The YummyData Initiative: How SPARQL-y is your biomedical endpoint?

Ivar Andrea Splendiani[1,2] and Johan Nystrom-Persson[3] and Michel Dumontier[4] and Yasunori Yamamoty[5]

[1] DERI, Galway, IE andrea.splendiani@deri.org
[2] intelliLeaf ltd, Cambridge, UK andrea.splendiani@intellilaf.com
[3] The National Institute of Biomedical Innovation, Osaka, Japan johan@nibio.go.jp
[4] Carleton University, Ottawa, Canada michel_dumontier@carleton.ca
[5] Database Center for Life Science, JP yy@dbcls.jp

Abstract. Although increasing amounts of biomedical data is being provided as structured content on the Semantic Web, there is currently no standardized way to monitor SPARQL endpoints for their availability, reliability or content flux. Importantly, there are additional issues relating to the provision of version-sensitive data republished by third parties or made available as part of a one off research project. All of these aspects have important consequences for users that rely on federated queries across distributed SPARQL endpoints.

1 Results

We describe the YummyData initiative to provide monitoring of biomedical SPARQL endpoints on a variety of factors including availability, reliability, content summarization, and content evolution. Our prototype website, yummydata.org, provides simple metrics relating to the availability and response status of a selected set of endpoints, the size of the data set, etc. In addition to these fundamental metrics, our long term goal is to compute a SPARKLE score, a composite metric combining measures such as the number of triples, size and frequency of updates, the number of links to other datasets, and the capabilities of the endpoint server. Although the SPARKLE score is not a measure of the quality of a dataset, it helps indicate whether the dataset changes over time, and whether these changes are likely to be negative or positive in nature. These statistics may possibly correlated with the declared update frequency of the datasets published by the endpoint of a given provider, thus providing an additional input to our score. Finally, yummydata.org also supports custom queries that track endpoint/data metrics, allowing new metrics to be computed and shared amongst participants. Although devising an objective measure of quality is controversial, we believe that the YummyData initiative will help users better understand the content that is currently available while also helping providers understand what kinds of metrics are important to users. We also believe that emergence of more and more third party SPARQL endpoint rating services like YummyData will bring about an environment where we can get more objective

evaluation of endpoints, and therefore qualities of the entire RDF-based biomedical data/services are becoming higher. As yummydata.org is an early prototype, we welcome suggestions to clarify existing metrics and to help develop additional metrics to tease out information of interest.

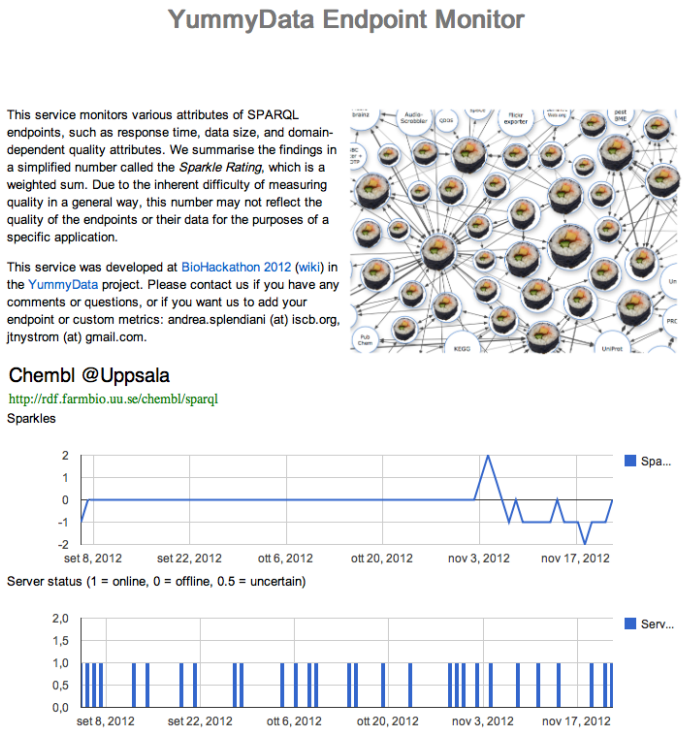

Fig. 1. A screenshot from YummyData.org, showing the evolution of parameters and score for a given endpoint

2 Acknowledgments

We wish to thanks the organizers of the BioHackathon 2012, during which this work originated.