A Linked Data Scalability Challenge: Concept Reuse Leads to Semantic Decay

Paolo Pareti
University of Edinburgh
United Kingdom
p.pareti@sms.ed.ac.uk

Ewan Klein
University of Edinburgh
United Kingdom
ewan@inf.ed.ac.uk

Adam Barker
University of St Andrews
United Kingdom
adam.barker@st-andrews.ac.uk

ABSTRACT

The increasing amount of available Linked Data resources is laying the foundations for more advanced Semantic Web applications. One of their main limitations, however, remains the general low level of data quality. In this paper we focus on a measure of quality which is negatively affected by the increase of the available resources. We propose a measure of semantic richness of Linked Data concepts and we demonstrate our hypothesis that the more a concept is reused, the less semantically rich it becomes. This is a significant scalability issue, as one of the core aspects of Linked Data is the propagation of semantic information on the Web by reusing common terms. We prove our hypothesis with respect to our measure of semantic richness and we validate our model empirically. Finally, we suggest possible future directions to address this scalability problem.

Categories and Subject Descriptors

I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Semantic networks*

Keywords

Linked Data; Semantic Richness; Information Content

1. INTRODUCTION

Recent years have seen a constant increase in the amount of available Linked Data resources. While the problem of data availability is gradually reducing, data quality remains one of the main limiting factors of Linked Data applications. At the root of this problem lies the open nature of the Web, where knowledge can be erroneous, incomplete, and where concepts can be used in different ways between different sources. In this paper we focus on a particular measure of quality, namely the semantic richness of Linked Data concepts. Following a number-of-features approach [10], we define our measure of semantic richness as the amount of facts that we can expect to infer from a concept.

Web-scale reuse of semantic formalisations, such as concepts and relations, is an essential aspect of Linked Data [2]. However the openness of Linked Data results in the paradox that, while the reuse of terms is considered a good practice, it also progressively decreases the semantic richness of the reused terms. This trend of diminishing semantics as the usage of a term increases is a severe limitation to the scalability of Linked Data. In fact, the less semantically rich a concept becomes the less assumptions can be safely made about its entities, thus reducing its potential applications.

For example, the concept of *Person* in a restricted domain, such as an individual organisation, might follow some particular conventions, such as having an office number and an address. However, if we extend the usage of this concept to more and more entities outside the organisation, then it might not be possible to assume that a *Person* has an office number, or maybe even an address. In the Semantic Web, by virtue of the fact that *anybody can say anything about anything*, an entity of type *Person* could be anything. This problem affects any Linked Data term, including relations. A notable example is the *owl:sameAs*[1] relation, originally intended to represent strict equality between two entities, which is frequently misused to represent a wide range of weaker similarity relations [5].

The main contributions of this work are two. The first is the validation of our hypothesis that the more datasets reuse a Linked Data concept, the less semantically rich it becomes. The second is a measure of semantic richness that can be used to compare the usage of Linked Data concepts between different datasets. This problem will be discussed in more detail in Section 2 and it will be contextualised with previous research in Section 3. In Section 4 we define our measure of semantic richness and we prove our hypothesis with respect to this measure. In Section 5 we validate our hypothesis empirically to provide insights on the extent of this problem. We will then suggest possible future directions to address this problem in Section 6.

2. PROBLEM DEFINITION

We define a measure of semantic richness Γ as a function that takes a concept α, such as *foaf:Person*[2] and a Linked Data graph S, such as the DBpedia graph [1], and returns a positive real number. This number quantifies the amount of information conveyed by the concept α within the graph S. In other words, it is a measure of how many facts we can expect to infer about the instances of that concept that are found in the graph. If $\Gamma(\alpha, S_1) > \Gamma(\alpha, S_2)$ then knowing that an entity is of type α entails more facts in the context of the graph S_1 rather than S_2.

When creating such measure it is important to consider scalability and robustness, as there is large variability in the size and quality of Linked Data resources. It should be possible to compute this measure effectively over large amounts of Linked Data. Moreover, this measure should be applicable to any dataset, even when taxonomic information

---

[1] http://www.w3.org/2002/07/owl#sameAs
[2] http://xmlns.com/foaf/0.1/Person

is not available or when all the available entities are of a single type. These requirements prevent the applicability of existing semantic measures such as the Information Content.

