Rating, recognizing and rewarding metadata integration and sharing on the semantic web

Francisco M. Couto

LASIGE, Dept. de Informática, Faculdade de Ciências, Universidade de Lisboa, Portugal
fcouto@di.fc.ul.pt

Abstract. Research is increasingly becoming a data-intensive science, however proper data integration and sharing is more than storing the datasets in a public repository, it requires the data to be organized, characterized and updated continuously. This article assumes that by rewarding and recognizing metadata sharing and integration on the semantic web using ontologies, we are promoting and intensifying the trust and quality in data sharing and integration. So, the proposed approach aims at measuring the knowledge rating of a dataset according to the specificity and distinctiveness of its mappings to ontology concepts.
The knowledge ratings will then be used as the basis of a novel reward and recognition mechanism that will rely on a virtual currency, dubbed KnowledgeCoin (KC). Its implementation could explore some of the solutions provided by current cryptocurrencies, but KC will not be a cryptocurrency since it will not rely on a cryptographic proof but on a central authority whose trust depends on the knowledge rating measures proposed by this article. The idea is that every time a scientific article is published, KCs are distributed according to the knowledge rating of the datasets supporting the article.

Keywords: Data Integration, Data Sharing, Linked Data, Metadata, Ontologies

1 Introduction

Research is increasingly becoming a data-intensive science in several areas, where prodigious amounts of data can be collected from disparate resources at any time [6]. However, the real value of data can only be leveraged through its trust and quality, which ultimately results in the acquisition of knowledge through its analysis. Since multiple types of data are involved, often from different sources and in heterogeneous formats, data integration and sharing are key requirements for an efficient data analysis. The need for data integration and sharing has a long-standing history, and besides the big technological advances it still remains an open issue. For example, in 1985 the Committee on Models for Biomedical Research proposed a structured and integrated view of biology to cope with the available data [8]. Nowadays, the BioMedBridges [1] initiative aims at constructing the data and service bridges needed to connect the emerging biomedical sciences research infrastructures (BMSRI), which are on the roadmap of the European Strategy Forum on Research Infrastructures (ESFRI). One common theme to

[1] www.biomedbridges.eu

all BMSRIs is the definition of the principles of data management and sharing [3]. The Linked Data initiative [2] already proposed a well-defined set of recommendations for exposing, sharing and integrating data, information and knowledge using semantic web technologies. In this paradigm data integration and sharing is achieved in the form of links connecting the data elements themselves and adding semantics to them. Following and understanding the links between data elements in publicly available Data Linked stores (Linked Data Cloud) enables us to access the data and knowledge shared by others. The Linked Data Cloud offers an effective solution to break down data silos; however the systematic usage of these technologies requires a strong commitment from the research community.

Promoting the trust and quality of data through their proper integration and sharing is essential to avoid the creation of silos that store raw data that cannot be reused by others, or even by the owners themselves. For example, the current lack of incentive to share and preserve data is sometimes so problematic, that there are even cases of authors that cannot recover the data associated with their own published works [5]. However, the problem is how to obtain a proactive involvement of the research community in data integration and sharing. In 2009, Tim Berners-Lee gave a TED talk[3], where he said: "you have no idea the number of excuses people come up with to hang onto their data and not give it to you, even though you've paid for it as a taxpayer." Public funding agencies and journals may enforce the data-sharing policies, but the adherence to them is most of the times inconsistent and scarce [1]. Besides all the technological advances that we may deliver to make data integration and sharing tasks easier, researchers need to be motivated to do it correctly. For example, due to the Galileos strong commitment to the advance of Science, he integrated the direct results of his observations of Jupiter with careful and clear descriptions of how they were performed, which he shared in Sidereus Nuncius [4]. These descriptions enabled other researchers not only to be aware of Galileos findings but also to understand, analyze and replicate his methodology. This is another situation that we could characterize with the famous phrase "That's one small step for a man, one giant leap for mankind." Now let us imagine if we could extend Galileos commitment to all the research community, the giant leap that it could bring to the advance of science.

