IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 10, October 2017

Linked Data in Modern Web Applications

Er Robin Kumar[1], Er Anupreet[2]

Assistant Professor, Computer Engineering, YCET, Jammu, India[1,2]

Abstract: Enabling the "Web of Data" has recently gained increased attention, particularly driven by the success of Linked Data. The agreed need for technologies from the database domain is therein often referred to as the "Web as a Database", a concept that is still more a vision than a reality. In this paper we have discussed the Technologies for Web 2.0 Applications. "Web of data" is also called as linked open data. We study the comparison between RDBMS and LOD. DATASET are the space provided online we have listed few of them. We have discussed the components of linked data and standards used in structured data. There is some challenge in linked data.

Keywords: Web of data, linked data, LOD, web as a database.

I. INTRODUCTION

In computing, linked data describes a method of publishing structured data so that it can be interlinked and become more useful. It builds upon standard Web technologies such as HTTP and URIs, but rather than using them to serve web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data from different sources to be connected and queried [5].

Tim Berners-Lee, director of the World Wide Web Consortium, coined the term in a design note discussing issues around the Semantic Web project. However, the idea is very old and is closely related to concepts including database network models, citations between scholarly articles, and controlled headings in library catalogs.

The World Wide Web is the largest collection of information created by mankind and bears potential for unthought-of applications. As of today, its potential still cannot be fully explored because current Web technologies mostly rely on manual efforts for integrating the huge amount of heterogeneous, loosely interrelated, and highly dynamic Web data. To address this problem, the Semantic Web community proposes a transition from the Web of documents to the "Web of Data", combining raw data with data describing its semantics. This enables an integrated view on the information of the World Wide Web for machines and humans alike. The required annotation and linkage approach is supported by the linked data initiative, a core building block of the Semantic Web. While this movement recently gained a wealth of attention and clearly marked its success, the often demanded idea of the Web as the largest existing "heterogeneous distributed database" is far from reality, mainly due to the lack of a wide range of technology and functionality established by the database community [4].

Meanwhile, the notion of data spaces was coined by Franklin, Halevy, and Maier as a new research agenda for the database community. This notion argues for the need of a different view on data management problems in the context of enterprise data from a range of heterogeneous and loosely connected sources. [4] They propose the idea of a data space support platform (DSSP) as a key element of a data space, hiding all the data management complexity behind a range of services for search, monitoring, integration, etc.

Up to now, much of the work from both directions has been carried out independently from each other, though following very similar objectives and applying similar approaches. Recently, Heath and Bizer compared the Web of Data to a global data space. The Linked Data guidelines indeed provide abstractions and technologies to publish, access and process linked data . Thanks to the Linked Data initiative, data from different domains is already accessible in (to some extent) integrated data spaces, such as government data and the different domains covered by the Linking Open Data cloud. Moreover, links between these different data spaces already exist and new links are continuously added. However, misses to address how a support platform for the Web of Data can be accomplished.

II. TECHNOLOGIES FOR WEB 2.0 APPLICATIONS

Web 2.0 is not just about new technologies; rather it is more related to how technology can be used. However there are few technologies which are associated with Web 2.0. The proper use of such technologies assists in developing Web 2.0 applications which conform to standards and fulfills its aims and objectives[1].

A. Web Ontology Language

Web Ontology Language (OWL) is the vocabulary extension of RDF and RDF Schema. It is used to add and represent the meaning of terms in vocabularies to describe classes and properties and relation with other classes in such a way

DOI10.17148/IARJSET.2017.41005

that information can be processed by applications. OWL documents conform to XML structure and rules, therefore OWL documents can be exchanged with other systems and applications. Unlike web pages, OWL documents are intended to be read by computers rather than humans. It is now widely used for information sharing for Semantic Web contents. OWL is much stronger than RDF in terms of syntax rules and structure, thus making it more machine interportable [1].