Being based on the usage of a concept, rather than on its position within a conceptual framework such as a taxonomy, this notion differs from the notion of semantic specificity. For example, while a concept is less semantically specific (or less informative) than its sub-concepts, it might be more semantically rich. This situation occurs when the probability of an entity being a member of a sub-concept depends on the other sub-concepts it is a member of, such as in case of mutual exclusivity of sub-concepts. For example, in a dataset S a semantically rich concept α might contain a small set of entities s of which nothing is known. If we consider those entities as members of concept β, we will have that $\Gamma(\beta, S) < \Gamma(\alpha, S)$ although β is a sub-concept of α.

We define S^{α_1} and S^{α_2} as the sets of entities of type α contained, respectively, in datasets S_1 and S_2. Our hypothesis is that the reuse of a Linked Data concept results in a decrease of its semantic richness. Assuming that the sets of entities of type α in two different datasets are disjoint ($S^{\alpha_1} \cap S^{\alpha_2} = \emptyset$), we express this hypothesis as follows:

$$\Gamma(\alpha, S_1 \cup S_2) \leq \frac{|S^{\alpha_1}|\Gamma(\alpha, S_1) + |S^{\alpha_2}|\Gamma(\alpha, S_2)}{|S^{\alpha_1}| + |S^{\alpha_2}|} \qquad (1)$$

Intuitively, this inequation states that the semantic richness of a concept with respect to the union of two graphs is at most equal to the average semantic richness of the graphs. The difference between these values represents the amount of information about a concept that is lost when merging two datasets. In Section 5 we will show how, in practice, the union of two graphs typically results in a significant decrease of semantic richness.

3. BACKGROUND

The notion of Information Content (IC) has been proposed to measure the informativeness of concepts within a taxonomy in the field of Information Theory [11]. This measure has been shown to be effective not only with respect to concepts, but also to measure the informativeness of resources and relations within a knowledge base [8]. The IC of a concept α within a knowledge base is defined as $IC(\alpha) = -log(p(\alpha))$, where $p(\alpha)$ is the probability of an entity belonging to concept α. One limitation to the applicability of this measure is that the probability $p(\alpha)$ might not be meaningful when considering domain specific datasets. For example, several existing datasets contain only instances of type *foaf:Person*. The IC of this concept computed over these datasets will always be equal to 0 since $p(foaf:Person) = 1$. The IC measure would not distinguish between the different datasets although they might have a different semantic interpretation of the concept *foaf:Person*. A different measure is therefore necessary to compare the amount of information conveyed by the concept in each dataset.

In the field of Cognitive Neuroscience several measures of semantic richness have been proposed [10]. These measures have been used to explain why certain concepts can be processed better or faster by humans to perform certain tasks. We adopt one such measure, called number-of-features (NF), by measuring the number of facts that characterise a Linked Data concept. The basic intuition behind the NF measure is that semantically rich concepts tend to be characterised with more attributes than less semantically rich ones.

Our metric to compute the semantic richness of a concept does not assume the availability of pre-existing inference rules. In order to calculate how many facts we can infer about the entities of a given concept we adopt an Inductive Learning approach. For example, if all the entities of concept α from a graph S share property i, then we can induce the rule that if an entity belongs to α then it will have property i. The applications of Inductive Learning for Linked Data have received considerable attention in the recent years thanks to an increasing availability of Linked Data resources. The main advantages of Inductive Learning approaches are scalability and robustness to errors and uncertainty [4].

Schema information can be inferred from a knowledge base by considering the frequency of certain patterns in the data, such as the common properties shared by the entities of the same type [3]. This schema information can then be used to evaluate the quality of the knowledge base. For example, it is possible to evaluate how well a Linked Data graph matches a set of schema rules by using SPARQL query patterns [7].

When considering data from multiple sources, the decrease of semantic richness of a concept can be alleviated by annotating data with contextual information. Several approaches have been proposed to represent contextual information about Linked Data, in particular with respect to Provenance [9]. For example, we might be able to infer more knowledge about a set of entities of a concept if we know that they all originate from DBpedia. One limitation of this approach is the difficulty of creating and maintaining contextual information. Moreover, contextual information does not explicitly describe how a concept is used within a dataset.

4. THE SEMANTIC RICHNESS MEASURE

Given a concept α, such as `foaf:Person`, we define a measure of its semantic richness G which quantifies the amount of information conveyed by the concept. Unlike IC-based measures [6], our measure is not proportional to the probability of an entity being a member of the concept. This probability cannot be estimated reliably for Linked Data due to its open nature and partial observability. When evaluating the semantic richness of a concept we only assume the availability of a representative subset of entities of that concept obtained from a finite number of sources. Taxonomical information about the concept is not required.