Thus the commitment of the research community to data integration and sharing is currently a major concern, and this explains why BMSRIs have recently included in their definition of the principles of data management and sharing the following challenge: "to encourage data sharing, systematic reward and recognition mechanisms are necessary". They suggest studying not only measurements of citation impact, but also highlighting the importance to investigate other mechanisms as well. Systematic reward and recognition mechanisms should motivate the researchers in a way that they become strongly committed in sharing data, so others can easily understand and reuse it. By doing so, we encourage the research community to improve previous results by replicating the experiments and testing new solutions. However, before developing a reward and recognition mechanism we must formally define: i) what needs to be rewarded and recognized; ii) and measure its value in a quantitative and objective way.

[2] http://linkeddata.org/

[3] http://www.ted.com/talks/tim_berners_lee_on_the_next_web

2 Metadata Quality

Proper data integration and sharing is more than storing the datasets in a public repository, it requires the data to be organized and characterized in a way that others can find it and reuse it effectively. In an interview[4] to Nature, Steven Wiley emphasized that sharing data "is time-consuming to do properly, the reward systems aren't there and neither is the stick". Not adding links to external resources hampers the efficient retrieval and analysis of data, and therefore its expansion and update. Making a dataset easier to find and access is also a way to improve its initial trust and quality, as more studies analyze, expand and update it. Like the careful and clear descriptions provided by Galileo, semantic characterizations in the form of metadata must also be present so others can easily find the raw data and understand how it can be retrieved and explored.

Metadata is a machine-readable description of the contents of a resource made through linking the resource to the concepts that describe it. However, to fully understand such diverse and large collections of raw data being produced, their metadata need to be integrated in a non-ambiguous and computational amenable way [9, 13]. Ontologies can be loosely defined as "a vocabulary of terms and some specification of their meaning" [7, 14]. If an ontology is accepted as a reference by the community (e.g., the Gene Ontology), then its representation of its domain becomes a standard, and data integration and sharing facilitated. The complex process of enriching a resource with metadata by means of semantically defined properties pointing to other resources often requires human input and domain expertise. Thus, the proposed approach assumes that by rewarding and recognizing metadata sharing and integration on the semantic web using standard and controlled vocabularies, we are promoting and intensifying scientific collaboration and progress.

Figure 1 illustrates the Semantic Web in action with two datasets annotated with its respective metadata using a hypothetical Metal Ontology. A dataset including Gold Market Stats contains an ontology mapping (e.g., an RDF triple) to the concept Gold, and another dataset Silver Market Stats contains an ontology mapping to the concept Silver. Given that Gold and Silver are both coinage metals, a semantic search engine is capable of identifying as relevant both datasets when asked for market stats of coinage metals.

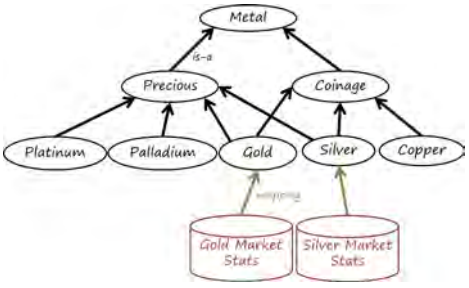

Fig. 1. An hypothetical metal ontology and dataset mappings.

Now, we need to define the value of metadata in terms of knowledge it provides about a given dataset. Semantic interoperability is a key requirement in the realization

[4] http://www.nature.com/news/2011/110914/full/news.2011.536.html

of the semantic web and it is mainly achieved through mappings between resources. For example, all dataset mappings to ontology concepts are to some extent important to enhance the retrieval of that dataset, but the level of importance varies across mappings. The proposed approach assumes that metadata can be considered as a set of links where all the links are equal, but some links are more equal than others (adaption of George Orwells quote). Thus, the proposed approach aims at measuring the knowledge rating of any given dataset through its mappings to concepts specified in an ontology.

