Adequate class assignments on Linked Data

Leandro Mendoza*†, Alicia Díaz †
*CONICET, Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina
†LIFIA, Facultad de Informática, Universidad Nacional de La Plata, Argentina
{lmendoza, alicia.diaz}@lifia.info.unlp.edu.ar

***Abstract*—In recent years *Semantic Web* technologies and the *Linked Data* paradigm have allowed the emergence of large interlinked knowledge bases as *Linked datasets*. These databases contain information that associates *Web* entities (called *resources*) with a well-defined semantics that specifies how these entities should be interpreted. A way to perform this task is through a *class assignment* process where *resources* are identified as members of certain classes described in *ontologies*. In order to improve the quality of the "meaning" of the data contained in *Linked datasets* a key challenge in the *Linked Data* community is to detect, assess and eventually fix wrong *class assignments*. In this sense, this work proposes an interpretation for *adequate class assignments* considering three quality dimensions from a semantic perspective: *redundancy*, *consistency* and *accuracy*. For each dimension, a formal definition is presented, then applied to *class assignments* and finally used as guideline to show how quality metrics and data *curation* strategies can be defined.**

I. Introduction

In recent years, *Semantic Web* technologies and the *Linked Data* paradigm [1] have allowed the emergence of large interlinked knowledge bases (also known as *Linked datasets*) conforming what is known as the *Web of Data*. These datasets contain information about *Web* entities (called *resources*) identified by unique HTTP URIs to which a well-defined semantics has been associated. This semantics specifies how *resources* should be interpreted and allows automatic knowledge discovering through *inference* techniques. Adding semantics to *Web* entities can be seen as a classification process: given a set of *resources* and *classes* (concepts) usually described in ontologies, *class assignment* assertions are created to specify *resources* as member of certain *classes*. This *class assignment* process forms part of the *Linked Data* life cycle [2] which, from an overall perspective, includes phases to: process data from a variety of unstructured or semi-structured data sources, identify entities as resources, define (or reuse) classes to model these resources, generate *class assignment* assertions and publish these data according to the *Linked Data* principles[1]. In this sense, *class assignments* information is essential in most *Linked datasets* and its quality analysis is a current challenge in the *Linked Data* community. From a semantic perspective, quality assessment in *Linked datasets* involves the development of mechanisms to detect, measure and fix errors in the *"meaning"* of the data. In this way, three related dimensions can be found in the *Linked Data* literature: *semantic redundancy*, *semantic consistency* and *semantic accuracy*. Although these dimensions have been addressed separately in different works, there are no techniques that consider all of them together. In contrast, our approach interpret them as complementary concepts that can be used together to achieve semantically *adequate class assignment* assertions in *Linked datasets*. The following sections are organized as follows: in section II some background definitions about *Semantic Web* and *Linked Data* are given. In section III related work is described. In section IV we explain our approach by interpreting *redundancy*, *consistency* and *accuracy* for *class assignments* in *Linked datasets*. In section V some brief discussions related to the addressed concepts are introduced. Finally, conclusions and future work are given in section VI.

[1]https://www.w3.org/DesignIssues/LinkedData.html

II. Background

The *Linked Data* paradigm allows us to describe everything that can be identifiable with HTTP URIs, from real world entities to intangible concepts. For example, in DBpedia (a *Linked data* version of Wikipedia), the URI `dbo:SoccerPlayer`[2] is used to identify the concept of *soccer player*, whereas the URI `dbr:Lionel_Messi`[3] is used to identify a real person. Everything identified by a URI is considered a *resource* and the information about it is described using the RDF[4] language. This information is stored following a triple pattern of the form "*subject*, *predicate*, *object*" conforming a large graph structure. Knowledge representation mechanisms like RDFS[5] and OWL[6] are layered on top of RDF and allow to augment these structure with more expressive semantics. For example, it is possible to define *classes* and relationships between them in ontologies (e.g. *dbo:SoccerPlayer*, `rdfs:subClassOf`, *dbo:Atlhete*) and to specify *resources* as member of that classes (e.g. *dbr:Lionel_Messi*, `rdf:type`, *dbr:SoccerPlayer*). Each mechanism has its own semantics which determines its inference capabilities and its complexity. From an overall perspective, a *Linked dataset* is a knowledge base of RDF triples which has been built following the *Linked Data* principles. The information contained in these datasets can be divided into two levels: *schema level* and *instance (or data) level*. *Schema level* refers to *terminological knowledge* (TBox), for example, classes, properties and their relationships normally defined through ontologies. In the other hand, *instance level* refers to *assertional knowledge* (ABox),

