The Semantic Web as a Platform for Collective Intelligence

Leandro Mendoza[1], Guido Zuccarelli[1],
Alicia Díaz[1], and Alejandro Fernández[1,2]

[1] LIFIA, Facultad de Informática, Universidad Nacional de La Plata, Argentina
[2] CIC, Comisión de Investigaciones Científicas, Argentina

Abstract. The *Semantic Web* constitutes a promising platform for the development of computer support for cooperative work. However, the maturity of the related technologies and available datasets poses new challenges. Knowing what these challenges are, and assessing their impact in advance can save effort and reduce the chance of failure. In this article we discuss the specific challenges in the development of an application that integrates collaborative product reviews available in the *Semantic Web*. The challenges we identify, if not tackled, translate to an additional effort in the integration process, the need to discard available data, and potential inconsistencies and lack of data-quality in the final product.

1 Introduction

The World Wide Web is currently an ecosystem where users contribute and consume content. Part of this content serves as input for collaborative decision making. Such is the case of collaborative reviewing sites for movies, books, and other products, where users share and discuss their opinions. Our ability to build systems that empower users' to exploit this socially created content is limited by our capacity to find and interpret the users' opinions. If users provide their opinions in natural language (i.e., plain English) our systems need to apply NPL techniques. If users publish their opinions in different web-sites our systems must retrieve, interpret and integrate these opinions. The *Semantic Web* [13] proposes methods and technologies to transform the current web in a *Web of (Linked) Data* that programs can more easily interpret and act upon.

Berners-Lee et al. [1] defined the *Semantic Web* as an extension of the current Web in which information is given well-defined meaning, better enabling computers and people to work in cooperation. A more programming-oriented definition given by the author Yu in [13] conceives the *Semantic Web* as a collection of technologies and standards that allows machines to understand the meaning of information on the Web. Two closely related concepts to the *Semantic Web* vision are *Linked Data* and *Web of Data*: while *Linked Data* refers to a set of best practices for publishing and connecting structured data on the Web [2], the term *Web of Data* can be viewed as the result of applying *Semantic Web* technologies

N. Baloian et al. (Eds.): CRIWG 2014, LNCS 8658, pp. 255–262, 2014.

to make *Linked Data* possible. The foundation of the *Semantic Web* is RDF[1], a set of standards published by the W3C that define a data model consisting of resources, properties and statements (triples that connect resources through properties), and the means to publish and access them.

We believe the *Semantic Web* constitutes a promising platform for the development of computer support for cooperative work. In this article we report on the challenges that the *Web of Data* poses on the development of CSCW applications that retrieve, integrate and interpret users contributions available on-line. To illustrate the discussion we introduce "*Collective opinions*", an application that aggregates *Reviews* and *Ratings* on the *Web of Data*. It is based on the architecture proposed by the LDIF project [12]. LDIF is the "Linked Data Integration Framework", a mature initiative in the *Semantic Web* community.

Following, we position our work in the context of existing research at the intersection of *Semantic Web* and CSCW. Then, we present the requirements for the "*Collective opinions*" application and the principles in its design. The combination of requirements, design approach, and nature of available data on the *Semantic Web* result in a set of challenges that we discuss in section 4. Finally, we summarise our findings, and provide an outlook.

2 Related Work

Within the context of CSCW, the most commonly explored contribution of the *Semantic Web* focuses on its power to model and store knowledge through the use of ontologies. Santos et. al. [11] show how *Semantic Web* technologies add quality to crowdsourced, geo-spacial annotations on maps. Ontologies and ontology modelling languages such as OWL[2] provide the formal semantics that allow for the verification of data consistency, increase interoperability, and enhanced information retrieval. Most importantly, ontologies represent a common, well defined language for users to contribute annotations.

