A rule-based approach to address semantic accuracy problems on Linked Data
(ISWC 2014 - Doctoral Consortium)

Leandro Mendoza[1]

LIFIA, Facultad de Informática, Universidad Nacional de La Plata, Argentina

1 Problem Statement

In 2001, Berners-Lee et al. [2] defined the *Semantic Web* (SW) as an extension of the current Web in which information is given well-defined meaning through the use of common standards and technologies to facilitate the sharing and reuse of data. In 2006, the related term *Linked Data* (LD) [3, 7] was proposed as a way to identify a set of best practices for publishing data using SW tools that allow to link these isolated datasets in a large network of distributed data [17]. Since then, the number of available datasets that follow the SW and LD ideas has been considerably increasing, leading to what is currently known as the *Web of Data* (WoD).

Although this WoD provides tons of information (see the LD cloud[1]), evidence shows that it is only as usable as its quality: there is a lot of noise in current SW datasets[2] and just a few applications can effectively exploit the inherent potential of this well-defined and structured information. These SW datasets covers different domains and have different levels of quality: from "high-quality" curated SW datasets (for example, in life-science domain) to those which were extracted from unstructured and semi-structured sources or were the result of a *crowdsourcing* process (for example, DBPedia [11]). Some of the data-quality problems that affects those datasets are out-of-date values, incomplete or incorrect data, inconsistencies, etc. Most of these problems arise during the creation process of SW data, due to errors in the original data source, the tools employed to convert or create SW data, misuse of ontologies, etc.

The main problem addressed in my PhD work is about to improve existing SW datasets (also new and emerging ones) that suffer from quality problems by taking advantage of information available in other SW datasets with presumed and relatively "high" quality. This work will be mainly focused in two related quality dimensions: *"semantic accuracy"* (values that do not correctly represent

[1] http://linkeddata.org/

[2] The term "Semantic Web (SD) dataset" used in this document is also referred in others works as "Linked Data (LD) set", "RDF dataset" or generally as "dataset in the WoD".

the real state of real-world objects) [6, 20] and *"interlinking"* (datasets that are not properly linked to another datasets) [9, 20]. The aim is to develop mechanisms to detect and evaluate these quality criteria and also make suggestions to **enrich** (complete or add relevant data) and **curate** (repair wrong or inconsistent data) SW datasets. To achieve this goal, existent SW datasets (that we call "*seeds*") will be used to derive "*dependency rules*" (DRs) (relationships between attributes of a schema or ontology) that then will be applied on other dataset (that we call "*target*") to detect, measure and fix quality problems. In order to clarify the ideas behind our approach, we propose a simple SW dataset as use case scenario:

SW dataset about books and its authors. For each book we have ISBN, keywords, publication date, language, topic, etc. For each author we have personal information like country and city of residence, work place, organization, etc. This will be our *"target"* dataset on which we want to improve quality.

According to the problem that we want to address on the *"target"* dataset, specific issues need to be tackled:

- *Identify "dependency rules" (DR) using "seeds" datasets.* For example, a DR could be "`country` and `city` names determine the `zip-code` value for the author's residence location. Another DR could be, "Author's `country` and `country-language` determines the `language` of author's books".

- *Detect inconsistencies, wrong values or incomplete data on the "target" dataset.* For example, if we have information about `country`, `city` and `zip-code` (and the corresponding DRs that relate them), we want to check if values for these attributes are consistent between them.

- *Make suggestions to improve data completeness of the* "target" *dataset.* For example, if the language of a book is not established, we want to derive this information from those attributes that provides information about author's residence `country` and `country-language` (first, we must detect the DRs that relate these attributes).

- *Make suggestions to improve interlinking between "target" and "seeds" datasets.* For example, if the `country` and `city` values are just string values like "Argentina" and "Buenos Aires", how can we suggest links to connect the *"target"* dataset with the *"seeds"* datasets that provides URIs for "Argentina" and "Buenos Aires" resources (for example, DBPedia).

2 Relevancy

As mentioned above, the *WoD* provides big amounts of information distributed over a large number of diverse datasets but the usefulness of this data depends on its quality. If I succeed, my PhD work will contribute in the SW data-quality research area and, more specifically, in the following related activities:

- ***Dataset enrichment and curation***. Enrichment refers to add relevant information to one dataset using data provided by other datasets. Curation refers to fix inconsistent or wrong data. Both activities are complementary.