[2]dbo: prefix for DBpedia ontology definitions http://dbpedia.org/ontology/
[3]dbr: prefix for DBpedia resources http://dbpedia.org/resource/
[4]https://www.w3.org/RDF/
[5]https://www.w3.org/TR/rdf-schema/
[6]https://www.w3.org/standards/techs/owl#w3c_all

978-1-5090-4470-2/16 $31.00 © 2016 IEEE
DOI 10.1109/WI.2016.75

that is, propositions about entities of a specific domain of interest like *class assignment* assertions. From a mathematical-logic perspective *schema* and *instance* level assertions are RDF triples considered as *propositions* and conform the basic elements of any *Semantic Web* reasoning process. For example, the notation $\{p_1, p_2\} \models \{p_3, p_4\}$, where $\models$ is called *entailment relation*, states that propositions p_3 and p_4 (also p_1 and p_2) are logical consequences of propositions p_1 and p_2 obtained under a certain set of rules. In the following sections we will use these definitions to explain our approach.

III. Related Work

Regarding *redundancy*, most of the work focused on RDF compression techniques from a syntactic aspect and only a few of them addressed this issue from a semantic perspective. In [3], the proposed compression technique is based on logical rules which are used to prune triples that then are inferred applying those rules during decompression. In [4], a systematic approach based on graph analysis strategies is proposed to make redundancy information explicit and available to *Linked Data* users. With respect to *consistency*, although there are several tools (such as reasoners) to check the validity of a knowledge base with respect to a formal specification, apply them to large datasets will require scalable solutions due to the complexity of the algorithms. For this reason, some works addressed *consistency* considering just a restricted problem like detecting *resources* as member of *classes* defined as disjointed using pattern-based techniques [5], [6]. Regarding *semantic accuracy*, relevant approaches used restrictions defined over a specific dataset to detect wrong data. These restrictions are basically conditions that data must meet and can be defined manually or detected automatically. To achieve this, some works implemented constraint rules (such as functional dependency rules) [7], [8] while others performed this task by developing techniques based on users evaluations [9], [10], statistical analysis [11] or schema enrichment [6]. Inspired in some ideas proposed on the mentioned works we restrict the study of *redundancy*, *consistency* and *accuracy* to a *class assignment* perspective. The aim of this limitation is to easily understand how these dimensions are related and facilitate the quality assessment in the *meaning* of the data.

IV. Class assignments

A *class assignment* (CA) is an *instance level* proposition that states that a *resource* belongs to a certain class (e.g. *dbr:Lionel_Messi*, `rdf:type`, *dbo:SoccerPlayer*). A *class assignment set* for a resource r (CAS_r) is then the set of all *class assignments* propositions in a *Linked dataset* that specifies the classes for r. From a semantic data quality perspective, it would be desirable that these CASs be non-redundant, consistent and accurate. In this way, we define an ***adequate class assignments set*** for r to those CASs that meet the mentioned conditions. In the following subsections we will show how *redundancy*, *consistency* and *accuracy* concepts can be interpreted in *class assignments* and how they can be used to evaluate, detect or even fix some errors in the data.

A. Non-redundant class assignments

The concept of *data redundancy* can be associated to what is known in mathematical logic literature as *independence*, that is, the ability to deduce a proposition from other propositions (or not). Formally, a set of propositions P is defined as *independent* if for all proposition $p_i \in P$ does not hold that $P - \{p_i\} \models p_i$. This means that if P is independent, each $p_i \in P$ can not be obtained from the remaining propositions of P. Furthermore, given a set of proposition P there is an independent set $Q \subseteq P$ such that $Q \models P$. This means that given a set of propositions P, a set Q can be obtained by reducing the elements of P but without missing information since both P and Q have the same *logical consequences*.

