Umaka-Yummy Data: Where providers and consumers in life sciences communicate with each other

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: Linked Data, RDF, SPARQL

Abstract. RDF datasets and SPARQL endpoints hosting them are increasingly available in the life science domain. One consequence is that users have increasing difficulties in finding those endpoints which are reliable and stable. For providers, it is not an easy task to provide appropriate metadata to be easily findable. To solve these issues, we propose a service to provide a place where data providers and data consumers can communicate with each other. Here we introduce YummyData, a service that monitors SPARQL endpoints that provide data of interest to the life science community. YummyData helps by providing a curated list of endpoints, and by monitoring over time their availability, updates rate, standard compliance, and other features that are important to data consumers. As we believe these data are valuable for both providers and consumers, YummyData provides a forum where providers and consumers of life science information in RDF can communicate and improve the usability of the web of life science data. You can freely access YummyData at http://yummydata.org/ .

1 Introduction

RDF datasets are being provided by major life science databases such as UniProt[1], PubChem[2], MeSH[3], or Ensembl[4]. In addition, other life science institutions have also started to release datasets in RDF. Such datasets are advantageous in that it is easy to integrate them, even when independently developed. But nevertheless users first need to know where their designated datasets reside, and second how much these data are reliable or useful. The former is the issue of findability and there are widely known vocabularies to be used for it. VoID[5] and Service Descriptions[6] (SD) are suited for SPARQL endpoints to provide data for findability, but only a few endpoints makes use of them. Furthermore, the richness of these data provided by endpoints differs

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

depending on the endpoints. As for the latter, it's an issue of data quality. A data quality here is mainly classified into two parts. The first 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].

To address these issues we have developed a service called Umaka-Yummy Data or YummyData for short[8]. YummyData has two main components. One is a data crawler, and the other is a website to present such data and to provide discussion forums. The crawler periodically accesses each of the curated list of endpoints and issues a series of SPARQL queries to measure different aspects related to "quality" and availability. The website shows the collected information and a synthetic Umaka Score for each end-point. The score is described in the following section. In addition, YummyData provides a forum for each endpoint to facilitate communications between its developers and consumers.

2 Umaka Score

We introduce the Umaka Score to facilitate comparison among endpoints and provoke communications between data providers and data consumers. This score consists of six aspects: availability, freshness, operation, usefulness, validity, and performance. An availability sub-score for an endpoint is defined as the ratio of the number of days it is alive in the last 30 days (a month). A freshness sub-score (for an endpoint) is defined as how often its contained datasets are updated. The Umaka crawler first looks up the property value of dcterms:modified assumed to be contained its SD and VoID data. However, as mentioned above, only a few endpoints have this value, and therefore the crawler issues SPARQL queries to see if something has been changed since the last access. An "operation" sub-score captures whether an endpoint provides SD or VoID. "Usefulness" is measured by how easily other datasets link to or use the dataset provided by an endpoint. The crawler issues SPARQL queries to obtain 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 how many datasets of an endpoint use shared vocabularies, which are defined in Linked Open Vocabularies[9] (LOV) or other datasets that YummyData collects. A validity sub-score reflects whether an endpoint follows the four Linked Data principles. "Performance" is measured by how fast an endpoint returns a result. The overall "Umaka score" of an endpoint is calculated by summing up those all of these aspects.

---

7 https://www.w3.org/DesignIssues/LinkedData.html

8 "Umaka" means yummy in Japanese.

9 http://lov.okfn.org/dataset/lov

3 Current Status

As of November 1, 2016, metadata of 78 endpoints that we specified have been collected. The result is updated every day, and you can check the obtained data in detail of each endpoint. In addition to the latest data, you can also directly access to ones of a specific day by clicking a point on the history graph at the top page or the page for an endpoint. Figure 1 shows the top page of YummyData, and Figure 2 shows a page for an endpoint as an example. By checking a history graph, you can learn how stable or how reliable an endpoint is. Moreover, there are links at the page for each endpoint to a page showing the issued queries and their responses. These data help both data providers and data consumers to know what is happened at an endpoint. A data provider may discover some new problems. A data consumer can learn differences among closely related endpoints from multiple aspects. To facilitate communications between data providers and data consumers, at a page for each endpoint, there is a link to its corresponding discussion forum at a GitHub issues page.