- ***Link discovery and interlinking***. One of the key principles of LD is to relate datasets between them. Thus, once a set of potential external sources to relate with is detected (links discovery), the publisher must face with the decision of which one choose to link (interlinking). I expect to contribute in the Link Discovering [4] research area by developing methods to detect errors in links (incomplete, invalid, out-of-date, etc.) or suggest new links.

As a direct consequence of the potential contribution in the areas mentioned above, my PhD work will also contribute in the following activities:

- ***Data publishing***. Currently, there is a growing interest by organizations in publish data using SW and LD principles. One of the most important and complex aspects to consider during this task is to ensure data quality. It is therefore essential that publishers have mechanisms to detect quality problems and, eventually, have the tools to fix them.

- ***Development of applications and software over the WoD (Semantic Web applications)***. SW aplications developers will be hampered their task when trying to build intelligent software agents that automatically collect information of the WoD in order to get an integrated knowledge base for a certain purpose. Data quality is a critical aspect in an integration scenario where the readiness of information needs to ensure that it can be efficiently exploited by applications.

3 Related Work

As the amount and usage of SW data grew, several works have been addressed the data quality aspect of datasets. Zaveri et al. [20] present the results of a systematic review of approaches for assessing data quality of *LD* identifying a core set of twenty-six data quality dimensions (criteria). Vrandecic's work [16] focuses on ontology evaluation and provide a theoretical framework defining a set of eight ontology quality criteria and ontology aspects that can be evaluated as well as related methods and evaluations. Regarding data quality assessment methods (also known as framework or methodologies) for SW datasets, existent approaches can be classified into semi-automated, automated and manual [12, 19, 10, 1]. Besides, there is a lot of research performed extensively to assess the quality and report commonly occurring problems of the existing datasets [8, 9]. Regarding to *"semantic accuracy"* assessment, Fürber and Hepp [6] propose SWIQA, a quality framework that employs data quality rule templates to express quality requirements which are automatically used to identify deficient data and calculate quality scores for five quality dimensions. *"Semantic accuracy"* is one of these dimensions and authors proposed to identify semantically incorrect values through the manual definition of functional dependency rules. Another work that is inspired in the "functional dependency" concepts was done by Yu and Heflin [18]. In that work, authors propose a clustering-based approach to facilitate the detection of abnormalities in SW data by computing functional dependencies like, for example,"The language of a book is determined by the author's country". Fleischhacker et al. [5] give an approach oriented to enrich the schema of a SW dataset with property axioms (based on association rule mining) by means of statistical schema induction and also discuss other approaches related with the research areas of "LD mining" and "association rule learning" [14].

4 Research Questions

The research questions that I plan to address are:

- **What are the implications of learning "dependency rules" (DRs) from existent SW datasets?**

  To answer this question we need to understand the mechanisms to learn DRs from SW datasets and what kind of data do we need to perform this task (schemas, instance data, etc.). Besides, some related questions also need to be answered: Are these DRs dependent on both *"seeds"* and *"target"* datasets? Can these DRs be reused for apply in different datasets? How the amount-of-data of the involve datasets does affect the detection of DRs?.

- **How existent data quality assessment metrics can be used in my approach to measure "Semantic accuracy" and "Interlinking"?**

  To answer this question we need to understand the quality problems related to *"semantic accuracy"* and *"interlinking"*, examine its causes and consequences and study the existent methods to deal with them. In this sense, it is important to see the relation of these two dimensions and the potential of work with them together to improve quality. Finally, determine in which way DRs can be used to build procedures that allow us to detect a quality problem and measure certain information of the mentioned dimensions.

- **How to suggest recommendations to enrich and curate a SW dataset?**

  To answer this question we need to separate both activities. To enrich a dataset we need to know how to detect what information is missing or incomplete, to then suggest not only new relevant information but also the way it should be used (completing a property value, adding a link, etc.). To curate a dataset, we need to detect wrong or inconsistent attribute values and suggest a way to correct them (deleting, replacing, etc.) giving new consistent values. For both scenarios, it is necessary to understand how DRs can be used with instance data of *"seeds"* datasets in order to make suggestions of new relevant data for the *"target"* dataset.

