Undefined 1 (2010) 1–5
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*
E-mail: {first.last}@deri.org
[b] *Institut of Applied Informatics and Formal Description Methods – AFIB, Karlsruhe Institute of Technology*
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 [8,6], 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 majority 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

[1] cf. `http://richard.cyganiak.de/2007/10/lod/`

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 it's 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. 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 ontology 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],

[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.

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 [31,33,32]. 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 excellent 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., lexi-

[4] `http://vocamp.org/`

[5] `http://oaei.ontologymatching.org/`

cally 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 [30,23,16] 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. [30,23,16]) and reasoning (cf. [33,14]) over Linked Data. A generic framework for querying and reasoning over annotations (including, e.g., provenance or trust values) of RDF [50,41] may also serve as a useful starting point.

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 lo-

[6] `http://pedantic-web.org/`

cate 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. [24,15]) 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. [33,14]); (iii) ranking: scoring the importance and relevance of given data artefacts for prioritisation of results (cf. [30,23,16]). 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 [24,20,17], querying [34,51] and reasoning [14,57,54,55,34] 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.

[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.

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. Towards Creating Knowledge out of Interlinked Data. *Semantic Web – Interoperability, Usability, Applicability,* 2010. In this issue.

[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. M. (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,* 2010. In this issue.
[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 interoperability problems instead of solving them. *Semantic Web – Interoperability, Usability, Applicability*, 2010. In this issue.
[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.