B. RDF and RDF Schema

OWL and RDF are the approved two key Semantic Web technologies. Primarily RDF was designed to represent information regarding sources on the Web. However its purpose is now evolved to represent and link data with physical entities. RDF Schema provides basic elements to define ontologies, also referred as RDF Vocabularies. Both RDF and RDF Schema documents must be valid XML documents. One can define a new RDF vocabulary by defining new RDFS and using existing RDFS such as Dublin Core Metadata Initiative [1]

C. SPARQL

SPARQL is commonly known as RDF Query Language. It is developed to work with any RDF source that can be mapped to RDF. SPARQL consists of two main components, the query language and data access protocol. SPARQL has capabilities to perform queries by triple patterns, conjunctions, disjunctions, and optional patterns [1]

D. Web Feeds (Atom and RSS)

Web Feeds are generally used to distribute and notify users regarding changes and addition of contents at some website. In future, web feeds will eventually replace most of the distribution of email notifications. They are based on the concept of users' pulling the resources and information they are interested in rather than data being pushed to the users. Most web feeds are provided by using Really Simple Syndication (RSS), Atom Syndication Format or both. Unlike RSS, Atom is proposed IEFT Standard and is defined with XML namespace [1].

III. LINKED OPEN DATA

A significant number of large-scale datasets (for example, http://dbpedia.org, http://www.bbc.co.
uk/music, http://linkedgeodata.org, http://data.nytimes.com) have been published, adhering to the Linked Data principles [2]:

A. All items in a dataset should be identified using URIs

B. All URIs should be dereference able: using HTTP URIs allows looking up an item identified through an URI.

C. When looking up an URI, it leads to more data(typically represented in RDF)

D. Links to URIs in other datasets should be included in order to enable the discovery of more data.

The Linking Open Data community project was kicked-off in 2007 and has at time of writing produced over 100 data sets, providing over 6.7 billion RDF triples, interlinked by approximately 160 million RDF links[2].
In contrast to the full-fledged Semantic Web vision, Linked Data is mainly about publishing structured data in RDF using HTTP URIs, hence lowering the entry barrier for data providers and data consumers. As much as the current Web (of documents) is mainly targeting human users, a particular strength of Linked Data is that applications can use it straightforward. Essentially, one can understand the LOD cloud as a single data-space, or simply put, as a single,
Web-scale database. Given that an increasing number of applications uses LOD, one has to wonder what it takes to migrate (parts of) the relational database to the LOD setup. We hence argue that a database perspective for LOD is required to ensure its acceptance and realise a low-barrier adoption process [2].
Although research into mapping relational data to RDF and distributed query processing over the Web of Data is known, a systematic analysis of LOD from a database perspective is, to the best of our knowledge, not available. We study requirements and challenges concerning "LOD as a database"; our main focus are: (i) to study comparison of relational databases and LOD from an application perspective, (ii) to highlight challenges in understanding LOD as a database. Eventually, we want to motivate both the database research community and the Linked Data research community to mutually benefit from the issues we raise here and potentially use them as a starting point for further, joint research and development.

IV. COMPARISION

Current RDBMS can be characterized by a set of features they support. Table I lists these features, based on the Codd's rules. In the following, we discuss the applicability of the relational rules to Linked Open Data. Where applicable, we

DOI10.17148/IARJSET.2017.41005

discuss how they manifest in LOD; note also that we not discuss the rows Security, Transactions, Synchronizations and Safety from Table I, as they do not apply to the read-only case of LOD)[2].

A. *Integration:*

Integration of data is a major requirement for Linked Data. From the application perspective, one expects to query physically distributed data comparable to what RDBMS provide locally. However, with Linked Data, typically an additional step—the entity consolidation has to be performed in order to provide a consistent view on the data [2].

B. *Operations and Language:*

Operations are the essence of structured queries and thus a crucial requirement for Linked Data as well. The SPARQL Query Language for RDF, a W3C standard, corresponds to the read-only part of SQL’s Data Manipulation Language. Ongoing standardization efforts (http://www.w3.org/TR/sparql11-update/) focus on the extension of SPARQL with update capabilities [2].