– **What are the methodologies issues to be considered when assessing the quality of SW datasets?**

  To answer this question it is important to understand the limitations and drawbacks of current data quality assessment methodologies in order to determine how can we improve (or extend) them to fit with the needs of our approach.

5 Hypotheses

The main idea behind the approach of my PhD work is to improve the data-quality (regarding to *"semantic accuracy"*) of a SW datasets (that we will call *"target"* dataset) through a strategy that will use existing datasets (that we will call *"seeds"* datasets). Assuming a certain level of related "high-quality" for *"seeds"* datasets, we will use them to learn *"dependency rules"* (DRs). These DRs will be used to measure *"semantic accuracy"* (detecting wrong or inconsistent values), curate data (suggest new correct values) and enrich data the *target* dataset (complete missing values for attributes and suggest links to others datasets). This approach to improve data-quality leads to a cycle strategy: existent high-quality datasets can be used to improve quality of new and emerging datasets, and these in turn can also be used by future and even existent datasets with the same purpose. This general idea takes data-quality as a "transferable property": the quality of a SW dataset depends not only on the quality of their own data, but also on the quality of the external sources which are related to.

6 Preliminary results

Recently, we have been working on challenges related with the development of an application that integrates product reviews available as SW data (microformats, RDFa, rdf files, etc.) [13]. In this experimental work, we studied the architectures available to build SW applications and we focused on the data integration process. We also studied how quality problems affect the development of these applications when trying to consume and integrate data from heterogeneous SW datasets. We used a set of quality criteria which we divided in three categories: data-provider quality, schema quality and instance-data quality. Regarding data-provider quality we addressed "accessibility", "amount-of-data" and "timeliness". For schema quality we analyzed "coverage" and "mappings". Finally, for instance-data quality we analyzed "accuracy" (syntactic accuracy and semantic accuracy) and "completeness" (property completeness and interlinking completeness). We got SW data about reviews using Sindice[3] and LOD-Cache[4] search engines. After analyze the retrieved data, we described common occurring errors for each criteria and their effects in the integration process. We found that most reviews have quality problems mainly related to incomplete data (reviews's text, language, rating or even a reference to the reviewed item is missing) and inconsistent values (for example, the text property has the value "This books is great" and rating property has value "0"). Although we did not propose a solution to the problems found, we noticed that many of them could be detected or even curated using information available in other datasets like DBPedia.

7 Approach

As mentioned in section 1, my PhD work will intend to address the data quality aspect of SW datasets by considering two quality dimensions: *"semantic accuracy"* and *"interlinking"*. The main idea behind this approach is to use existent SW datasets as *"seeds"* to learn DRs. Then, apply these DRs over a *"target"* dataset to detect incomplete, erroneous or inconsistent data and finally, make suggestions to curate and enrich the "*target*" dataset using instance values of the *"seeds"* datasets. In order to facilitate the understanding of the main problem, it was divided into more specific sub-problems. The first and most important task is related with how to get *"dependency rules"* (DRs) from *"seeds"* datatsets. *"Dependency rules"* concept is inspired in "data dependency" concept (well-known in relational databases domain and already used in [18] to detect abnormal data in RDF Graphs). With the DRs obtained, we will work on:

- *Detection and measurement of "Semantic accuracy" and "Interlinking"*. Although both dimensions will be treated separately, the idea is to take as reference quality evaluations performed by related work (see section 3) and adapt them to our approach (using DRs, *"seeds"* and *"target"* dataset).

- *Suggest recommendations to "enrich" and "curate" data.* Although both activities will be treated separately, the idea is to use a "Content-based Recommender System" approach [15] that uses DRs and *"seeds"* datasets to suggest new relevant data, either to complete or replace erroneous and inconsistent values.

The novel contribution of this work lies in extending current quality assessment methodologies, using existent SW datasets to get DRs and apply them to other datasets in order to detect and fix quality problems to increase data quality levels.

