Competencies as a cost driver of querying Linked Data

Jan Kučera, Vojtěch Svátek, Jindřich Mynarz

University of Economics, Prague

{jan.kucera, svatek, jindrich.mynarz} [at] vse.cz

Abstract

Linked Data Fragments brought an attention to distribution of costs associated with querying Linked Data between servers and clients. Costs associated with consumption of infrastructure resources such as CPU or RAM are often discussed. However from the business perspective the IT infrastructure is not the only factor in deciding the type of interface for publishing data. Competencies also play an important role and developing them therefore acts as a cost driver. We discuss how the necessary competencies vary between three different Linked Data querying scenarios: data dumps, SPARQL endpoint and triple pattern fragments. Results of our analysis could serve as an input into the competency gap analysis.

Keywords

costs, competencies, data consumer, data publisher, Linked Data, semantic web

Introduction

Linked Data (LD) is provided through various types of interfaces such as data dumps, SPARQL endpoints or triple pattern fragments (TPF) [1]. These interfaces could be characterized with server costs in terms of CPU, RAM, and disk use and client costs associated with processing the responses of the server. The type of interface affects the distribution of the total costs of querying LD between servers and clients. For example, data dumps are cheap for servers to provide, but higher costs are usually incurred to clients, whereas providing data through a SPARQL endpoint typically results in higher server costs and lower client costs [1].

Research aimed at the Linked Data Fragments, and specifically the TPF, has already resulted in promising outcomes. From the business perspective, however, costs measured as consumption of information technology (IT) resources are too narrowly defined: publishing and consuming LD also requires relevant competencies. According to [2] the lack of the competencies is one of the roadblocks for publishing and reuse of openly licensed LD. Investments into developing the necessary competencies should therefore be considered.

Our objective is to discuss how these competencies vary between the three LD querying scenarios mentioned. The motivation is that the competencies act as cost a driver and therefore play an important role when deciding about the type of interface to publish LD. With this paper we aim to complement the ongoing discussion about the technical implications of different kinds of interfaces.

Research approach

By competencies we understand abilities and skills of people to effectively perform some tasks. Developing the IT human capital portfolio involves, among other activities, identification of gaps between the desired and the current state of this portfolio and subsequent identification of

scenarios to close the gaps, e.g. training, hiring or outsourcing [3]. Implementing these scenarios requires resources and therefore it is a source of costs. These implications are illustrated in Figure 1.

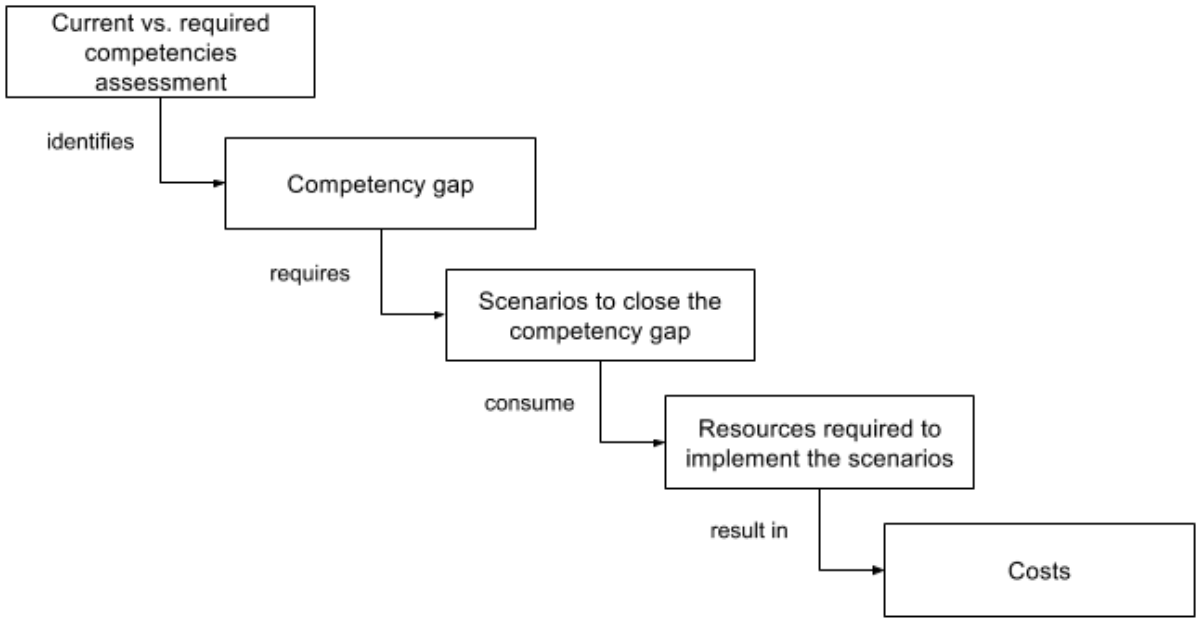

Figure 1: Competency gap as a cost driver

Competency requirements to publish and consume LD could lead to the competency gap. Our analysis highlights the elements that should be taken into account during the gap analysis as a prerequisite of the cost estimation.