In [7], Tom Gruber argues that the current web (the web 2.0, the *Social Web*) provides "collected intelligence" instead of "collective intelligence". That is, the value of the current web is that it collects the contributions of users and aggregates them into community- or domain- specific sites such as Flicker or Youtube. However, to attain real collective intelligence new levels of understanding on this content should emerge. He presents *RealTravel.com*, an example of a collective knowledge system for the domain of travel, based on the *Semantic Web* principles.

Di Noia and Mirizzi [4] argue that, although the web of data provides tons of data, only few applications exploit this potential. They implemented a content based, movie recommender system that leverages the data available in the *Semantic Web*. They focus on three popular dataset; DBPEDIA [10]; Freebase [3], and LinkedMDB [8]. They construct a content based recommender algorithm

[1] `http://www.w3.org/standards/techs/rdf#w3c_all` - Last accessed on May 1st, 2014.

[2] `http://www.w3.org/TR/owl-features/` - Last accessed on May 1st, 2014.

that performs well in terms of precision and recall on the dataset of the Lenskit project [5]. Their work experimentally shows that these three datasets are mature enough and rich in high quality data to serve as the main data source for the proposed recommender system. Moreover they conclude that combining information from various datasets improves recommendations and does not add noise.

Summarising, the work of Santos and colleagues shows how *Semantic Web* technologies support the collaborative construction of data models within the boundaries of a single system. Gruber illustrates how the *Semantic Web* supports the emergence of new knowledge from the contributions of users in a collaborative system. Dinoia and Mirizzi, demonstrate that well curated semantic datasets can be combined to build effective recommender systems to support decision making. Our goal is to understand what is involved in implementing collective intelligence systems that can cope with the open, distributed, variable in quality, large scale nature of the Web of Data.

3 Collective Opinions - The Case Study

Our goal is to illustrate, with a concrete example, the challenges that currently face those that attempt to build groupware applications that exploit the potential of the *Semantic Web* as a repository of "collectively" constructed knowledge.

"Collective opinion" is a system that crawls the (semantic) Web to harvest what users say about books, movies and products to construct a "collective opinion". It processes textual reviews and numeric ratings, focusing on the requirements that pose the most interesting and diverse problems.

- **R1: Trending opinions** - rank the items that users are talking about the most these days. Identify reviews that were published recently, and aggregate them by the item they refer to. That a review was discovered recently is not enough to infer that it is a recent review. It might be the case of a dataset that was recently published to the *Semantic Web* with reviews from the past year.
- **R2: Ranking surprises** - list the items whose aggregated rating plummeted or skyrocketed in the last days. Correctly calculate the rating taking into account that: a) reviews could come in different scales, b) there are individual reviews and aggregated reviews.
- **R3: Associations** - provide associations in the form "users who liked this, also liked ...". A user might express an opinion on several sites; match reviews and ratings to users, and identifying the same user on various sites.

We adopt the architecture proposed by the LDIF project [12]. It models a *Semantic Web* application as consisting of a sequence of phases:

- **Access and retrieve data:** Linked data is published in various forms. RDF documents (using, for example, RDF/XML serialisation[3]) publish a

[3] `http://www.w3.org/TR/REC-rdf-syntax/` - Last accessed on May 1st, 2014.

collection of RDF triples that systems retrieve through http requests – complex data models are split into various RDF documents. Sparql end-points provide SQL-like query access to (normally large) RDF datasets. It is also possible to embed RDF statements within HTML documents using RDFa[4], Microformats[5] and Microdata[6].

- **Translate to a common vocabulary:** Various models, called schemas or vocabularies, can represent the same data in terms of resources and properties (much like various entity-relation models can represent the same data in the relational database world). They differ, for example, in the level of detail they provide. Vocabularies emerge and evolve independently which means that, at a given movement, several vocabularies for the same domain might coexist. To exploit this data we first need to translate it to one common vocabulary.
- **Resolve identities:** In the *Semantic Web anyone can say anything about anything.* Statements about a resource can be distributed in multiple datasets, in multiple locations. Moreover, there is no central authority to ensure the existence of a unique identifier for each resource. Applications must realise when two statements refer to the same resource, and act accordingly.
- **Fuse data and assure quality:** Once data is represented in a common vocabulary, and identities are resolved, only statements that comply with predefined quality criteria get fused into an integrated dataset.
- **Exploit data:** The final application (in our case, "Collective opinions") works on the resulting dataset to exploit the available, integrated, curated data.