In the *Linked Data* context, redundancy has been considered as "*wasted* space to represent certain meaning in the *Web of Data* environment" [4] and it is related with the concept of *conciseness* [12]. From a semantic perspective this means that certain data can be removed without causing any changes in its meaning. Considering the above, the following definition is proposed:

*Definition 1 (**Non-redundant class assignments set - NRCAS**):* given a resource r and its *class assignment set* (CAS_r), it is *non-redundant* if it is *independent*.

Example 1. Consider the following *schema level* propositions extracted from the DBpedia ontology[7] about classes *dbo:Person*, *dbo:Athlete* and *dbo:SoccerPlayer*:

- p_1: *dbo:Athlete* `rdfs:subClassOf` *dbo:Person*,
- p_2: *dbo:SoccerPlayer* `rdfs:subClassOf` *dbo:Athlete*,

Then, consider the following CAS for a given resource `r`:

- p_3: *r* `rdf:type` *dbo:Person*
- p_4: *r* `rdf:type` *dbo:Athlete*
- p_5: *r* `rdf:type` *dbo:FootballPlayer*

As we can see, the given CAS_r is redundant because if we remove propositions p_3 and p_4 from CAS_r, it can be deduced from the remaining proposition p_5 if transitivity of `rdfs:subClassOf` predicate is considered (e.g. if *r* belongs to class `C` and `C` is a subclass of `B` then *r* belongs to `B`). But if p_5 is removed, it can not be obtained from propositions p_3 and p_4. Thus, a *non-redundant class assignment set* for the resource r ($NRCAS_r$) would be $\{p_5\}$.

Example metric 1. Given a resource r, a redundancy score RS_r can be computed as follows:

$$RS_r = 1 - \frac{\#NRCAS_r}{\#CAS_r}$$

For the given example, our RS_r would be 0,66 ($1 - \frac{1}{3}$), which can be interpreted as 66% of *class assignment* for resource r are redundant. The symbol $\#$ specifies cardinality.

[7] http://mappings.dbpedia.org/server/ontology/classes/

Data curation strategy: in order to achieve a *non redundant class assignment* for a resource r, we can get the CAS_r, then compute the $NRCAS_r$ and finally set up this set as the new CAS_r. Note that to recover the complete *class assignment* information related to r inference techniques must be applied.

B. Consistent class assignments

From a model-theoretics semantics perspective, a set of propositions P of a knowledge base is *consistent* (or *satisfiable*) if it has at least one model. Besides, a set of propositions has a model if and only if every finite subset of it has a model (compactness property). In the other hand, it is *inconsistent* (or *contradictory*), if it is not satisfiable [13]. Furthermore, P is considered *maximal consistent* if P is consistent and for any other set R, if P is a proper subset of R ($P \subset R$) then R is inconsistent.

In the *Linked Data* context, *consistency* means that "a knowledge base is free of logical contradictions with respect to a particular knowledge representation and inference mechanisms" [12]. In *Linked datasets*, one of the most common forms of inconsistencies comes from the use of *disjoint classes* [14]. In this case, the OWL predicate `owl:disjointWith` is used to relate classes whose intersection is empty and inconsistencies at *instance level* occur when a resource is defined as member of classes that should not have elements in common. Considering this, the following definition is proposed:

*Definition 2 (**Consistent class assignments set - CCAS**):* given a resource r and its *class assignment set* (CAS_r) it is *consistent* ($CCAS_r$) if contains the maximal amount of propositions from CAS_r and for each pair of them (p_j, p_k) does not hold that (c_j, `owl:disjointWith`, c_k) where c_k and c_j are the classes specified in proposition p_k and p_j, respectively.

Example 2. Consider the following knowledge base fragment with one resource r and three classes *dbo:Person*, *dbo:Athlete* and *dbo:Place*:

- p_1: *dbo:Person* `owl:disjointWith` *dbo:Place*
- p_2: *r* `rdf:type` *dbo:Person*
- p_3: *r* `rdf:type` *dbo:Athlete*
- p_4: *r* `rdf:type` *dbo:Place*

Proposition p_1 corresponds to a *schema level* assertion meanwhile the last three conform the *class assignments set* of r (CAS_r). As *dbo:Person* and *dbo:Place* are defined as disjointed classes, CAS_r is an inconsistent set. Note, for example, that if we leave out proposition p_4 (or p_2 and p_3), the resulting set would be consistent. Thus, the sets of propositions $\{p_2\}$,$\{p_3\}$,$\{p_4\}$ and $\{p_2,p_3\}$ are $CCAS$s for r but only the last two are maximal consistent sets ($CCAS_r$ does not include the inferred propositions).