Our measure of semantic richness G is defined as a function of the number of facts that we can infer about its instances, and about their probability of being correct. These facts are represented as graph patterns. For example, the fact that persons have a birth date can be represented by the graph pattern: `{ ?person a foaf:Person . ?person foaf:birthday ?date . }` here represented as a SPARQL[3] graph pattern using the FOAF[4] ontology. We define the probability of an entity of type α from dataset S matching pattern i as $p(i, \alpha, S)$. The average number of patterns that we can observe in the entities of type α is the expected value μ of a Poisson binomial distribution of the probabilities p. We only need to compute this value over the finite set of patterns I that have been observed among the entities of the dataset, namely when $p(i, \alpha, S) > 0$. The expected value of

---

[3]`http://www.w3.org/TR/rdf-sparql-query/`
[4]`http://xmlns.com/foaf/spec/`

this distribution can then be computed as follows:

$$\mu = \sum_{i \in I} p(i, \alpha, S) \qquad (2)$$

Equation 2 gives us a measure of how many patterns, in average, we can observe from the entities of type α in dataset S. This measure, however, does not tell us how well we can characterize a concept with a set of common features. In fact, a high value of μ could be equally a result of a large number of infrequent patterns or of a small number of frequent patterns. However, frequent patterns are more useful than infrequent ones for the purpose of characterising a concept, for example to learn schema information using Inductive Learning. Taking this into consideration, we define our measure of semantic richness with respect to a set of characteristic patterns $Y \subseteq I$ that define which patterns we can expect entities of type α to have.

If a pattern i belongs to Y, then whenever we observe an entity ϵ of type α we can infer that it would match pattern i. We base our measure of semantic richness on the correctness of such inferences. If an entity ϵ really matches pattern i then we count it as a correct inference, we count it as incorrect otherwise. For example, we might consider including the pattern "has a dedicated Wikipedia page" in the set of characteristic patterns of concept *foaf:Person*. Whenever we observe an entity of type *foaf:Person* which has a dedicated Wikipedia page we say that inferring this pattern is correct; we say it is incorrect otherwise.

Our measure of semantic richness $G(\alpha, S)$ of a concept α with respect to a dataset S and a set of characteristic patterns Y is the difference between the expected number of correct inferences and the expected number of incorrect ones that we can make about an entity:

$$G(\alpha, S) = \sum_{i \in Y} p(i, \alpha, S) - \sum_{i \in Y}(1 - p(i, \alpha, S))$$
$$G(\alpha, S) = \sum_{i \in Y} 2p(i, \alpha, S) - 1 \qquad (3)$$

From equation 3 it follows that, in order to maximize the value of $G(\alpha, S)$, Y should be the set of patterns which occur in the majority of the entities ($\forall i \in Y.p(i, \alpha, S) > 0.5$). Therefore equation 4 can be rewritten as follows:

$$G(\alpha, S) = \sum_{i \in I} \begin{cases} 2p(i, \alpha, S) - 1 & \text{if } p(i, \alpha, S) > 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad (4)$$

The particular threshold of 0.5 is a result of the fact that in this measure, which aims at being generic, correct and wrong inferences carry the same (absolute) weight. Similarly, all patterns carry equal weight. This formulation of semantic richness is not meant to be unique and variations are possible. For example, equation 4 could be modified to penalize wrong inferences more. Also, different patterns could be given different weights to represent how important or informative they are for a specific application.

In this work we use use our generic measure of semantic richness to analyse our hypothesis, namely that as we consider more entities as members of a concept, the semantic richness of that concept decreases. In fact, assuming that the two sets of entities are disjoint ($S^{\alpha_1} \cap S^{\alpha_2} = \emptyset$) the semantic richness of the union of two sources of data is always inferior or equal to the average individual semantic richness of the sources:

$$G(\alpha, S_1 \cup S_2) \leq \frac{|S^{\alpha_1}|G(\alpha, S_1) + |S^{\alpha_2}|G(\alpha, S_2)}{|S^{\alpha_1}| + |S^{\alpha_2}|} \qquad (5)$$