4 Collective Opinion - The Challenges

Building *"Collective opinions"* confronted us with challenges inherent to the nature and maturity (or lack thereof) of the *Semantic Web* and the LDIF architecture, and challenges specific to the datasets available in the domain of our case study. Next, we report on those we faced when selecting the input and common vocabularies. Then we discuss the challenges in finding useful data. Finally, we discuss the challenges for data retrieval and fusion.

4.1 Challenges for the Selection of Vocabularies

A prerequisite to build an application that uses the LDIF framework is to select the vocabularies accepted during retrieval (input vocabularies), and the common vocabulary to use for the fused dataset. We evaluated each vocabulary's popularity in existing datasets as well as the nature of its supporting community of users, to select only those that added more value. To decide on the common

[4] `http://www.w3.org/TR/xhtml-rdfa-primer/` - Last accessed on May 1st, 2014.
[5] `http://microformats.org` - Last accessed on May 1st, 2014.
[6] `http://www.w3.org/TR/microdata/` - Last accessed on May 1st, 2014.

vocabulary we analysed how well it modelled our domain (*coverage* [14]), and how good it could map data published in the remaining vocabularies (*mappability* [14]). This challenging selection process involved extensive review of scientific publications, technology web-sites, and technical, specialised discussion forums.

The are three main alternatives to publish data about reviews. The **Review Vocabulary**[7] (also known as Review Ontology) was one of the earliest vocabularies to publish reviews and ratings using RDF. It can be traced back to the work of Heath and Motta [9] in the Revyu.com system for collaborative rating an reviewing. The "Microformats community" puts forward its own vocabulary for marking up reviews. **hReview**[8], is a simple, open format, suitable for embedding reviews (of products, services, businesses, events, etc.) in HTML, XHTML, Atom, RSS, and arbitrary XML. **Schema.org**[9] is a *Semantic Web* initiative led by Google, Bing, Yahoo and Yandex to help authors embed semantics into HTML pages. It concentrates on simplicity and on a well understood set of abstractions (including Reviews) that these big search companies think can have special treatment in their search engines, for example showing rich snippets. Microdata is the recommended mechanism so publish Schema.org data within HMTL pages, although RDFa and Microformats are also applicable.

Through API queries to two widely used semantic search engines, *LOD Cloud cache (LODC)*[10] and *Sindice (SIND)*[11], we observed that, in their search indexes, the three vocabularies appeared frequently enough to justify including them as input. There were also traces of a predecessor of Schema.org called **data-vocabulary** that we decided to ignore as most sites should eventually upgrade.

We compared the three vocabularies to assess coverage of the data needed to implements the application requirements, and to establish alignments or mappings [6] between equivalent concepts (with similar meaning). The three vocabularies describe the person who creates a review (requirement R3), the date of creation (R1 and R2), a personal opinion in form of text, and a rating that corresponds to a numeric value within a given range (R1, R2, and R3). All of them also foresee a mechanism to associate a review with the resource being reviewed (R1, R2, and R3), the difference being that for the Review Vocabulary, this is achieve through a relationship from the resource to the review (i.e., backwards).

We conclude that Review Vocabulary, hReview, Schema.org are mappable to one-another. Moreover, Review Vocabulary is formally defined in RDFS[12]. It can be used to publish not only within HTML (with RDFa) but also in RDF documents, and in Sparql endpoints. There are already tools that map hReview to Review Vocabulary. Based on this observations, we choose **Review Vocabulary** as the base vocabulary to represent our integrated data.