C. *Catalogue:*

Based on the RDF model and semantics all data inherently follows a common schema, such as the notion of objects and data types. In order to represent domain-specific schema information, one uses RDF vocabulary languages such as RDF-S or OWL. However, concerning schema consolidation, additional efforts are necessary in the LOD case, which will be discussed below [2].

D. *User Views:*

SPARQL, the de facto standard query language for Linked Data, does not define views explicitly as they are known from RDBMS. With the SPARQL CONSTRUCT operation a basic support for views is available. However, what is currently missing are definitions about the up-to-date character of the so generated RDF graphs. This is mandatory for the support of views as known from the RDBMS world [2].

E. *Integrity:*

Enabling a database-like view on Linked Data requires actions for enabling integrity as a fundamental prerequisite. In RDBMS, the main building blocks to enable integrity guarantees by the system are primary keys, foreign keys, integrity constraints and normal-forms of relations. These building blocks have to be defined independently from the application and have to be stored in the database catalogue. Primary and foreign keys refer to the problem of entity recognition on URI basis. Besides this, in the Linked Data scenario the problem of integrity is mitigated, as updates and deletions are not supported [2].

F. *Null Values:*

In contrast to relational databases, LOD operates under the Open World Assumption (in contrast to the Closed World Assumption, which essentially says that a statement that is not known to be true is explicitly assumed to be false). It typically requires schema knowledge to decide how null values from RDBMS are exposed in RDF [2].

G. *Physical/Logical Data Independence:*

Logical data independence is a non-issue in LOD, as the third Linked Data principle ensures that the underlying data model in any case is RDF. Physical data independence is guaranteed through the fact that processing of LOD data is independent of the used representation (such as RDF/XML, RDFa, Turtle, etc.)[2].

H. *Distribution Independence:*

This is a very crucial principle in the context of Linked Data, as it is inherently distributed.

V. DATASETS

A. *CKAN:*

REGISTRY OF OPEN DATA AND CONTENT PACKAGES PROVIDED BY THE OPEN KNOWLEDGE FOUNDATION

B. *DBPEDIA :*

A DATASET CONTAINING EXTRACTED DATA FROM WIKIPEDIA; IT CONTAINS ABOUT 3.4 MILLION CONCEPTS DESCRIBED BY 1 BILLION TRIPLES, INCLUDING ABSTRACTS IN 11 DIFFERENT LANGUAGES

C. *GEONAMES :*

PROVIDES RDF DESCRIPTIONS OF MORE THAN 7,500,000 GEOGRAPHICAL FEATURES WORLDWIDE.

D. *UMBEL:*

A LIGHTWEIGHT REFERENCE STRUCTURE OF 20,000 SUBJECT CONCEPT CLASSES AND THEIR RELATIONSHIPS DERIVED FROM OPENCYC, WHICH CAN ACT AS BINDING CLASSES TO EXTERNAL DATA; ALSO HAS LINKS TO 1.5 MILLION NAMED ENTITIES FROM DBPEDIA AND YAGO

E. *FOAF :*

A DATASET DESCRIBING PERSONS, THEIR PROPERTIES AND RELATIONSHIPS [5]

The unprecedented availability of data promised by linked data on the Web represents a major paradigm shift over the existing Web's structure. By building on Web infrastructure (URIs and HTTP), Semantic Web standards (such as the Resource Description Framework and RDF Schema [RDFS]), and vocabularies, linked data can effectively reduce barriers to data publication, consumption, and reuse, adding a rich layer of fine-grained, structured data to the Web. At its core, linked data exposes previously siloed databases as data graphs, which can be interlinked and integrated with other datasets, creating a global-scale interlinked data space.

However, linked data poses challenges inherent to querying highly heterogeneous and distributed data. To query linked data on the Web today, users must first be aware of which exposed datasets potentially contain the data they want and what data model describes these datasets, before using
this information to create structured queries. This query paradigm is deeply attached to the traditional perspective of structured queries over databases and doesn't suit the linked data Web's heterogeneity, distributiveness, or scale. It's impractical to expect Web data consumers to have a previous understanding of available linked datasets' structure and location. Letting users expressively query relationships in the data while abstracting them from the underlying data model is a fundamental problem for Web-scale data consumption, which, if not addressed, will ultimately limit linked data's utility for consumers.