[3] http://sindice.com/

[4] http://lod.openlinksw.com/

8 Evaluation Plan

To facilitate the evaluation of my PhD approach, I will divide the task in the same way as Section 7. The proposed solutions for each sub-problem will be evaluated using offline experiments performing on pre-collected datasets that must meet certain requirements. For *"seeds"* datasets, it is neccessary to ensure a minimum level of data-quality, at least, for those attributes that will be considered in the DR, and will be used to make recommendations (for enrich and curate data). Both types of datasets, *"seeds"* and *"target"* must have a controlled size (in terms of amount-of-data) according to the complexity of the algorithms and hardware limitations. Attributes of interest of the involved schemas (or ontologies) must be mappeable. I pretend to evaluate my approach by comparing how many correct and useful DRs have been detected and how they can be used in detection and recommendations tasks:

- A DR is correct if the involved attributes represents a consistent relation according to *"seeds"* and *"target"* dataset (instance data and schema). We must check manually if a DR is correct (for example, having a set of predefined DRs we can test if our approach generates similar DRs).

- A DR is useful (for detection) if it can be used to detect wrong values (test *"semantic accuracy"*) or missing values (incomplete properties).

- A DR is useful (for prediction) if it can be used by recommendation algorithms to provide new attributes values and suggest potential relevant links to other datasets.

Note that the evaluation plan should include the test of algorithms used to derive DRs, detect wrong and incomplete values and generate recommendations. Traditional "precision", "recall" and other related approaches [15] can be used in these tasks.

9 Reflections

My PhD approach is based on the fact that there is a huge amount of information published following SW and LD principles and also that quality problems affects these diverse datasets to a greater or lesser extent. I also understand that data quality in SW datasets is an emerging research area of great interest with applications in domains like e-science, e-government and even e-commerce. Although many works have addressed the SW data-quality problem, most of them proposes methodologies to evaluate specific quality-criteria and report common occurring errors on a particular dataset. Only a few mention mechanisms to deal with incomplete or inconsistent data. The development of mechanisms and scalable tools to effectively solve these problems is still an open challenge.

References

1. Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., Lehmann, J.: Crowdsourcing linked data quality assessment. In: The Semantic Web–ISWC 2013, pp. 260–276. Springer (2013)
2. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american 284(5), 28–37 (2001)
3. 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)
4. Ferraram, A., Nikolov, A., Scharffe, F.: Data linking for the semantic web. Semantic Web: Ontology and Knowledge Base Enabled Tools, Services, and Applications p. 169 (2013)
5. Fleischhacker, D., Völker, J., Stuckenschmidt, H.: Mining rdf data for property axioms. In: On the Move to Meaningful Internet Systems: OTM 2012, pp. 718–735. Springer (2012)
6. Fürber, C., Hepp, M.: Swiqa - a semantic web information quality assessment framework. In: Tuunainen, V.K., Rossi, M., Nandhakumar, J. (eds.) ECIS (2011)
7. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology 1(1), 1–136 (2011)
8. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web (2010)
9. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of linked data conformance. Web Semantics: Science, Services and Agents on the World Wide Web 14, 14–44 (2012)
10. Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.J.: Test-driven evaluation of linked data quality. In: Proceedings of the 23rd international conference on World Wide Web (2014), to appear
11. 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)
12. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 116–123. ACM (2012)
13. Mendoza, L., Zuccarelli, Díaz, A., Fernández, A.: The semantic web as a platform for collective intelligence (2014)
14. Nebot, V., Berlanga, R.: Mining association rules from semantic web data. In: Trends in Applied Intelligent Systems, vol. 6097, pp. 504–513. Springer Berlin Heidelberg (2010)
15. Shani, G., Gunawardana, A.: Evaluating recommendation systems. In: Recommender Systems Handbook, pp. 257–297. Springer (2011)
16. Vrandečić, D.: Ontology evaluation. Springer (2009)
17. Yu, L.: A developer's guide to the semantic Web. Springer (2011)
18. Yu, Y., Heflin, J.: Extending functional dependency to detect abnormal data in rdf graphs. In: The Semantic Web–ISWC 2011, pp. 794–809. Springer (2011)
19. Zaveri, A., Kontokostas, D., Sherif, M.A., Bühmann, L., Morsey, M., Auer, S., Lehmann, J.: User-driven quality evaluation of dbpedia. In: Proceedings of the 9th International Conference on Semantic Systems. pp. 97–104. ACM (2013)
20. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment methodologies for linked open data. Submitted to SWJ (2012)