in the large. Unfortunately, the systematic evaluation of semantic systems at web scale is an open research issue. There is a need for reaching a common understanding of how the claim that a system scales to 'web-scale' can be established beyond doubt, because running real experiments at the scale of the web - whatever that might be - is infeasible. Such an agreement on a feasible way of establishing the claim has to be implemented in terms of datasets and benchmarks that can be used for testing, leading to the problem that there are many different ways of representing and processing semantics as we have seen above. This means that it is very hard top agree on common benchmarks. The good news is that work in this direction has already started to benefit from new developments in the area of distributed computing that has led to the availability of huge computing resources as bookable services that can be used to test systems at large scale.

6.2 Multi-Source Extraction and Bootstrapping

The Knowledge Acquisition Bottleneck is a well known problem associated with semantic models. Creating semantic models by hand is extremely time intensive and costly. This effort is sometimes put up with if the resulting model is highly used and plays an important role in a community. Examples include the international classification of diseases (ICD) and the gene ontology. In the large, however, the manual creation of models by experts is not an option. As discussed in this article, various approaches have been proposed for semi-automatically creating semantic models. So far, there is no silver bullet for creating semantic representations and known methods can often only successfully cover a very specific aspect of semantic models. It seems that pushing the boundaries in this area cannot be achieved by working on a single method. A promising direction is rather to find intelligent ways of combining supplementary approaches and information types. Successful examples include the combination of machine learning and reasoning or the combination of crowdsourcing and information extraction. While so far, approaches typically focus on a single type of information like text or structured data. In order to fully capture the semantics of a domain, any available source can provide a contribution. In particular the use of different types of information or different extraction methods can benefit from bootstrapping approaches that enhance on method with the results of other. While combined methods have bene investigated,work in this area would benefit from a systematic approach to combining different approaches. This way, the web could be populated with a new generation of semantic models that do not reach the level of quality of manually curated models, but that can further enhance the capabilities of semantic search and information integration.

6.3 Combined Representations

We can observe a similar situation with respect to research on the representation of semantic models. While constant progress is made in the different areas of modeling semantics mentioned in this article, it is clear that none of the representations that stays within one of the areas may it be logic, lexical semantics or models of uncertainty will be able to cover all aspects of the semantics of web data. While logic-based representations have proven useful in the context of using structured data on the web and lexical and distributional models provide a well founded basis for capturing the semantics of text, none of these 'pure' formalisms is able to provide solutions for the mix of different types of data that co-exist on the Web. What is needed is a stronger focus on formalisms that combine aspects of these different schools. In particular many applications on the web demand for solutions that can adequately describe the semantics of both structured and unstructured data and are also able to take uncertainty into account. This means that more research is necessary in combining the different approaches. A good example of how useful this combination is the increased use of approaches like Markov-Logic that combine first-order logic with probabilistic models in a theoretically sound way while still supporting practical implementations of inference and learning. It has been shown that such formalisms can improve methods in various areas including natural language processing, knowledge acquisition and data integration. Form a theoretical point of view, combined formalisms need fundamental research on tractable fragments and reasoning methods. On the more practical side, there is a need to integrate combined models with established standards for representing semantics that so far have not made the move beyond purely logical representations.

6.4 Open Analytics

Maybe the most interesting opportunity that comes with the publication of structured data and its semantics on the web is the opportunity to use this data as a basis for analyzing complex problems that could not be addressed so far, because they require the combined observation of data from multiple data sources that were extremely hard or even impossible to access and combine in a meaningful way. The availability of data like public statistics like it has been published at the large in the context of the open government data is of high interest for researchers from various disciplines like economics of social sciences. While publishing this data on the Web already provides significant benefits in terms of transparency, an adequate representation of the meaning of the published data is necessary for being able to draw conclusions across different datasets and supports a semi-automatic analysis of data using analytical methods. This opportunity, we refer to as open analytics could provide significant benefits for different scientific disciplines and public decision making. Developing a reliable infrastructure for capturing the semantics of data on the web and drawing conclusions from the integrated data is a huge opportunity.

7 Conclusions

In summary, Data semantics on the web is both a challenging research topic that needs ideas from different fields of computer science. It thus provides an opportunity to create radically new approaches on the boundaries of disciplines and test results from fields such as databases, information retrieval and artificial intelligence in a new challenging setting, leading to new research questions in the different areas. On the other hand, being able to capture and represent the Semantics of Data on the Web has a huge potential for advanced applications in an area that rapidly gains importance in almost all ares of business and society including electronic commerce, political discourse and scientific exchange. This combination of a long term research challenge and practical significants makes Data Semantics on the Web a topic that promises to remain long-term relevance and is clearly set apart from short term hypes that come and go in the process of scientific discovery.