In addition to data model awareness, users querying linked data must master the syntax of structured query languages such as SPARQL. Most Web users aren't comfortable with structured queries, thus creating a usability barrier for the linked data Web. From a user perspective, natural language queries emerge as a simple and intuitive alternative. Previous investigations have empirically confirmed natural language's suitability for search and query tasks.

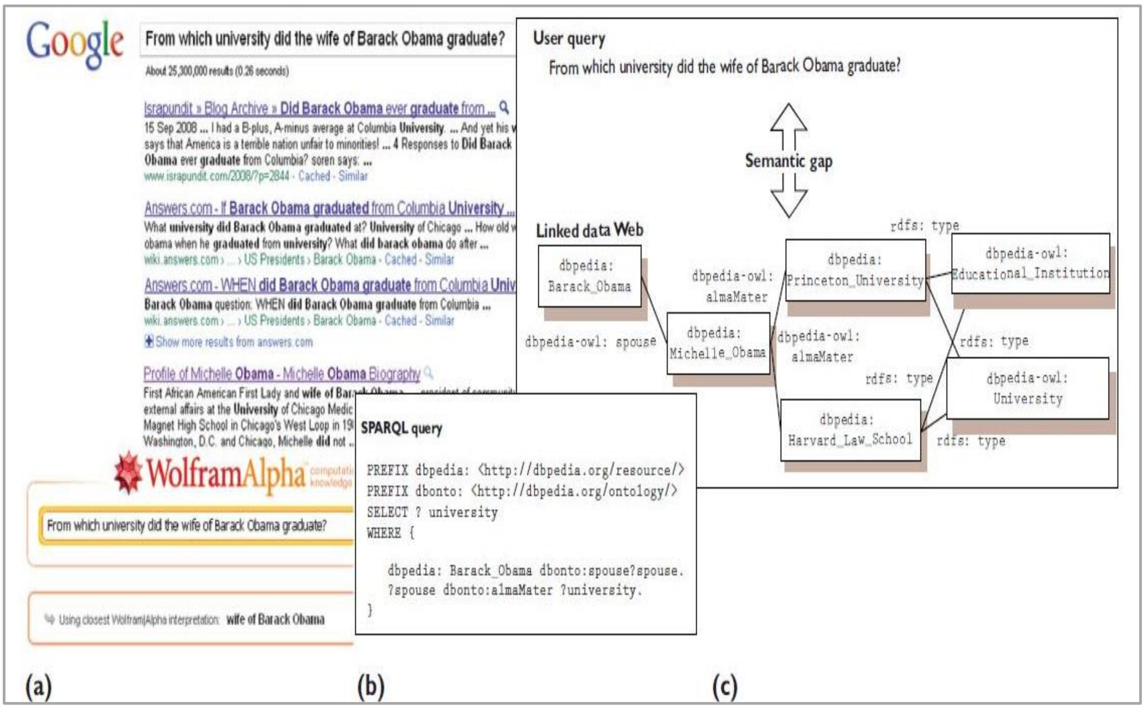
VI. LINKED DATA SPACE ENVIRONMENT

Linked data provides a data layer on the Web that represents objects and relations. The availability of Web-scale information in a structured and fine-grained representation could generate a paradigmatic shift in how applications and users consume data. Consider a journalist compiling a list of facts regarding public personalities and those personalities' previous academic
Affiliations. The journalist can express his or her information needs as natural language queries, such as "From which university did the wife of Barack Obama graduate?" Document search engines can't currently provide a level of query interpretation that could point directly to the final answer. With a traditional search engine, the journalist must navigate through the links and read the content of each candidate page the search engine returns. Modern search engines such as Wolfram Alpha, which relies on manually curated structured knowledge sources don't provide a sufficiently comprehensive solution to answer this query (see Figure 1a)[3].
The information that can answer this query is already available on the Web as linked data. However, to access it, users must know datasets location and structure, and the syntax of the SPARQL query language (see Figure 1b). Figure 1c shows the semantic gap between the user's
Information needs expressed in a generic natural language query and the data representation in the target dataset. The query's terms and structure differ from the data representation in the dataset.
The linked data Web already contains valuable data in diverse areas, such as e-government, e-commerce, and the biosciences. Additionally, the number of available datasets has grown solidly since its inception. The provision of