We also build upon the roles in the Linked Data Value Chain proposed in [4]. Since we focus on querying LD, we are interested in Linked Data Publishers and Linked Data Consumers. In contrast to [4] we do not use the term Linked Data Application Provider because we do not presume the purpose for which a subject queries LD. Raw data transformation into LD is beyond the scope of our analysis, too.

First, we described the LD querying scenarios and analysed the required competencies. Then, we selected a subset of the case studies from the W3C Semantic Web Education and Outreach (SWEO) Interest Group portal[1] in order to examine domain-specific constellations of potential LD Publishers and Consumers. Since a subject currently consuming human-readable data could possibly consume LD in the future, the End User role was considered as well.

Despite the limited representativeness of this collection (its bulk had been contributed before the uptake of the LD technology), it hints to areas where structured data amenable to processing in RDF abound. In order to focus on case studies featuring (ideally, public) instance data querying, rather than, e.g., reasoning over OWL taxonomies, we selected a subset based on the "SW technologies used" metadata facet on the portal. We required one of the facet's value to be "SPARQL" or "public datasets", which yielded 20 cases.

The analysis of one study took approximately 20-30 minutes to one researcher (co-author of this paper); each analysis was then verified by one different researcher, and a consensual view was subsequently formed.

[1] https://www.w3.org/2001/sw/sweo/public/UseCases/

Linked Data querying scenarios

We did not intend to perform an exhaustive analysis of all possible interfaces and thus we limited our analysis to three LD querying scenarios. Data dumps and SPARQL endpoints were selected because they represent boundary scenarios in the continuum of LD interfaces described in [1]. In order to investigate whether a similar distribution pattern as in the case of the infrastructure resource consumption applies to the required competencies for querying LD, the triple pattern fragments scenario with HDT (Header, Dictionary, Triples, see [5]) files as backend was added (the same TPF backend as in [1]).

In the data dumps scenario (Figure 2) LD Consumer accesses data published by LD Publisher in a form of downloadable files. LD Consumer uses its own IT infrastructure to make the data queryable.

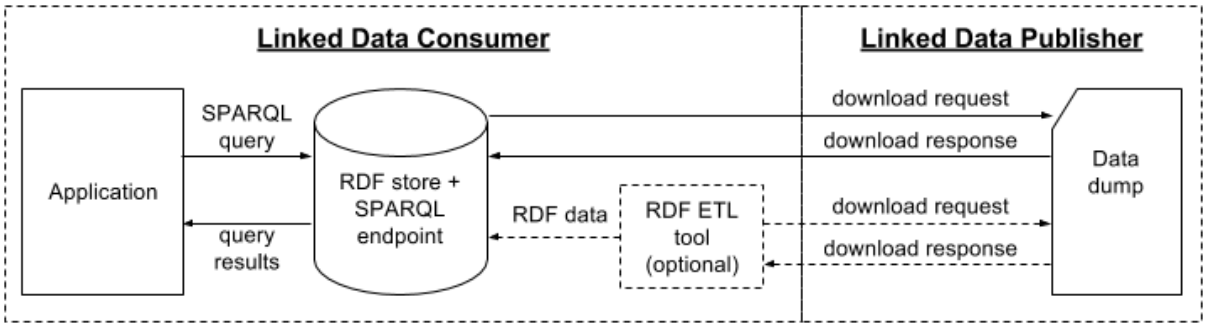

Figure 2: Data dumps

In the TPF with HDT files scenario (Figure 3) LD Publisher makes LD queryable using a TPF server and LD Consumer access this data using a TPF client.