References

[Aleksovski et al(2006)Aleksovski, Klein, and Warner ten Kate] Aleksovski Z, Klein M, Warner ten Kate Fv (2006) Matching unstructured vocabularies using a background ontology. In: Staab S, Svtek V (eds) Managing Knowledge in a World of Networks, 15th International Conference, EKAW 2006, Podebrady, Czech Republic, Lecture Notes in Computer Science, vol 4248, pp 182–197

[Ashburner et al(2000)Ashburner, Ball, Blake, Botstein, Butler, Cherry, Davis, Dolinski, Dwight, and Eppig] Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, Davis A, Dolinski K, Dwight S, Eppig J (2000) Gene ontology: tool for the unification of biology. Nature Genetics 25:25–29

[Baader et al(2003)Baader, Calvanese, McGuinness, Nardi, and Patel-Schneider] Baader F, Calvanese D, McGuinness D, Nardi D, Patel-Schneider P (2003) The Description Logic Handbook: Theory, Implementation, Applications. Cambridge University Press, Cambridge, UK

[Berendt et al(2011)Berendt, Hollink, Hollink, Luczak-Rsch, Mller, and Vallet] Berendt B, Hollink L, Hollink V, Luczak-Rsch M, Mller K, Vallet D (2011) Usage analysis and the web of data. SIGIR Forum 45(1):63–69

[Berners-Lee et al(1994)Berners-Lee, Cailliau, Luotonen, Nielsen, and Secret] Berners-Lee T, Cailliau R, Luotonen A, Nielsen HF, Secret A (1994) The world-wide web. Communications of the ACM 37(8):76–82

[Bouquet et al(2004)Bouquet, Giunchiglia, van Harmelen, Serafini, and Stuckenschmidt] Bouquet P, Giunchiglia F, van Harmelen F, Serafini L, Stuckenschmidt H (2004) Contextualizing ontologies. Web Semantics: Science, Services and Agents on the World Wide Web 1(4):325–343

[Buitelaar et al(2005)Buitelaar, Cimiano, and Magnini] Buitelaar P, Cimiano P, Magnini B (eds) (2005) Ontology Learning from Text: Methods, Evaluation and Applications. IOS Press

[Cali et al(2009)Cali, Lukasiewicz, Predoiu, and Stuckenschmidt] Cali A, Lukasiewicz T, Predoiu L, Stuckenschmidt H (2009) Tightly coupled probabilistic description logic programs for the semantic web. Journal on Data Semantics XII:95–130

[Calvanese et al(2007)Calvanese, De Giacomo, Lembo, and und Riccardo Rosati] Calvanese D, De Giacomo G, Lembo D, und Riccardo Rosati ML (2007) Tractable reasoning and efficient query answering in description logics: The dl-lite family. Journal of Automated Reasoning 39(3):385–429

[Ceri et al(1989)Ceri, Gottlob, and Tanca] Ceri S, Gottlob G, Tanca L (1989) What you always wanted to know about datalog (and never dared to ask). IEEE Transactions on Knowledge and Data Engineering 1(1):146–166

[Dalvi et al(2009)Dalvi, Re, and Suciu] Dalvi N, Re C, Suciu D (2009) Probabilistic databases: diamonds in the dirt. Communications of the ACM 52(7):86–94

[Decker et al(2000)Decker, Melnik, van Harmelen, Fensel, Klein, Broekstra, Erdmann, and Horrocks] Decker S, Melnik S, van Harmelen F, Fensel D, Klein M, Broekstra J, Erdmann M, Horrocks I (2000) The semantic web: The roles of xml and rdf. IEEE Internet Computing 4(5):63–74

[Doan et al(2003)Doan, Domingos, and Halevy] Doan A, Domingos P, Halevy AY (2003) Learning to match the schemas of data sources: A multistrategy approach. Machine Learning 50(3):279–301

[Eckert et al(2010)Eckert, Niepert, Niemann, Buckner, Allen, and Stuckenschmidt] Eckert K, Niepert M, Niemann C, Buckner C, Allen C, Stuckenschmidt H (2010) Crowdsourcing the assembly of concept hierarchies. In: Hunter J, Lagoze C, Giles CL, Li YF (eds) Proceedings of the Joint Conference on Digital Libraries JCDL-2010, ACM, Gold Coast, Queensland, Australia

[Euzenat and Shvaiko(2007)] Euzenat J, Shvaiko P (2007) Ontology Matching. Springer