DOI10.17148/IARJSET.2017.41005

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

intuitive and flexible query mechanisms that can approximate users from an unconstrained amount of data represents a fundamental challenge, which, if not addressed, could affect the linked data Web's growth and adoption [3].



From which university did the wife of Barack Obama graduate?

About 25,300,000 results (0.26 seconds)

Israpundit » Blog Archive » Did Barack Obama ever graduate from ...
15 Sep 2008 ... I had a B-plus, A-minus average at Columbia University. ... And yet his w says that America is a terrible nation unfair to minorities! ... 4 Responses to Did Barack Obama ever graduate from Columbia? soren says: ...
www.israpundit.com/2008/?p=2844 - Cached - Similar

Answers.com - If Barack Obama graduated from Columbia University ...
What university did Barack Obama graduated at? University of Chicago ... How old w obama when he graduated from university? What did barack obama do after ...
wiki.answers.com › ... › US Presidents › Barack Obama - Cached - Similar

Answers.com - WHEN did Barack Obama graduate from Columbia Univ
Barack Obama question: WHEN did Barack Obama graduate from Columbia ...
wiki.answers.com › ... › US Presidents › Barack Obama - Cached - Similar

Show more results from answers.com

Profile of Michelle Obama - Michelle Obama Biography
First African American First Lady and wife of Bar... external affairs at the University of Chicago Medic... Magnet High School in Chicago's West Loop in 19... Washington, D.C. and Chicago, Michelle did not ...

From which university did the wife of Barack Obama graduate?

Using closest Wolfram|Alpha interpretation: wife of Barack Obama

User query
From which university did the wife of Barack Obama graduate?

SPARQL query

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbonto: <http://dbpedia.org/ontology/>
SELECT ? university
WHERE {