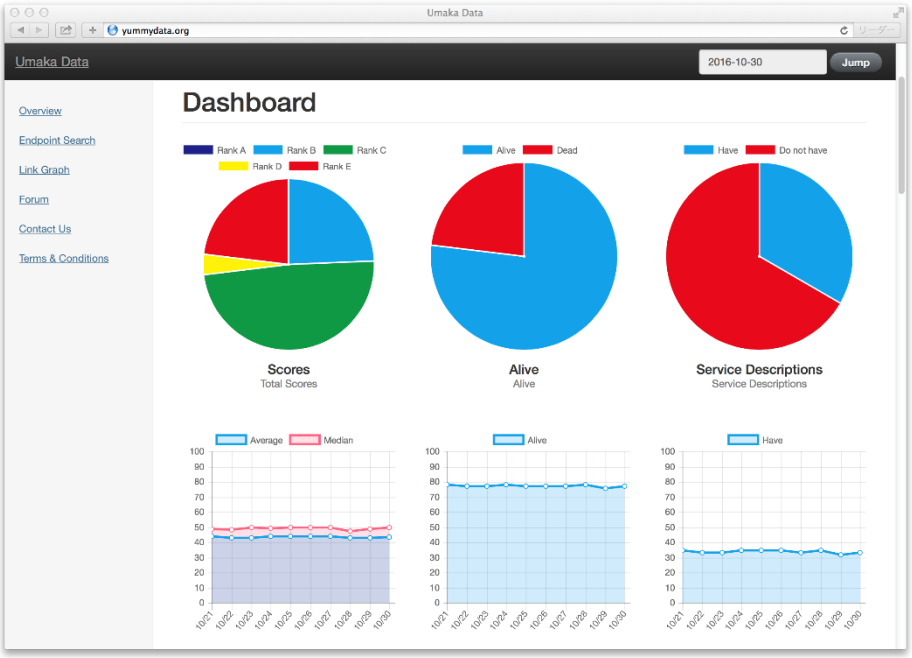

Fig. 1. Umaka-Yummy top page.

4 Discussions and Conclusion

Umaka Score is a measure that is under development, and the aspects and scoring methods are rather ad-hoc. We continue to discuss and update them to make the score more effective. Our goal is to make RDF datasets more “tasty,” which means that we

want to contribute to make more datasets easily findable and used. To that end, we provide a forum where data providers and data consumers can communicate with each other along with several measures for each endpoint that can be a trigger to begin a discussion. We are making our service as transparent and reliable as possible to build a consistent user community.

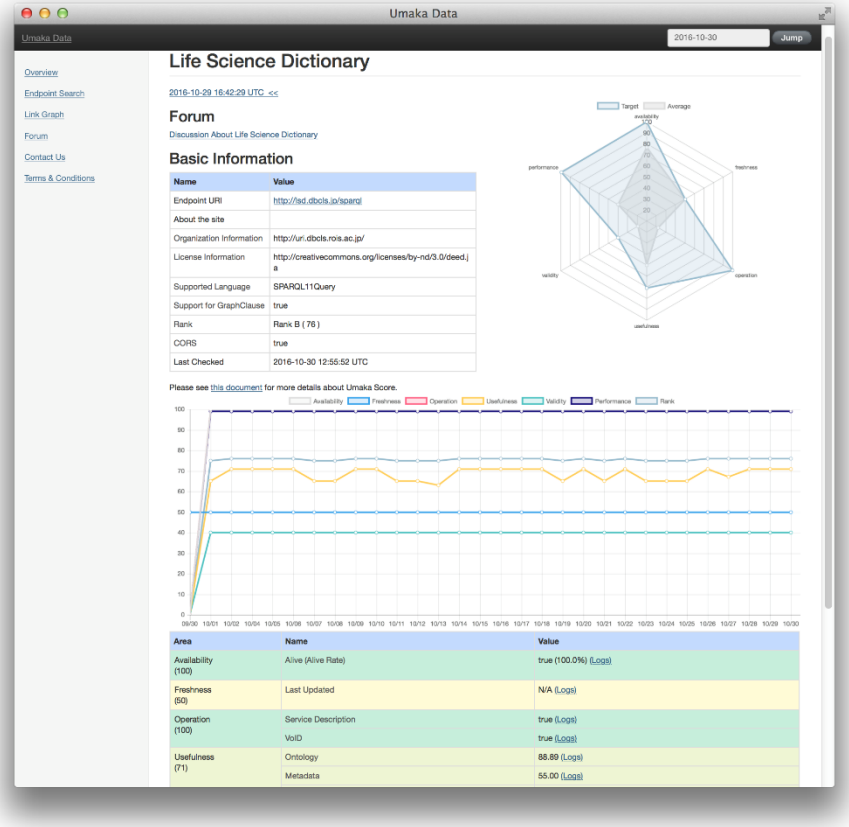

Fig. 2. An example of obtained data for an endpoint.

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