Big Data Quality Assessment

PI: Michel Dumontier, Postdoc: Amrapali Zaveri

The development and standardization of Semantic Web technologies has resulted in an unprecedented volume of data being published openly on the Web, specifically in the Life Science domain. However, we observe varying quality of the data ranging from extensively curated datasets to extracted data of relatively low quality. A survey unified and formalized commonly used terminologies across 30 core data quality assessment approaches and provided a comprehensive list of 18 quality dimensions and 69 metrics. Additionally, a set of metrics have been recently published to ensure that open data is maximally Findable, Accessible, Interoperable and Reusable. However, there is no standardized tool that implements all of these metrics that can be used to assess the quality of a dataset, specifically in the Life Science domain. Moreover, there is a need for not only assessing but also improving the quality by reporting on the root case of the quality issue. The thesis will be focused towards development and implementation of metrics and a tool for big data quality assessment as well as improvement, specifically for Life Science Data.

Publications:

1. Quality assessment for linked data: A survey. **A Zaveri** , A Rula, A Maurino, R Pietrobon, J Lehmann, S Auer. Semantic Web 7 (1), 63-93. **Cited by 214.**
2. Test-driven evaluation of linked data quality. Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen, **Amrapali Zaveri** . Proceedings of the 23rd international conference on World Wide Web, 747-758. **Cited by 138.**
3. The FAIR Guiding Principles for scientific data management and stewardship, Mark D Wilkinson, **Michel Dumontier,** et al. Scientific data, 2016. **Cited by 237.**
4. Crowdsourcing linked data quality assessment. M Acosta, **A Zaveri** , E Simperl, D Kontokostas, S Auer, J Lehmann International Semantic Web Conference, 260-276. **Cited by 81.**
5. Bio2RDF release 2: improved coverage, interoperability and provenance of life science linked data A Callahan, J Cruz-Toledo, P Ansell, **M Dumontier.** Extended Semantic Web Conference, 200-212. **Cited by 70.**