[Euzenat et al(2011)Euzenat, Meilicke, Stuckenschmidt, Shvaiko, and Trojahn] Euzenat J, Meilicke C, Stuckenschmidt H, Shvaiko P, Trojahn C (2011) Ontology alignment evaluation initiative: Six years of experience. Journal of Data Semantics XV(158-192)

[Fielding and Taylor(2002)] Fielding RT, Taylor RN (2002) Principled design of the modern web architecture. ACM transactions on Internet technology 2(2):115–150

[Flouris et al(2006)Flouris, Huang, and Jeff Z. Pan] Flouris G, Huang Z, Jeff Z Pan HW Dimitris Plexousakis (2006) Inconsistencies, negations and changes in ontologies. In: Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference AAAI-06, AAAI Press, Boston, Massachusetts, USA

[Fuhr(2000)] Fuhr N (2000) Probabilistic datalog: Implementing logical information retrieval for advanced applications. Journal of the American Society for Information Science 51(2):95–110

[Gruber(2008)] Gruber T (2008) Collective knowledge systems: Where the social web meets the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 6(1):4–13

[Gulli and Signorini(2005)] Gulli A, Signorini A (2005) The indexable web is more than 11.5 billion pages. In: Ellis A, Hagino T (eds) Proceedings of the 14th international conference on World Wide Web (WWW 2005) - Special interest tracks and posters, ACM, Chiba, Japan

[Halpin et al(2010)Halpin, J. Hayes, McCusker, McGuinness, and Thompson1] Halpin H, J Hayes P, McCusker J, McGuinness D, Thompson1 H (2010) When owl:sameas isn't the same: An analysis of identity in linked data. In: Patel-Schneider P, Pan Y, Hitzler P, Mika P, Zhang L, Pan J, Horrocks I, Glimm B (eds) The Semantic Web - 9th International Semantic Web Conference, ISWC 2010, Springer, Shanghai, China, Lecture Notes in Computer Science, vol 6496, pp 305–320

[Hartig et al(2009)Hartig, Bizer, and Freytag] Hartig O, Bizer C, Freytag JC (2009) Executing sparql queries over the web of linked data. In: Bernstein A, Karger D, Heath T, Feigenbaum L, Maynard D, Motta E, Thirunarayan K (eds) The Semantic Web, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, Lecture Notes in Computer Science, vol 5823, pp 293–309

[Hassell and und I. Budak Arpinar(2006)] Hassell J, und I Budak Arpinar BAM (2006) Ontology-driven automatic entity disambiguation in unstructured text. In: Cruz I, Decker S, Allemang D, Preist C, Schwabe D, Mika P, Uschold M, Aroyo L (eds) The Semantic Web, 5th International Semantic Web Conference, ISWC 2006, Athens, GA, USA, Lecture Notes in Computer Science, vol 4273, pp 44–57

[He et al(2007)He, Patel, Zhang, and Chang] He B, Patel M, Zhang Z, Chang KCC (2007) Accessing the deep web. Communications of the ACM 50(5):95–101

[Hendler(2001)] Hendler J (2001) Agents and the semantic web. Intelligent Systems 16(2):30–37

[Horrocks et al(2003)Horrocks, Patel-Schneider, and van Harmelen] Horrocks I, Patel-Schneider P, van Harmelen F (2003) From shiq and rdf to owl: the making of a web ontology language. Web Semantics: Science, Services and Agents on the World Wide Web 1(1):7–26

[Huang et al(2005)Huang, van Harmelen, and ten Teije] Huang Z, van Harmelen F, ten Teije A (2005) Reasoning with inconsistent ontologies. In: Pack-Kaelbling L, Saffiotti A (eds) Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, IJCAI-05, Edinburgh, Scotland, UK, pp 454–459

[Kashyap and Sheth(1998)] Kashyap V, Sheth A (1998) Semantic heterogeneity in global information systems: the role of metadata, context and ontologies. In: Cooperative Information Systems Trends and Directions, Elsevier

[Lehmann et al(2011)Lehmann, Auer, Bhmann, and Tramp] Lehmann J, Auer S, Bhmann L, Tramp S (2011) Class expression learning for ontology engineering. Web Semantics: Science, Services and Agents on the World Wide Web 9(1):71–81

[Lindberg et al(1993)Lindberg, Humphreys, and McCray] Lindberg D, Humphreys B, McCray A (1993) The unified medical language system. Methods of Information in Medicine 32(4):281–291

[Lukasiewicz and Straccia(2008)] Lukasiewicz T, Straccia U (2008) Managing uncertainty and vagueness in description logics for the semantic web. Web Semantics: Science, Services and Agents on the World Wide Web 6(4):291308

