Decentralizing the Semantic Web: Who will pay to realize it?

Tobias Grubenmann, Daniele Dell'Aglio, Abraham Bernstein,
Dmitry Moor, Sven Seuken

Department of Informatics, University of Zurich, Switzerland,
{grubenmann, dellaglio, bernstein, dmoor, seuken}@ifi.uzh.ch

Abstract. Fueled by enthusiasm of volunteers, government subsidies, and open data legislation, the Web of Data (WoD) has enjoyed a phenomenal growth. Commercial data, however, has been stuck in proprietary silos, as the monetization strategy for sharing data in the WoD is unclear. This is in contrast to the traditional web where advertisement fueled a lot of the growth. This raises the question how the WoD can (i) maintain its success when government subsidies disappear and (ii) convince commercial entities to share their wealth of data.
In this paper, we propose a marketplace for decentralized data following basic WoD principles. Our approach allows a customer to buy data from different, decentralized providers in a transparent way. As such, our marketplace presents a first step towards an economically viable WoD beyond subsidies.

1 Introduction

The *Web of Data* (WoD) is a machine-readable alternative to the traditional World Wide Web. In the WoD, data is exposed in a semantically annotated format which allows machines to easily access the information they need according to the task they are performing. Due to the ease of integration given by the underlying Semantic Web technologies, data sources can be queried in a federated fashion without agreeing on a common scheme beforehand. Hence, the WoD can be seen as one big, decentralized database which can be queried over the Web.

Without financial incentives, many promising datasets will be poorly maintained or be unavailable as relying on volunteers is not enough to keep the data up-to-date and the endpoint running. Indeed, as [2] points out, only a third of all known public endpoints have an uptime of 99% and above.

One main reason is a lack in financial incentives for people to provide data in a semantic format. Unlike in the traditional Web, semantic data is accessed primarily by automatic agents rather than human ones. Therefore, *advertisements* are completely ignored while accessing the data. An alternative to advertisement is to charge a fee for accessing the data. Such strategies are already pursued in the traditional Web by companies like *Bloomberg*, *LexisNexis*, and *Thomson Reuters*. Also, marketplaces like the *Azure DataMarketplace* allow different publishers to sell data with different subscriptions. So far, none of these implemented

markets allow users to buy data in an integrated way from decentralized data providers. Specifically, it is not possible for a user to buy data which constitutes of a join between different datasets from different sources. Hence, applying the aforementioned subscription-based monetization strategies to the WoD is not compatible with the idea of a decentralized Semantic Web. emphHow can we wean the WoD from government subsidies or federation-averse centralization? Finding an answer to this question is crucial to fulfill the promise of the data economy [1].

Our vision is to create a marketplace where decentralized data providers can offer their data and customers can buy answers to SPARQL queries. Such a marketplace is one possible way to make the Semantic Web independent from subsidies and financially sustainable.

2 A marketplace for Semantic Data

We propose to build a marketplace which allows combining data from different sources in an integrated way. Depending on the query and the providers involved, there might be different combination of providers' data that yield non-empty query answers. The market hence needs to (1) decide which datasets to include—a process that is akin to source selection but needs to consider the prices for the different results and (2) determine optimal payments to each of the providers ensuring their participation in the marketplace.

Most data in the Semantic Web is not located in a single endpoint but distributed over several endpoints. Each endpoint can, potentially, contribute to a given query answer. Searching for endpoints offering the required data becomes cumbersome if the number of endpoints increases. A customer may need to access a lot of data out of which only a few (if any at all) end up in the result. Given these problems, we argue for a marketplace which is able to assess individual endpoints on their usefulness for a given query and which can help the customer to decide which data should be bought. As we have shown in [4], deciding whether accessing a certain combination of endpoints would yield a big enough result which is worth the involved costs is a challenging task. As the WoD gets more decentralized, it becomes unlikely that it is possible to accurately evaluate the contribution of a single endpoint towards a query answer without actually executing the query. Join estimation techniques for SPARQL queries might help to sort out endpoints which can hardly contribute towards a query answer. However, for the remaining endpoints, only a query execution can reveal the true contribution and value of an endpoint's data. Hence, we argue that a market for Semantic Data in a decentralized setting has to execute a given query on all promising endpoints *before* the decision can be made which part of the data should be bought by the customer. Of course, this raises interesting questions about *trust*, since sellers need trust that the market will not forward the data to the customer without payment.

Once a query is executed on promising endpoints, the result can be rated by the marketplace and either a buying decision can be made by the market on

behalf of the customer, or a summary of the findings can be given to the customer who can then make a buying decision. Only after the buying decision has been made and the involved payments have been completed, the customer will receive the actual data. Again, *trust* is an important property of our market, as the customer has to rely on the market providing accurate information about the offered data and making buying decisions in the best interest of the customer.