   dbpedia: Barack_Obama dbonto:spouse?spouse.
   ?spouse dbonto:almaMater ?university.
}
```

Fig. 1. Querying data over the web

VII. COMPONENTS OF LINKED DATA AND DATASPACES

The Linked Data initiative aims at enabling the Web of Data a single global network of human and machine readable information, based on the core Linked Data principles. Linked Data standards and technologies resemble some of the services proposed for a DSSP. In fact, the Web of Data can be seen as a collection of interlinked Web data spaces, in which support platforms can be achieved via combining Linked Data and database technologies. To illustrate this, Figure 2 gives an example of such a Web data space and the components of a support platform, aligned to the original data space concept in. The figure shows some of the originally proposed data space features, such as interlinked heterogeneous data sources as data space participants, surrounded by a set of services that hide the complexity of data integration and management from the data consumers and publishers. In contrast, data integration and access is driven and eased by standardizations (W3C), there is a lack of central control, and available knowledge is usually incomplete with respect to data dynamics[4].

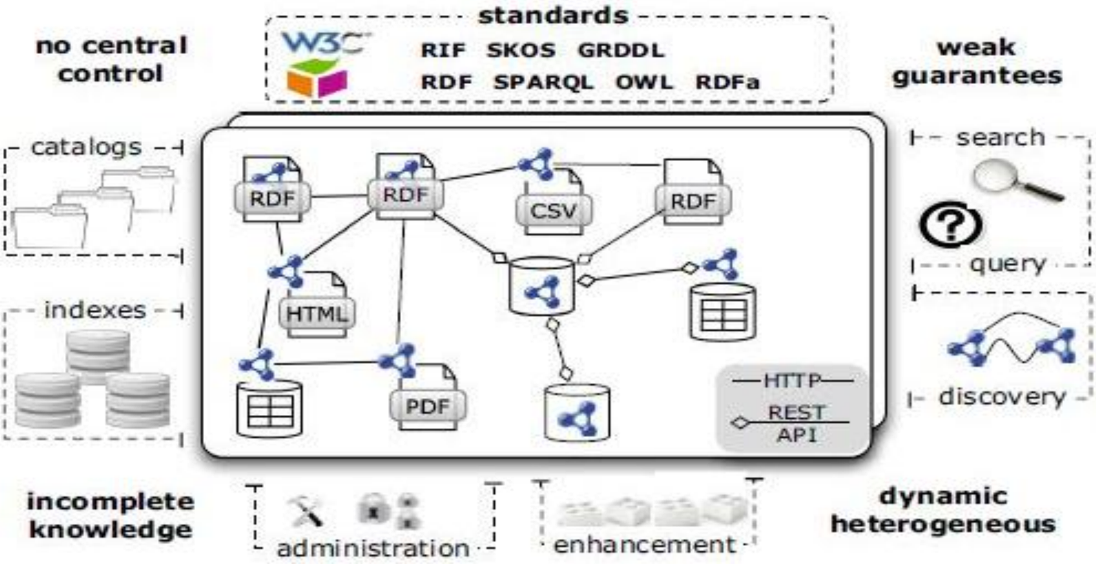


Fig. 2. Web dataspace environment

DOI10.17148/IARJSET.2017.41005

We base our argument that Linked Data partially implements some of the concepts of support platforms for Web data spaces on the following observations:

A. At the core of dataspaces are participants (data sources with different formats and different processing capabilities) and relationships. Linked Data is built around the same concepts, although they are called resources (documents, database endpoints, Web services, etc.) and links[4].

B. While proposes the usage of XML for interchanging data, Linked Data relies on RDF for this purpose, which offers even more advantages. RDF is not tied to a specific schema; it is self-describing, and allows a mix of structured and semi-structured data. It acts as the least common denominator between different data models, formats, and views, i.e., supporting the DSSP idea of being able to "query everything". This is further supported by a wide range of tools and wrappers for transforming and accessing various common data models into RDF (Figure 2)[4]

C. Describing the relationships between participants as explicit RDF triples serves the integration aspect of dataspaces by materialising "hard-wired" join structures, but also enables on-the-fly integration, e.g., by reasoning techniques that rely on query rewriting[4] .

D. As required for a support platform, a standardised access method (based on HTTP) and a common query language (SPARQL) exist for Linked Data. As for the actual data model, mappings between alternative query languages and SPARQL (e.g., W3Cs RDB2RDF) exist.[4]

E. URIs serve as global keys and support the integration of different sources and the identification of inconsistencies.

F. The discovery component of a support platform offers services to locate participants and data and to create and refine relationships among them. For Linked Data, resources can be discovered by Web crawlers and sources can acknowledge their existence and updates with so called "ping services", e.g. as used by major search engines. Link traversal and reasoning support the automatic identification and inference of links (i.e. relations).

G. Integration of different Web dataspaces is further encouraged and eased by techniques and tools for entity (a realworld object identified by a URI) recognition and by the provision of (optional) vocabularies and ontologies.

H. As required for dataspaces Linked Data can be streamed or stored, enabling, for instance, the integration of sensor data [4].

VIII. OPEN CHALLENGES

This list above illustrates that the tools, techniques, guidelines and principles of Linked Data represent a major step towards enabling support platforms for Web data spaces. However, while the Linked Data community successfully approaches the scale and openness of the Web, several, particularly database oriented, services of support platforms are still missing. In this section, we give a short overview of the main open challenges in this context and why the underlying concepts are essential to make the Web of Data a real success [4].