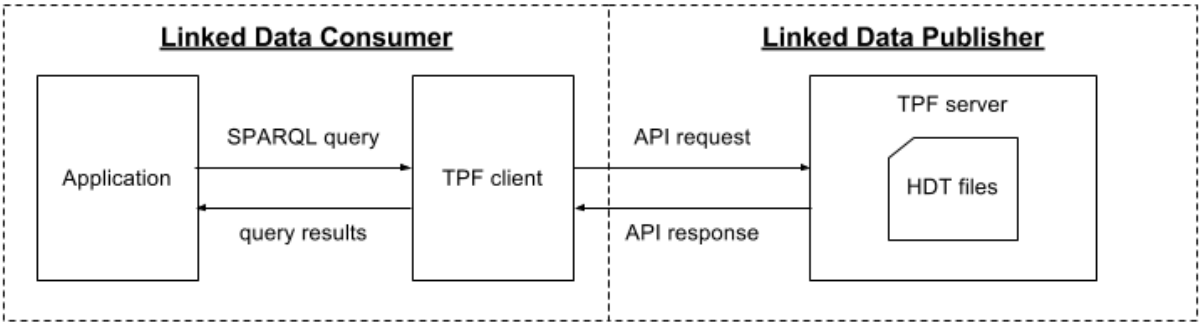

Figure 3: TPF with HDT files

In the RDF store with a SPARQL endpoint scenario (Figure 4) LD Publisher makes data accessible via a SPARQL endpoint and LD Consumer can query this endpoint directly from an application.

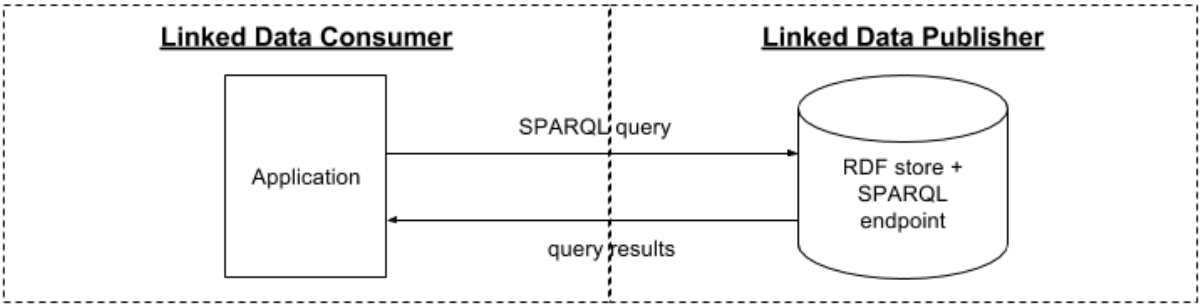

Figure 4: RDF store with a SPARQL endpoint

Required competencies for querying Linked Data

Table 1 summarizes competencies directly required to publish and query LD based on the components of the LD querying scenarios. We assumed that LD Publisher would be able to create RDF data. Likewise, LD Consumer might perform various other tasks besides querying LD such as data cleansing or enrichment, but competencies required for these tasks were not taken into account.

Differences in the required competencies result from the different nature of the interfaces, e.g., publishing downloadable files is a common competency non-specific to querying LD, whereas publishing data using a SPARQL endpoint or TPF server requires competencies to install and maintain the respective software components.

The costs of developing the necessary competencies would depend on the available learning materials, courses, or on the selected HW and SW components and the associated services provided by their vendors. According to [2] people without prior knowledge of LD sometimes find the existing tutorials difficult to understand, which contributes to the perceived lack of the LD-related competencies.

Table 1: Competencies required to publish and query Linked Data

Subject	Data dumps	TPF with HDT files	RDF store with a SPARQL endpoint
LD Publisher	Ability to publish downloadable files	Ability to create HDT files Ability to install and maintain a TPF server	Ability to install and maintain a RDF store and a SPARQL endpoint Ability to load RDF data into an RDF store
LD Consumer	SPARQL skills Ability to install and maintain an RDF store and a SPARQL endpoint Ability to load RDF data into a RDF store Ability to develop/ implement and maintain an RDF ETL tool (optional) Ability to develop/ implement and maintain application querying RDF data	SPARQL skills Ability to develop/ implement and maintain a TPF client Ability to develop/ implement and maintain application querying RDF data	SPARQL skills Ability to develop/ implement and maintain application querying RDF data

Legacy case studies: first touch on publisher and consumer competence requirement analysis

Two out of the twenty selected case studies turned out to be irrelevant since they described a platform rather than a consumption of some datasets. The remaining studies are listed in Table 2. The studies are identified with their acronym and can be accessed at the URL that conforms to the pattern: `https://www.w3.org/2001/sw/sweo/public/UseCases/{acronym}`.

We distinguished between cases where the same subject acted as LD Publisher and Consumer and cases where different subjects acted in these roles ("P=C" column). Sometimes both situations were present in a single case. We classified the involved LD Publishers, Consumers and End Users. The typology of the subjects was not normalized but we tried to classify the subjects as described in the case studies. Finally, we marked whether End User is affiliated with LD Consumer or not.