3 Knowledge rating

The proposed approach assumes that the metadata integration and sharing value of a dataset, dubbed as **knowledge rating**, is proportional to the **specificity** and **distinctiveness** of its mappings to ontology concepts in relation to all the others datasets in the Linked Data Cloud.

The specificity of a set of ontology concepts can be defined by the information content (IC) of each concept, which was introduced by [11]. For example, intuitively the concept dog is more specific than the concept animal. This can be explained because the concept animal can refer to many distinct ideas, and, as such, carries a small amount of information content when compared to the concept dog, which has a more informative definition. The distinctiveness of a set of ontology concepts can be defined by its conceptual similarity [2, 12] to all the others sets of ontology concepts, i.e. a distinctiveness of a dataset is high if there are no other semantically similar datasets available. Conceptual similarity explores ontologies and the relationships they contain to compare their concepts and, therefore, the entities they represent. Conceptual similarity enables us to identify that arm and leg are more similar than arm and head, because an arm is a limb and a leg is also a limb. Likewise, because an airplane contains wings, the two concepts are more related to each other than wings is to boat.

Most implementations of IC and conceptual similarity only span a single domain specified by an ontology [10]. However, realistic datasets frequently use concepts from distinct domains of knowledge, since reality is rarely unidisciplinary. So, the scientific challenge is to propose innovative algorithms to calculate the IC and conceptual similarity using multiple-domain ontologies to measure the specificity and distinctiveness of a dataset. Similarity in a multiple ontology context will have to explore the links between different ontologies. Such correspondences already exist for some ontologies that provide cross-reference resources. When these resources are unavailable, ontology matching techniques can be used to automatically create them.

4 Reward and recognition mechanism

The reward and recognition mechanism can rely on the implementation of a new virtual currency, dubbed KnowledgeCoin (KC), that will be specifically designed to promote and intensify the usage of semantic web technologies for scientific data integration and sharing. The idea is that every time a scientific article is published, KCs are distributed according to the knowledge rating of the datasets supporting that article. Note that KCs

should by no means be a new kind of money and the design of KC transactions will focus on the exchange of scientific data and knowledge.

After developing the knowledge rating measures, they can be used to implement the supply algorithm of a new virtual currency, KC. This will not only aim at validating the usefulness of the proposed knowledge ratings but also deliver an efficient reward and recognition mechanism to promote and intensify the usage of semantic web technologies for scientific data integration and sharing. Unlike conventional cryptocurrencies, the KCs will rely on a trusted central authority and not on a cryptographically proof. But even without being a cryptocurrency, the KC will take advantage of the technical solutions provided by existing cryptocurrencies, such as bitcoin[5].

The scientific challenge is to create a trusted central authority that issue new KCs when new knowledge is created in the form of a scientific article, as long as it references a supporting dataset properly integrated in the Linked Data Cloud. If there is no reference to the dataset in the Linked Data Cloud no KCs will be issued. This way, researchers will be incentivized to publically share the dataset, including the raw data or at least a description of the raw data, in the Linked Data Cloud. If a dataset is shared through the Linked Data Cloud then its level of integration will be measured by its knowledge rating. This way, researchers will be encouraged to properly integrate their data. The success of this mining process will rely on the trustworthiness of the knowledge ratings, and therefore will further validate the developed measures.

From recognition researchers may get reputation, and from reputation they may get a reward. For example, researchers recognize the relevance of a research's work by citing it, and by having a high number of citations the researcher obtains a strong reputation, which may in the end help him to be rewarded with a project grant. Thus, KCs can be interpreted as a form of reputation that in the end can result in a reward. However, we can also design and implement direct reward mechanisms through KCs transactions as a way to establish a virtual marketplace of scientific data and knowledge exchanges. The main scenario of a KCs transaction is to represent the exchange of datasets identified by an URI from the data provider to the data consumer, which may include recognition statements.