A. Graph-Based Data Model: RDF marks its impact for solving the problems of data integration and exchange of information, a fundamental requirement for overcoming disconnected data silos. The RDF data model is a form of a highly-normalized relational model with binary relations between global unique identifiers. However, without a deeper understanding of its theoretical foundations we will not be able to exploit the full potential of the Web of Data. For instance, without the identification of reoccurring structures and motifs, as well as the support of efficient graph traversals, indexing and query processing approaches will hardly achieve the scalability and efficiency required for the scale of the Web. Several research fields are approaching these issues. The logics and reasoning communities focus on the theoretical aspects, several recent data-mining efforts focus on graph analysis, and novel graph databases provide a good starting point for efficient graph management and processing. These directions, traditionally related to data management, have to be enriched by other research directions, such as behavioral analysis to help coping with the openness of the Web.

B. Search and Query: A fundamental requirement that has to be adopted from dataspaces is the urgent need for combining structured queries with unstructured (e.g., keywordbased) search. Without search functionality users will not be able to efficiently explore and discover the knowledge available. Beyond that, only structured querying will allow users to perform complex data analytics. By today, none of the established search engines for Linked Data offers efficient structured and meta-data queries, while SPARQL endpoints (HTTP interfaces for remotely posting SPARQL queries to RDF repositories) often lack in the support of efficient keyword-based search. Luckily, first approaches to support combined search and query are being discussed, designed and expectedly advanced and further developed. Apart from that, in a data collection as huge as the Web, users will usually be overwhelmed by the vast amount of somewhat relevant results. Modern Web technologies show that ranking is mandatory in such a situation. Some of the most prominent ranking algorithms for the Web have been adapted for

DOI10.17148/IARJSET.2017.41005

the Web of Data, but they only partially explore the richness of relationships embedded in RDF links. Further, there is the need for supporting ranking at different levels (domains, resources, etc) and for incorporating trust and personalization. All these requirements are already researched by the information retrieval, Web, and database communities, but to an extent that has to be leveraged. This also holds for the crucial requirement of supporting additional "meta" information with query results. For instance, provenance (or lineage) and trustworthiness are essential for assessing quality of data and query results, a feature that is mandatory for being able to handle amounts of data as large and diverse as produced in the Web. Similarly, approaches for identifying inconsistencies, object consolidation and entity resolution represent a good starting point, but have to be advanced regarding the scalability of actually fixing encountered problems. These areas can strongly benefit from fundamental database research.

C. Guarantees: Guarantees are essential for dataspaces, for enterprise data and Web data alike. If there is no assurance that the recent data is received, that updates do not get lost, that results are complete (with respect to the currently available data), no meaningful interchange and interaction between the participants in the Web will ever get into place. However, considering the character of the Web, we cannot achieve full guarantees but have to aim for loosened features like eventual consistency. Issues like access control & policies and the data sovereignty of the publishers are already intensely discussed, but a full assessment of An available guarantee is mostly missing [4].

IX. CONCLUSION

This paper is a comparative study of RDBMS and LOD. We argue that the combination of Linked Data and database technologies qualify to establish missing services and to overcome the challenges we identified. Linked data as an individual can be used as a data base but with the proper use of technology.

REFERENCES

[1] Andre Freitas, Edward Curry, Joao Gabriel Oliveira, and Sean O'Riain. Querying heterogeneous datasets onthe linked data web: Challenges, approaches, and trends. IEEE Internet Computing, 16(1):24{33, January 2012.

[2] Michael Hausenblas and Marcel Karnstedt. Understanding linked open data as a web-scale database. In Proceedings of the 2010 Second International Conference on Advances in Databases, Knowledge, and Data Applications, DBKDA '10, pages 56{61, Washington, DC, USA, 2010. IEEE Computer Society.

[3] Omer Mahmood. Developing web 2.0 applications for semantic web of trust. In Proceedings of the Inter-national Conference on Information Technology, ITNG '07, pages 819{824, Washington, DC, USA, 2007. IEEE Computer Society.

[4] Jurgen Umbrich, Marcel Karnstedt, Josiane Xavier Parreira, Axel Polleres, and Manfred Hauswirth. Linked data and live querying for enabling support platforms for web dataspaces. In Proceedings of the 2012 IEEE 28th International Conference on Data Engineering Workshops, ICDEW '12, pages 23{28, Washington, DC, USA, 2012. IEEE Computer Society.

DOI10.17148/IARJSET.2017.41005