4.1 Proof of Semantic Richness Decrease

Formula 5 can be proved according to our definition of semantic richness defined in Formula 4. It can be observed that the semantic richness of a concept is the sum of the individual semantic richness of each pattern $G'(\alpha, S, i) = (2p(i, \alpha, S) - 1)[p(i, \alpha, S) > 0.5]$. Therefore Formula 5 can be proved by proving that the inequation holds for every individual pattern:

$$G'(\alpha, S_1 \cup S_2, i) \leq \frac{|S^{\alpha_1}|G'(\alpha, S_1, i) + |S^{\alpha_2}|G'(\alpha, S_2, i)}{|S^{\alpha_1}| + |S^{\alpha_2}|} \qquad (6)$$

Having defined $w_1 = p(i, \alpha, S_1)$ and $w_2 = p(i, \alpha, S_2)$, then the probability $p(i, \alpha, S_1 \cup S_2)$ can be defined as ratio between the number of entities of each set which match pattern i and the number of all the entities in the union of the sets:

$$p(i, \alpha, S_1 \cup S_2) = \frac{|S^{\alpha_1}|w_i + |S^{\alpha_2}|w_2}{|S^{\alpha_1}| + |S^{\alpha_2}|} \qquad (7)$$

Formula 6 can then be written as follows:

$$(2\frac{|S^{\alpha_1}|w_i+|S^{\alpha_2}|w_2}{|S^{\alpha_1}|+|S^{\alpha_2}|} - 1)[\frac{|S^{\alpha_1}|w_i+|S^{\alpha_2}|w_2}{|S^{\alpha_1}|+|S^{\alpha_2}|} > 0.5] \leq$$
$$\frac{|S^{\alpha_1}|(2w_1-1)[w_1>0.5]+|S^{\alpha_2}|(2w_2-1)[w_2>0.5]}{|S^{\alpha_1}|+|S^{\alpha_2}|} \qquad (8)$$

From Formula 8 it follows that in case $w_1 \leq 0.5$ and $w_1 \leq 0.5$ then $p(i, \alpha, S_1 \cup S_2) < 0.5$ and therefore Formula 8 is reduced to the tautology $0 \leq 0$. This matches the intuition that if a pattern was not frequent in any of the original datasets, it will not be frequent also in the union of the datasets.

In case $w_1 > 0.5$ and $w_1 > 0.5$ then $p(i, \alpha, S_1 \cup S_2) > 0.5$ and therefore equation 8 becomes:

$$2\frac{|S^{\alpha_1}|w_i+|S^{\alpha_2}|w_2}{|S^{\alpha_1}|+|S^{\alpha_2}|} - 1 \leq \frac{|S^{\alpha_1}|(2w_1-1)+|S^{\alpha_2}|(2w_2-1)}{|S^{\alpha_1}|+|S^{\alpha_2}|} \qquad (9)$$

This equation can be simplified to the tautology $0 \leq 0$. This matches the intuition that a pattern which was frequent in both of the original datasets will still be frequent in the union of the datasets.

We then consider the last case in which a pattern is frequent only in one of the original datasets: when $w_1 > 0.5$ and $w_2 \leq 0.5$. This requires us to consider two sub-cases, depending on the value of $p(i, \alpha, S_1 \cup S_2)$. If $p(i, \alpha, S_1 \cup S_2) \leq 0.5$, then Formula 8 can be simplified to:

$$0 \leq \frac{|S^{\alpha_1}|(2w_1-1)}{|S^{\alpha_1}|+|S^{\alpha_2}|} \qquad (10)$$

This formula can be further simplified to $w_1 > 0.5$ which is true by assumption. Finally, in case $w_1 > 0.5$ and $w_2 \leq 0.5$ and $p(i, \alpha, S_1 \cup S_2) > 0.5$ Formula 8 can be simplified to:

$$2\frac{|S^{\alpha_1}|w_i+|S^{\alpha_2}|w_2}{|S^{\alpha_1}|+|S^{\alpha_2}|} - 1 \leq \frac{|S^{\alpha_1}|(2w_1-1)}{|S^{\alpha_1}|+|S^{\alpha_2}|} \qquad (11)$$

This equation can be simplified to $w_2 \leq 0.5$ which is true by assumption. Having proved Formula 6, and therefore Formula 5, we have proved our hypothesis (Formula 1) with respect to our measure of semantic richness.