Besides the buying decision, the market has to determine (1) how much a customer has to pay for the query answer and (2) how much payment each provider's contribution to a query answer warrants.

3 Costs of a Query Answer

To discuss the costs involved in producing a query answer, we distinguish between two different roles on the sellers' side: *Provider* and *Host*.

A provider is the originator of data, which is used in the production of a query answer. Providers are responsible for the *quality* of data, including *recentness*, *consistency* and *accuracy* [3]. Providers do not serve their data; this is done by separate entities, the hosts. Hosts operate computers that run SPARQL endpoints for querying data products. They provide the computational and network resources needed to query the providers' data products. Hence, they ensure the reliability, availability, security, and performance, which are usually specified as *Quality of Service* [3].

The separation between host and provider enables more flexible business models for data provision, as some providers might have an initial budget to create data (e.g., government subsidies) but do not have the funds to cover the *operating costs* for running a SPARQL endpoint or may have other reasons to outsource the actual data provision. Providers can decide to act at the same time as a host for their own and/or other provider's data. Nevertheless, we will distinguish between these two different roles and treat them as separate entities.

Data providers might have large fixed costs, which typically accrue whilst *creating* the data. The marginal costs of offering data, however, is (effectively) zero for the provider. This is because, as discussed above, the data is not served by providers but by hosts. Any cost that might occur while offering data is inflicted on the host. It is important to note that even if a provider acts as its own host, the marginal costs are only inflicted on the entity acting as a host, not as a provider.

Like cloud service providers, hosts incur the fixed cost of operating the infrastructure, possibly some variable cost relative in the size of the data they store, and some marginal cost in form of the computational resources spent for each executed query. The host's marginal costs occur whenever the providers' data are queried, independently of whether any data will eventually be bought by a customer.

Data providers rely on the hosts to make their data available to the marketplace and thus, enable customers to buy their data. Similar to a Web host for traditional Web content, hosts in our market concept are paid by the provider,

based on some service agreement. Hence, the providers have to include the hosting costs into their pricing decision. The hosts' costs are already compensated by the providers prior to query execution by the market. Thus, the hosts' costs become transparent to the market and, as a result, the market and customer do not have to take them into account. This facilitates the buying decision.

4 Outlook

To continue growing and being able to serve as a high-quality, decentralized data source, the WoD has to find the means to fund the creation, serving, and maintenance of data sources. In this paper, we proposed a new vision for funding these activities in the form of a marketplace for Semantic Data.

As a precursor to our research, we conducted a pilot study simulating a market platform for the WoD [6]. In [5], we introduced the idea of using a double-auction for the WoD and showed the deficiency of the threshold rule in this setting, together with three ways to correct them. However, our approach assumed that we have access to accurate join-estimates to produce satisfying results – an assumption which might be hard to enforce in the WoD.

Based on our research, we foresee the following challenges in building a market place for Semantic Data:

- Different market mechanisms have to be explored to understand their trade-offs under various market settings.
- Given a market mechanism, providers of Semantic Data have to decide how they will bundle and price the data they are selling. The challenge is to find prices which will satisfy the costumers and allow the providers to cover their costs.
- Our market idea introduces a new metric for source selection, query optimization, and query execution: the financial profitability. Revisiting known techniques and developing new techniques with respect to this new metric will undoubtedly open interesting opportunities for research.

This paper is a first step in the direction of finding stable financing for the WoD. We plan to address the aforementioned challenges in future work and believe that our vision of a marketplace for Semantic Data is a promising way to ensure the financial sustainability of decentralized providers of Semantic Data.

Acknowledgments This work was partially supported by the Swiss National Science Foundation under grant #153598.

References

1. Fuel of the future: Data is giving rise to a new economy. *The Economist*, 2017(May 6th), 2017.
2. C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbussche. SPARQL web-querying infrastructure: Ready for action? In A. H. et al., editor, *The Semantic Web – ISWC 2013.*, volume 8219, pages 227–293, 2013.

3. S. Dustdar, R. Pichler, V. Savenkov, and H.-L. Truong. Quality-aware service-oriented data integration: Requirements, state of the art and open challenges. In *ACM SIGMOD Record*, volume 41, pages 11–19. ACM New York, NY, USA, 2012.
4. T. Grubenmann, A. Bernstein, D. Moor, and S. Seuken. Challenges of source selection in the WoD. In *Proceedings of the International Semantic Web Conference ISWC '17*, Forthcoming 2017.
5. D. Moor, T. Grubenmann, S. Seuken, and A. Bernstein. A double auction for querying the web of data. In *The Third Conference on Auctions, Market Mechanisms and Their Applications*, 2015.
6. M. Zollinger, C. Basca, and A. Bernstein. Market-based sparql brokerage with matrix: Towards a mechanism for economic welfare growth and incentives for free data provision in the web of data. Technical Report IFI-2013.4, 2013.