Umaka-Yummy Data: A Place to Facilitate Communication between Data Providers and Consumers

Yasunori Yamamoto[1], Atsuko Yamaguchi[1], and Andrea Splendiani[2]

1. Database Center for Life Science, 2. A BioHackathon Participant
{yy,atsuko}@dbcls.rois.ac.jp, andrea@sgtp.net

Keywords: RDF, SPARQL, Linked Data

Abstract. A consequence of the increasing amount of information available in RDF is that it is getting harder, for users, to find which sources (often SPARQL endpoint) are the most appropriate, reliable and up to date for some sought information. Here we introduce YummyData, a service that monitors and assess the "quality" of endpoints providing datasets of interest to the biomedical community. It helps biomedical researchers in two ways: by providing a curated list of endpoints and by enriching it with information on their availability, updates rate, standard compliance, and other features that are relevant to users. Since we believe this assessment is valuable for both researchers or consumers and providers of biomedical RDF data, YummyData provides a forum where they can communicate and improve the usability of the web of (bio) data.

1 Introduction

Major life science databases such as UniProt[1], Ensembl[2], MeSH[3], or PubChem[4] are offering their datasets as RDF, setting a trend that is followed by other life science institutions. RDF enables researchers to simply assemble independently developed datasets. However, researchers need to know first where the data they need reside, and second how much these data is reliable and useful. The former is an issue of findability. Standard ways to describe datasets have been developed to address this issue, for instance VoID[5] and Service Descriptions[6] (SD) standards, but their usage to annotate SPARQL endpoints is still limited.

The issue of assessing the usefulness of a dataset is more complex, as it relates to the richness of the data provided and its overall quality. Data quality has many dimensions, but regarding the provision of information in RDF, we can identify two main aspects: one is the quality of endpoints, that is, how long an endpoint is up and running or whether an endpoint provides the above mentioned metadata or not. The second is the quality of provided datasets themselves, that is, how much a dataset uses well-defined ontologies or vocabularies or how well a dataset follows the Linked Data principles[7].

1 http://www.uniprot.org/downloads
2 https://www.ebi.ac.uk/rdf/services
3 https://id.nlm.nih.gov/mesh/
4 https://pubchem.ncbi.nlm.nih.gov/rdf/
5 http://www.w3.org/TR/void/
6 http://www.w3.org/TR/sparql11-service-description/
7 https://www.w3.org/DesignIssues/LinkedData.html

adfa, p. 1, 2011.

To help users in assessing aspects of the quality of a "dataset" (as provided over a SPARQL endpoint and/or following linked-data principles) we have developed a service called Umaka-Yummy Data or YummyData for short (umaka means yummy in Japanese). YummyData has two main components. One is a data crawler, and the other is a website to summarize collected data and to provide a discussion space on issues found in endpoints. The crawler periodically accesses a curated list of endpoints and issues a series of SPARQL queries to inspect a variety of features. The website shows Umaka Score for each endpoint. The score is described in the following section. In addition, YummyData provides a forum for each endpoint to facilitate communications between its developer and consumers.

We introduce the Umaka Score to facilitate comparison among endpoints. The score relates to six aspects of "data provision quality": *availability*, *freshness*, *operation*, *usefulness*, *validity*, and *performance*. The *availability* of an endpoint is defined as the ratio of the number of alive days over 30 (a month). The *freshness* of an endpoint is defined as how often its datasets are updated. The Umaka crawler looks up the property value of dcterms:modified, which is assumed to be in its SD or VoID data. However, only a few endpoints provide this value. When missing, the crawler issues SPARQL queries to see whether some randomly obtained data has been changed since the last access or not. The "*operation* score" of an endpoint relates to whether it adheres to standards that facilitate its consumption (e.g.: whether the endpoint provides SD or VoID or not). *Usefulness* is defined as how easy other datasets can link to or can reuse the dataset provided by an endpoint. The crawler issues SPARQL queries to measure how much each instance is typed (*i.e.*, whether it has the rdf:type property) or labeled (*i.e.*, whether it has the rdfs:label property). In addition, we check the extent to which datasets of an endpoint contains shared vocabularies that are stored in Linked Open Vocabularies[8] (LOV) or other datasets that YummyData collects. *Validity* is defined as whether an endpoint follows the four Linked Data principles. Finally, *performance* is defined how fast an endpoint returns a result. The Umaka score for an endpoint is calculated by aggregating sub-scores for each of the above aspects.

2 Discussions and Conclusion

The Umaka Score is a measure which is under development, and some aspects of scoring methods are rather ad-hoc. We intend to engage with both consumers and providers to evolve this score as to better reflect the "quality" of an endpoint to users. Our goal is to make RDF datasets “tastier,” which means that we want to contribute to making more datasets easily findable and usable. To this end, we provide a forum where providers and consumers communicate with each other. We believe that by publishing a measure on the standard compliance and overall quality of SPARQL endpoints, we can trigger a dialogue between consumers and providers that will be mutually beneficial. We are making our service as transparent as possible to gain reliability of our service and attract more users.

Acknowledgements. This work was supported by the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST).

[8] http://lov.okfn.org/dataset/lov