5 Future Directions

The design of the approach is ongoing work and its direction depends on a more detailed analysis of many social and technical challenges that its implementation poses. For example, some of the issues that need to be further studied and discussed: i) knowledge ratings implementation, i.e. their validation, aggregation, performance, exceptions, and extension to any mappings besides the ontological ones; ii) potential abuses, such as the creation of spam mappings and other security threats; iii) central trusted authority for the KC vs. the peer-to-peer mechanisms used by bitcoin; iv) use case scenarios for the KC, e.g. exchange of datasets and their characterization based on KC transactions.

In a nutshell, this paper presents the guidelines for delivering sound knowledge rating measures to serve as the basis of a systematic reward and recognition mechanism

[5] `http://bitcoin.org/`

based on KCs for improving the trust and quality of data through proper data integration and sharing on the semantic web. The proposed idea aims to be the first step in providing an effective solution towards data silos extinction.

Acknowledgments

The anonymous reviewers for their valuable comments and suggestions. Work funded by the Portuguese FCT through the LASIGE Strategic Project (PEst-OE/EEI/UI0408/2014) and SOMER project (PTDC/EIA-EIA/119119/2010).

References

1. Alsheikh-Ali, A.A., Qureshi, W., Al-Mallah, M.H., Ioannidis, J.P.: Public availability of published research data in high-impact journals. PloS one 6(9), e24357 (2011)
2. Couto, F.M., Pinto, H.S.: The next generation of similarity measures that fully explore the semantics in biomedical ontologies. Journal of bioinformatics and computational biology 11(05) (2013)
3. ELIXIR, EU-OPENSCREEN, BBMRI, EATRIS, ECRIN, INFRAFRONTIER, INSTRUCT, ERINHA, EMBRC, Euro-BioImaging, LifeWatch, AnaEE, ISBE, MIRRI: Principles of data management and sharing at European Research Infrastructures (DOI:105281/zenodo8304, Feb 2014)
4. Galilei, G.: Sidereus Nuncius, or The Sidereal Messenger. University of Chicago Press (1989)
5. Goodman, A., Pepe, A., Blocker, A.W., Borgman, C.L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., et al.: Ten simple rules for the care and feeding of scientific data. PLoS computational biology 10(4), e1003542 (2014)
6. Hey, A.J., Tansley, S., Tolle, K.M.: The fourth paradigm: data-intensive scientific discovery (2009)
7. Jasper, R., Uschold, M., et al.: A framework for understanding and classifying ontology applications. In: Proceedings 12th Int. Workshop on Knowledge Acquisition, Modelling, and Management KAW. vol. 99, pp. 16–21 (1999)
8. National Research Council (US). Committee on Models for Biomedical Research: Models for biomedical research: a new perspective. National Academies (1985)
9. Noy, N.F.: Semantic integration: a survey of ontology-based approaches. ACM Sigmod Record 33(4), 65–70 (2004)
10. Pedersen, T., Pakhomov, S.V., Patwardhan, S., Chute, C.G.: Measures of semantic similarity and relatedness in the biomedical domain. Journal of biomedical informatics 40(3), 288–299 (2007)
11. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th international joint conference on Artificial intelligence-Volume 1. pp. 448–453. Morgan Kaufmann Publishers Inc. (1995)
12. Rodríguez, M.A., Egenhofer, M.J.: Determining semantic similarity among entity classes from different ontologies. Knowledge and Data Engineering, IEEE Transactions on 15(2), 442–456 (2003)
13. Uschold, M., Gruninger, M.: Ontologies and semantics for seamless connectivity. ACM SIGMod Record 33(4), 58–64 (2004)
14. Uschold, M., Gruninger, M.: Ontologies: Principles, methods and applications. The knowledge engineering review 11(02), 93–136 (1996)