Example metric 2. Given a resource r, a consistency score CS_r can be computed as follows:

$$CS_r = 1 - \frac{\#CCAS_r}{\#CAS_r}$$

For the given example, if we take $\{p_2,p_3\}$ as the $CCAS_r$ our CS_r would be 0,33 ($1 - \frac{2}{3}$), which can be interpreted as 33% of *class assignment* for resource r are inconsistent. Note that we may have more than one maximal $CCAS_r$ and the score value will depend on the strategy selected to choose the most appropriate of these sets.

Data curation strategy: in order to achieve a *consistent class assignment* for a resource r, we can get the CAS_r, compute consistent subsets of CAS_r, determine which of them are maximal consistent and finally set up one of these sets as the new CAS_r. Determining which maximal $CCAS$ is the most appropriate for a given resource may require additional analysis that includes some extra information about the resource context.

C. Accurate class assignments

In contrast to the concepts of *redundancy* and *consistency*, *accuracy* can not be formally described without considering the data context. In *Linked datasets*, *semantic accuracy* refers to "the degree to which data correctly represent real world facts" [12]. To describe what a "real world fact" is we need to consider a specific *Linked dataset* scenario and define some restrictions. These restrictions could be, for example, constraint rules predefined manually or temporal axioms detected automatically by enrichment strategies. For *class assignments*, restrictions must describe the conditions or requirements that a *class assignment* for a given resource must meet to be valid. Considering this, the following definition is proposed:

*Definition 3 (**Accurate class assignments set - ACAS**):* given a resource r and its *class assignment set* (CAS_r) it is accurate if each $p \in CAS_r$ meets a set of predefined *class assignment restrictions*.

Example 3. Consider the following knowledge base fragment with classes *dbo:Artist*, *dbo:Athlete*, *dbo:SoccerPlayer* and *dbo:Book*, and some *class assignments* for a resource r:

- p_1: *r* `rdf:type` *dbo:Artist*
- p_2: *r* `rdf:type` *dbo:Athlete*
- p_3: *r* `rdf:type` *dbo:SoccerPlayer*
- p_4: *r* `rdf:type` *dbo:Book*

The above four propositions conform the *class assignments set* for r (CAS_r) and we want to detect which subsets are accurate *class assignments* and which are not. A simple strategy consists on grouping and counting similar classes to detect outliers. In the mentioned example, *dbo:Artist*, *dbo:Athlete* and *dbo:SoccerPlayer* can be considered as similar classes and conform a group meanwhile *dbo:Book* is not semantically related with them. As the first group contains more elements than the second one, we can take it

as the *accurate class assignments set* for r ($ACAS_r$). This simple technique not ensures that class *dbo:Book* will be inaccurate for r but shows how restrictions can be defined to detect it: we first detect a group of prevalent classes and define a restriction that states that a *class assignment* is valid if the class involved belongs to that group.

Example metric 3. Given a resource r, an accuracy score AS_r can be computed as follows:

$$AS_r = 1 - \frac{\#ACAS_r}{\#CAS_r}$$

For our example, if the $ACAS_r$ have the propositions $\{p_1, p_2, p_3\}$, the CS_r value would be 0,25 ($1 - \frac{3}{4}$) which can be interpreted as 25% of *class assignment* for resource r are inaccurate with respect to certain *class assignment restrictions*.

Data curation strategy: computing *accurate class assignment* for a resource r implies that restrictions must be predefined. These restrictions can be defined manually by users or automatically discovered using different techniques. Then, we can get the $ACAS_r$ from CAS_r by applying these constraints and set it up as the new CAS_r.