[7] `http://purl.org/stuff/rev#` - Last accessed on May 1st, 2014.
[8] hReview `http://microformats.org/wiki/hreview` - Last accessed on May 1st, 2014.
[9] `http://schema.org/` - Last accessed on May 1st, 2014.
[10] `http://lod.openlinksw.com` - Last accessed on May 1st, 2014.
[11] `http://sindice.com/` - Last accessed on May 1st, 2014.
[12] `http://www.w3.org/TR/rdf-schema/` - Last accessed on May 1st, 2014.

4.2 Challenges for the Selection of Data Sources

Most semantic information about reviews and ratings is currently embedded in HTML documents. Semantic search engines are the standard mechanism to find these documents. Using semantic search engines is challenging in terms of *availability* [14] of these services and the *timeliness* [14] of their responses. The alternative (impracticable for most scenarios) is to implement an ad-hoc web crawler. Both, SIND and LODC, provide a SPARQL endpoint to query its RDF datasets. SIND provides a search API that we can use by calling it programmatically. During our experiments, SIND suffered frequent shutdowns which spanned weeks. LODC remained accessible for the whole duration of our study (two months). In order to assess the amount of available data, we performed a query to retrieve all those documents that contains data about Reviews, using the *Review Vocabulary* as the baseline. The results showed that LODC reports 5,014,468 documents using the *Review Vocabulary*, whereas SIND reports 10,216,632 documents. It is important to note that each document could contain information about more than one Review. To assess data *timeliness*, we searched for any document containing information about Reviews (using any of our input vocabularies). We took a random sample of size 1000 from the results obtained in each engine. We immediately downloaded those documents and inspect them. The percentage of documents from the result set that was still available on-line was 69% for SIND and 54% for LODC. Moreover, 72% of the documents in SIND's result set that were on-line, still had relevant semantic content; in comparison only 40% of those in LODC's result set did, which indicates that search engine's indexes are largely outdated.

4.3 Challenges for Data Retrieval and Fusion

Using URIs to identify resources is a key principle of the *Semantic Web*. Our input vocabularies foresee that the subject of the review and its author are resources. Our application depends on this principle to uniquely identify items and persons for all three requirements. However, we found multiple cases where a string (the name of the person) is used to specify `author` when the expected value for this property is a resource (i.e., a URI) typed as `Person` or `Organisation`. Our approach in these cases was to discard the data. The same problem was present when the item of the review was a string (e.g., the title of the movie) instead of a URI. If we knew the domain was restricted to books or movies, we could guess the identity of the item via comparison to labels of known books or movies in curated datasets such as DBPEDIA. However, this approach would require additional effort and is error prone and of limited applicability.

Our input vocabularies define that the rating value should be numerical and must be in the range defined by the *min* and *max* values. The use of non-numerical values for rating (or rating range) is a recurring problem in the available data. For example, a rating value described using a string such as "rating 1 of 5" instead of a numerical value and a valid range (rating value: 1, Min rating value:1, Max rating value: 5). Reviews that presented this problem were

discarded as they are of no use to implement our requirements. For other scenarios they might still be valuable.

In RDF, a properties can take a literal value (instead of a URI that identifies a resource). The name of a person and the numeric value of a rating are typed literals. Typed literal values consist of a string (the lexical form of the literal) and a datatype (identified by a URI). Knowing how dates are represented is critical for our application (specially R1 and R2). It is common practice to use XML schema data types; and the convention[13] to represent dates and times is to use the ISO 8601 Date and Time Formats. When the date was not available or did not follow the conventions, we considered for the total ratings (thus R3) but not for trends and surprises (R1, and R2).

Our input vocabularies have a mechanism to indicate the type of resource being reviewed (i.e., a movie, a book, a restaurant). In Schema.org and hReview the type of resource is a property of the review itself. In the Review Vocabulary, the type is a property of the resource that identifies the item. Being able to tell the type of the reviewed item lets us implement R3, suggesting only resources of the same type (i.e., users who liked this *book*, also liked these *books*). The data we obtained varied widely regarding this aspect therefore we had to resort a more open version of R3.