5. EXPERIMENTAL EVALUATION

Having defined our model to measure semantic richness we perform an empirical validation of this measure and of our hypothesis. This experimental evaluation, performed over several large datasets, also provides evidence of the scalability of our measure. These experiments are based on an implementation of a system that can sample entities from a dataset and calculate the semantic richness of a concept within the dataset.
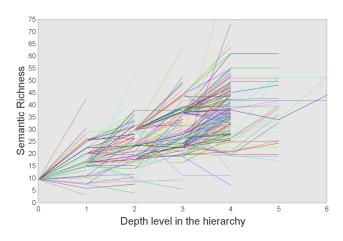
Figure 1: The DBpedia ontology tree represented in function of the semantic richness of each concept. Each segment represents a subclass relation.

Figure 2: Decrease of semantic richness of the concept *foaf:Person* of ten different datasets as we add more entities from the other datasets.

5.1 Validation of the Semantic Measure

Our approach uses a numerical measure of the semantic richness of a concept using equation 4. But how well does this measure actually capture the concept of semantic richness? We evaluate this measure according to the intuition that concepts are, in average, less semantically rich than their sub-concepts. A similar property is also found in the IC measure, as the IC of a concept is, by definition, always inferior or equal to the IC of its sub-concepts.

We have evaluated how well our measure matches this intuition on the concepts of the DBpedia ontology[5] by calculating the semantic richness of the entities of each concept.[6] To reduce the computational complexity, all concepts with over 1000 entities have been replaced by a randomly sampled set of 1000 entities. Also, we have considered only simple graph patterns which involve a single triple. More specifically, we have considered patterns as pairs $< p, o >$ which match an entity ϵ if the triple $< \epsilon, p, o >$ is asserted in the dataset. Improving the quality of this measure by considering more complex patterns is left for further research.

Figure 1 shows the change of semantic richness between each concept and its sub-concepts. This graph shows a significant tendency of concepts to be less semantically rich than their sub-concepts. In fact, 90% of the subclass relations resulted in an increase of semantic richness from a concept to its sub-concepts. Also, in all cases concepts resulted in a lower semantic richness than the average semantic richness of their sub-concepts, thus supporting our hypothesis. Only in 10% of the cases a subclass relation resulted in a decrease of semantic richness from a concept to a sub-concept. This situation occurred because in the DBpedia ontology, sibling concepts are mutually exclusive. Therefore, the knowledge that an entity belongs to a certain concept affects the probability of it being a member of another, potentially more semantically rich, concept.

5.2 Validation of the Hypothesis

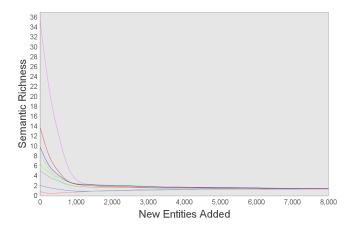We have run an experiment to quantify the decrease in semantic richness as concepts are reused. In this experiment

---

[5] http://dbpedia.org/ontology/
[6] The entities were extracted from DBpedia version 3.5.1

Table 1: SPARQL endpoints used in the experiment (accessed on 7/3/2015)

http://dbpedia.org/sparql
http://data.nobelprize.org/sparql
http://data.archiveshub.ac.uk/sparql
http://dati.camera.it/sparql
http://data.utpl.edu.ec/ecuadorresearch/lod/sparql
http://lod.sztaki.hu/sparql
http://semantic.eea.europa.eu/sparql
http://data.open.ac.uk/query
http://semanticweb.cs.vu.nl/europeana/sparql
http://services.data.gov.uk/reference/sparql

we have considered the concept *foaf:Person* as it is used in ten different datasets. The list of the endpoints used to access the datasets is shown in Table 1. From each endpoint we have extracted information about 1000 randomly selected entities.

For each dataset, we have calculated its semantic richness at different states. Starting from the semantic richness of the original dataset, we have recalculated this measure after adding progressively more entities from the other datasets. Figure 2 shows the progressive decay of semantic richness as we add new entities of the same type from other sources. The semantic richness of each dataset quickly converges to a value which is significantly inferior to the average measure. While the average semantic richness of all the individual sources is 9.6, the semantic richness of the merged set results is 1.3. This experiment validates our hypothesis since the semantic richness of the union of a number of sources is inferior to their average individual semantic richness. Another fact that can be observed is that different datasets represent the same concept with significantly different levels of semantic richness. In this case, the values ranged from 0.7 of `services.data.gov.uk` to 35.2 of `dbpedia.org`.

6. RETAINING SEMANTIC RICHNESS

Since Linked Data does not allow negation, it is not possible to prevent an entity from becoming a member of a concept.