[Lutz et al(2009)Lutz, Toman, and Wolter] Lutz C, Toman D, Wolter F (2009) Conjunctive query answering in the description logic el using a relational database system. In: Boutilier C (ed) Proceedings of the 21st International Joint Conference on Artificial Intelligence IJCAI 2009, Pasadena, California, USA, pp 2070–2075

[Mangold(2007)] Mangold C (2007) A survey and classification of semantic search approaches. Journal International Journal of Metadata, Semantics and Ontologies 2(1):23–34

[Matuszek et al(2006)Matuszek, Cabral, Witbrock, and Deoliveira] Matuszek C, Cabral J, Witbrock M, Deoliveira J (2006) An introduction to the syntax and content of cyc. In: Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering, AAAI Press, vol 3864, pp 44–49

[Meilicke et al(2007)Meilicke, Stuckenschmidt, and Tamilin] Meilicke C, Stuckenschmidt H, Tamilin A (2007) Repairing ontology mappings. In: Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, AAAI Press, Vancouver, British Columbia, Canada, pp 1408–1413

[Motik and Rosati(2010)] Motik B, Rosati R (2010) Reconciling description logics and rules. Journal of the ACM 57(5):1–63

[Niles and Pease(2001)] Niles I, Pease A (2001) Towards a standard upper ontology. In: Welty C, Smith B (eds) Proceedings of the 2nd International Conference on Formal Ontology in Information Systems, ACM, Ogunquit, Maine, USA

[Novacek et al(2011)Novacek, Handschuh, and Decker] Novacek V, Handschuh S, Decker S (2011) Getting the meaning right: A complementary distributional layer for the web semantics. In: 10th International Semantic Web Conference ISWC 2011 TSW (ed) Lora Aroyo and Chris Welty and Harith Alani and Jamie Taylor and Abraham Bernstein and Lalana Kagal and Natasha Fridman Noy and Eva Blomqvist, Springer, Bonn, Germany, Lecture Notes in Computer Science, vol 7031, pp 504–519

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

[Saint-Dizier and Viegas(1995)] Saint-Dizier P, Viegas E (eds) (1995) Computational lexical semantics. Cambridge University Press

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

[Sheth(1996)] Sheth A (1996) Panel: Data semantics: what, where and how? In: Meersman R, Mark L (eds) Database Applications Semantics, Proceedings of the Sixth IFIP TC-2 Working Conference on Data Semantics (DS-6), Chapman & Hall, Stone Mountain, Atlanta, Georgia, USA, IFIP Conference Proceedings, vol 74

[Siorpaes and Hepp(2008)] Siorpaes K, Hepp M (2008) Games with a purpose for the semantic web. IEEE Intelligent Systems 23(3):50–60

[Specia and Motta(2007)] Specia L, Motta E (2007) Integrating folksonomies with the semantic web. In: Franconi E, Kifer M, May W (eds) The Semantic Web: Research and Applications, 4th European Semantic Web Conference, ESWC 2007, Springer, Innsbruck, Austria, Lecture Notes in Computer Science, vol 4519, pp 624–639

[Stoilos et al(2006)Stoilos, Simou, Stamou, and Kollias] Stoilos G, Simou N, Stamou G, Kollias S (2006) Uncertainty and the semantic web. IEEE Intelligent Systems 21(5):84–87

[Tablan et al(2008)Tablan, Damljanovic, and Bontcheva] Tablan V, Damljanovic D, Bontcheva K (2008) A natural language query interface to structured information. In: Bechhofer S, Hauswirth M, Hoffmann J, Koubarakis M (eds) The Semantic Web: Research and Applications, 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, Lecture Notes in Computer Science, vol 5021, pp 361–375

[Thomas K. Landauer and Laham(1998)] Thomas K Landauer PF, Laham D (1998) An introduction to latent semantic analysis. Discourse Processes 25(2-3):259–284, special Issue: Quantitative Approaches to Semantic Knowledge Representations

[Turney and Pantel(2010)] Turney P, Pantel P (2010) From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37:141–188

[Veltman(1984)] Veltman F (1984) Data semantics. In: Groenendijk J, Janssen T, Stokhof M (eds) Truth, interpretation, and information: selected papers from the third Amsterdam colloquium, Walter de Gruyter, pp 43–65

[Vlker and Niepert(2011)] Vlker J, Niepert M (2011) Statistical schema induction. In: Proceedings of the 8th Extended Semantic Web Conference (ESWC), Lecture Notes in Compuer Science, vol 6643, pp 124–138