Despite the selected case studies referred to SPARQL at the SWEO portal, in five cases the LD publishing method was unobvious. In another five cases SPARQL was mentioned but it was not clear how the data was actually queried. SPARQL endpoint was mentioned in three cases. Other mentioned publishing methods included content negotiation, machine-readable feeds, web service or use of RDFa.

In seven cases distinct subjects acted as LD Publisher and Consumer. Out of this group Kisti, ZBW and Faviki represent a constellation with no overlap between these roles. In this constellation subjects can focus on developing the required competencies based on their role and type of the LD publishing interface.

In case of iLaw, LD Publisher first consumed LD internally but subsequently started to provide data to distinct LD Consumers as well. In such a constellation the subject combining roles should develop a wider range of competencies, but this would allow it to benefit from its own published data.

In the Europeana and BBC cases, subjects are also combining the roles of LD Publisher and Consumer, but they enrich their own data with data published by a distinct LD Publisher. Therefore developing of the competencies partly depends on interfaces through which they consume the external data.

Finally, the Volkswagen case represents a constellation with mutual LD Publisher-Consumer relationships. The first subject publishes data in RDF that are used by the second subject to add semantics to some primary data. This enriched data is subsequently consumed by the first subject. Both subjects need to develop the required competencies.

Table 2: Characteristics of the analysed case studies

Acronym	Year	P=C	LD Publisher type	LD Consumer type	End user type	End user affiliation
NRK	2007	yes	broadcasting company	x – same as LD publisher	journalists, program makers	internal
Talis	2009	yes	university	x	instructors, students	internal
IOInformatics	2011	yes	consortium including hospital, university and research centre	x	researchers, clinicians	internal
Volkswagen	2011	both	car producer, used car seller	car producer, used car seller	web portal users	both
Zaragoza-2	2008	yes	local government	x	tourists	external

Acronym	Year	P=C	LD Publisher type	LD Consumer type	End user type	End user affiliation
EDF	2008	yes	energy company	x	employees	internal
Europeana	2012	both	non-profit foundation	unspecified	unspecified	external
iLaw	2010	both	ministry	government ministries and agencies	legal experts	both
Kisti	2010	no	government agency	public, businesses	unspecified	external
Aquaring	2009	yes	unspecified	x	unspecified users of the portal	external
Nasa	2008	yes	public institution	x	staffers, workforce planners, analysts, related personnel	internal
Lilly	2007	yes	pharmaceutical company	x	researchers, scientists	internal
ZBW	2009	no	library	other people and institutions	other people and institutions	external
Faviki	2008	no	community project	IT company	unspecified users of the tagging service	external
UniZheijang	2007	yes	academic institution	x	unspecified users	external
Twine	2009	yes	IT company	x	unspecified users of the tagging service	external
BBC	2010	both	broadcasting company, other subjects	broadcasting company	broadcasting company teams	internal
OntoFrame	2009	yes	public institution	x	researchers	external

Conclusions

LD could be published using various interfaces. The type of the interface affects the costs incurred by the subjects publishing and querying data. Not only technologies but also competencies are necessary for querying LD and therefore developing these competencies could be a source of costs. We analysed three LD querying scenarios from the perspective of the necessary competencies, and investigated constellations of LD Publishers and Consumers based on a subset of the legacy SWEO case studies.

Learning materials could play a significant role in developing the necessary competencies. Future research will therefore address the question of availability and quality of the learning materials. We will also proceed to other collections of case studies to collect a wider set of LD Publisher-Consumer constellations and we will study the competencies for querying LD in more detail.

Acknowledgements

This research has been supported by the H2020 project no. 645833 (OpenBudgets.eu) with contribution of the long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

1. Verborgh, R. et al.: Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web. *Journal of Web Semantics*, vol. 37–38, pp. 184–206 (2016)
2. Archer, P. et al.: *Study on business models for Linked Open Government Data* (2013), https://joinup.ec.europa.eu/sites/default/files/85/31/25/Study_on_business_models_for_Linked_Open_Government_Data_BM4LOGD_v1.00.pdf
3. Maizlish, B., Handler, R.: *IT Portfolio Management Step-by-Step*. John Wiley & Sons, Hoboken, New Jersey (2005)
4. Latif, A. et al.: The Linked Data Value Chain: A Lightweight Model for Business Engineers. In: *SEMANTiCS 2009*, pp. 568-575
5. Fernández, J.D. et al.: Binary RDF representation for publication and exchange (HDT). *Journal of Web Semantics*, vol. 19, pp. 22-41 (2013)