It is however possible to predict the effect that adding an entity to a dataset would have on the semantic richness of a concept. We define E as the set of patterns that are matched by entity ϵ. The degree of typicality of an entity with respect to a concept α can then be defined as $\delta^\epsilon = |E \cap Y|/|Y|$ where Y is the set of frequent patterns of α. Given two entities ϵ and ϵ', we can say that ϵ is more typical than ϵ' if $\delta^\epsilon > \delta^{\epsilon'}$. We define an entity to be *atypical*, meaning that including it into a concept would result in a decrease of its semantic richness, if $\delta < 0.5$. If $\delta \geq 0.5$ we say that the entity is *typical*, meaning that its inclusion in the concept would maintain or increase its semantic richness.

This typicality measure is scalable and easily verifiable because it can be computed automatically over a subset of the data. In fact, any agent can calculate whether an entity is a typical or an atypical member of a concept. In a closed domain, this measure could be used to determine which entities should be considered as members of a concept, and which should not. In Linked Data, however, it is not possible, nor desirable, to prevent entities from becoming members of a concept. It is however possible to add more knowledge about the typical entities to preserve their semantic richness. A possible approach to do this is to create a sub-concept which groups together the typical entities.

For example, the particular way in which the concept *foaf:Person* is used in DBpedia results in the property that the majority of those entities have an associated Wikipedia page. If we merge these entities with another source about persons which are not related with a Wikipedia page, then this property will no longer hold, thus reducing the semantic richness of the concept. In order to preserve this property, we can create a new sub-concept of *foaf:Person* that captures the particular interpretation of this concept adopted by DBpedia.

7. CONCLUSION

In this paper we have proposed a measure of semantic richness of Linked Data concepts that quantifies the amount of facts that can be inferred from a concept within the scope a particular dataset. This measure differs from previous semantic measures, such as Information Content, as it is meant to compare the difference in semantic richness of the same concept between different datasets. We have validated this measure with respect to the DBpedia ontology by showing that it captures the expected intuition that sub-concepts are, in average, more semantically rich than their super-concepts. We have used the mathematical formulation of this measure to prove our hypothesis that the more a concept is reused, the less semantically rich it becomes. To verify the extent of this problem we have compared the semantic richness of the concept *foaf:Person* from ten different existing datasets. Our preliminary results show that the semantic richness of a concept quickly decreases when merging entities of multiple datasets together. We have also shown that the same concept can be used in different datasets with significantly different levels of semantic richness. Since it is not desirable to prevent concepts from being reused we propose to preserve their semantic richness by organising their heterogeneous set of entities in a number of sub-concepts. These sub-concepts can be generated and populated automatically by considering the effect on their semantic richness that would result after the inclusion of a new entity.

8. REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *The Semantic Web*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer Berlin Heidelberg, 2007.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.

[3] L. Bühmann and J. Lehmann. Pattern based knowledge base enrichment. In *The Semantic Web - ISWC 2013*, volume 8218 of *Lecture Notes in Computer Science*, pages 33–48. Springer Berlin Heidelberg, 2013.

[4] C. d'Amato, N. Fanizzi, and F. Esposito. Inductive learning for the semantic web: What does it buy? *Semantic Web*, 1(1,2):53–59, 2010.

[5] H. Halpin, P. J. Hayes, J. P. McCusker, D. L. McGuinness, and H. S. Thompson. When owl:sameAs Isn't the Same: An Analysis of Identity in Linked Data. In *The Semantic Web - ISWC 2010*, volume 6496 of *Lecture Notes in Computer Science*, pages 305–320. Springer Berlin Heidelberg, 2010.

[6] S. Harispe, S. Ranwez, S. Janaqi, and J. Montmain. Semantic Measures for the Comparison of Units of Language, Concepts or Instances from Text and Knowledge Base Analysis. *arXiv preprint arXiv:1310.1285*, pages 70–76, 2013.

[7] 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*, WWW '14, pages 747–758, New York, NY, USA, 2014.

[8] R. Meymandpour and J. Davis. Linked Data Informativeness. In *Web Technologies and Applications*, volume 7808 of *Lecture Notes in Computer Science*, pages 629–637. Springer Berlin Heidelberg, 2013.

[9] L. Moreau. The Foundations for Provenance on the Web. *Foundations and Trends in Web Science*, 2:99–241, Feb. 2010.

[10] P. M. Pexman, I. S. Hargreaves, P. D. Siakaluk, G. E. Bodner, and J. Pope. There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, 15(1):161–167, 2008.

[11] P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 448–453, San Francisco, CA, USA, 1995.