V. Discussion

In previous sections we have seen that *redundancy* and *consistency* are intrinsic dimensions (can be defined regardless data context) meanwhile *accuracy* is a dependent data context dimension. Although quality levels related to these dimensions may vary widely and its relevance depends on the application at hand, the main challenge is to make this information explicit to users. Considering this, some issues need to be taken into consideration:

- *Linked databases* are huge heterogeneous knowledge bases. This implies that to check *consistency* and *redundancy* it is necessary to take into account scalable *reasoning* techniques [15]. The challenge is to find a balance between the expressiveness of the underlying semantics and the complexity of the reasoning algorithm.
- *Inconsistency* does not necessary means wrong data. As we are working with interlinked heterogeneous data, information about a same resource in different datasets can be inconsistent if it is integrated [5]. The challenge is to detect which *class assignments* need to be removed in order to achieve an acceptable level of *consistency*.
- *Redundancy* does not necessary implies wasted space. In some cases, additional data can be used to improve performance of knowledge bases in query response times or it can be useful to detect inaccurate *class assignments*. The challenge is to detect when redundancy is desirable and when it cause negative effects in our datasets.

VI. Conclusions and future work

In this work, we proposed a semantic interpretation of *accuracy*, *consistency* and *redundancy* quality dimensions in order to achieve *adequate class assignments* in *Linked Data* knowledge bases. Addressing these dimensions together as complementary concepts allowed us to cover a range of quality problems related to the "meaning" of the information at *instance* level. The quality assessment from a *class assignment* perspective may facilitate the definition of quality metrics and the development of mechanisms to detect and eventually fix wrong data. In future work, we plan to implement these mechanisms and evaluate our approach in real use case scenarios with the aim of improve the existing metrics and understand how they correlate with each other. Furthermore, we will study how the addressed quality dimensions are related with *schema level* characteristics of class hierarchies (e.g. number of classes, class specificity, etc.).

References

[1] T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space*, 1st ed. Morgan & Claypool.

[2] S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. van Nuffelen, C. Stadler, S. Tramp, and H. Williams, "Managing the life-cycle of linked data with the lod2 stack," in *Proceedings of the 11th International Conference on The Semantic Web - Volume Part II*, ser. ISWC'12, 2012, pp. 1–16.

[3] A. K. Joshi, P. Hitzler, and G. Dong, *The Semantic Web: Semantics and Big Data: 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. Logical Linked Data Compression, pp. 170–184.

[4] H. Wu, B. Villazón-Terrazas, J. Z. Pan, and J. M. Gómez-Pérez, "How redundant is it? - an empirical analysis on linked datasets," in *COLD*, ser. CEUR Workshop Proceedings, O. Hartig, A. Hogan, and J. Sequeda, Eds. CEUR-WS.org.

[5] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres, "Weaving the Pedantic Web," in *Linked Data on the Web Workshop (LDOW2010) at WWW2010*.

[6] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri, "Test-driven evaluation of linked data quality," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14, 2014, pp. 747–758.

[7] C. Fürber and M. Hepp, "Swiqa - a semantic web information quality assessment framework," in *ECIS*, V. K. Tuunainen, M. Rossi, and J. Nandhakumar, Eds.

[8] B. He, L. Zou, and D. Zhao, "Using conditional functional dependency to discover abnormal data in rdf graphs," in *Proceedings of Semantic Web Information Management on Semantic Web Information Management*, ser. SWIM'14, 2014, pp. 43:1–43:7.

[9] M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, and J. Lehmann, *The Semantic Web – ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. Crowdsourcing Linked Data Quality Assessment, pp. 260–276.

[10] A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, and J. Lehmann, "User-driven quality evaluation of dbpedia," in *Proceedings of the 9th International Conference on Semantic Systems*, ser. I-SEMANTICS '13, 2013, pp. 97–104.

[11] H. Paulheim and C. Bizer, "Improving the quality of linked data using statistical distributions," *Int. J. Semant. Web Inf. Syst.*, vol. 10, no. 2, pp. 63–86, Apr. 2014.

[12] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer, "Quality assessment for linked data: A survey," *Semantic Web Journal*, 2015.

[13] P. Hitzler, M. Krötzsch, and S. Rudolph, *Foundations of Semantic Web Technologies*. Chapman & Hall/CRC, 2009.

[14] A. Hogan, J. Z. Pan, A. Polleres, and Y. Ren, *Scalable OWL 2 Reasoning for Linked Data*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 250–325.

[15] A. Polleres, A. Hogan, R. Delbru, and J. Umbrich, *Reasoning Web. Semantic Technologies for Intelligent Data Access: 9th International Summer School 2013, Mannheim, Germany, July 30 – August 2, 2013. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, ch. RDFS and OWL Reasoning for Linked Data, pp. 91–149.