Search engines indicate that web sites might be more prominently displayed in search results if they provided semantic markup for their content. This situation motivated web-sites creators to indiscriminately copy and republish content from others sites (particularly movie reviews). There are currently no consistent mechanisms to identify and discard exact content replicas. If the date and time of the review, and its author are not available we cannot tell if two reviews are the same or not.

5 Conclusions

The *Semantic Web* can foster the creation of CSCW applications that exploit users' generated content. Existing work shows that, in controlled scenarios, these technologies support the emergence of new knowledge from the contributions of users in a collaborative system. In this work, we discuss some of the challenges of implementing collective intelligence systems that can cope with the open, distributed, variable in quality, large scale nature of the Web of Data. Focusing on the development of an application that integrates product reviews we learnt that: a) the distributed and collaborative nature of the *Semantic Web* originated a variety of alternative vocabularies (and supporting communities) to model the same domain, that developers must find, evaluate, select and combine, which demands considerable effort; b) semantic search engines, a common mechanism to identify data sources, lack stability and timeliness – therefore, alternative mechanism are called for; and c) many existing datasets lack quality and cannot

[13] XML Schema: `http://www.w3.org/TR/xmlschema-2/` - Last accessed on May 1st, 2014.

be effectively used in applications that aim at doing rich integration – strategies for the early identification of quality problems and for data curation are needed.

We continue studying the domain of user generated semantic content and the implications of using it for collective intelligence. Next steps are the compilation of a formal model for quality in *Linked Data* that can serve as the basis for automated evaluation of datases, and potentially, automated curation.

In this work we focused on CSCW applications that take the *Semantic Web* as a source of data. In an additional line of research, we explore the potential of RDF as a flexible modelling framework to enable Group Decision Support Systems.

References

1. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific American 284(5), 28–37 (2001)
2. Bizer, C., Heath, T., Berners-Lee, T.: Linked data-the story so far. International Journal on Semantic Web and Information Systems 5(3), 1–22 (2009)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data, pp. 1247–1250. ACM (2008)
4. Di Noia, T., Mirizzi, R., Ostuni, V.C., Romito, D., Zanker, M.: Linked open data to support content-based recommender systems. In: Proceedings of the 8th International Conference on Semantic Systems, I-SEMANTICS 2012, p. 1 (2012)
5. Ekstrand, M.D., Ludwig, M., Kolb, J., Riedl, J.T.: Lenskit: a modular recommender framework. In: Proceedings of the Fifth ACM Conference on Recommender Systems, pp. 349–350. ACM (2011)
6. Euzenat, J., Shvaiko, P.: Ontology matching, 2nd edn. Springer, Heidelberg (2013)
7. Gruber, T.: 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 (2008)
8. Hassanzadeh, O., Consens, M.P.: Linked movie data base. In: LDOW (2009)
9. Heath, T., Motta, E.: Revyu.com: a reviewing and rating site for the Web of Data. In: Aberer, K., et al. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 895–902. Springer, Heidelberg (2007)
10. Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., et al.: Dbpedia-a large-scale, multilingual knowledge base extracted from wikipedia. Semantic Web Journal (2013)
11. Gonzalez, A.L., Izidoro, D., Willrich, R., Santos, C.A.S.: OurMap: Representing Crowdsourced Annotations on Geospatial Coordinates as Linked Open Data. In: Antunes, P., Gerosa, M.A., Sylvester, A., Vassileva, J., de Vreede, G.-J. (eds.) CRIWG 2013. LNCS, vol. 8224, pp. 77–93. Springer, Heidelberg (2013)
12. Schultz, A., Matteini, A., Isele, R., Mendes, P.N., Bizer, C., Becker, C.: Ldif-a framework for large-scale linked data integration. In: 21st International World Wide Web Conference (WWW 2012), Developers Track, Lyon (2012)
13. Yu, L.: A developer's guide to the semantic Web. Springer (2011)
14. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment methodologies for linked open data. Submitted to SWJ (2012)