# The State of Linked Data

## An Extensible Framework to Asses and Build Dataset Profiles

Ahmad Assaf[1][2], Aline Senart[2] and Raphaël Troncy[1]

[1] EURECOM, Sophia Antipolis, France. `<firstName.lastName@eurecom.fr>`
[2] SAP Labs France. `<firstName.lastName@sap.com>`

**Abstract.** Linked Open Data (LOD) has emerged as one of the largest collection of interlinked datasets on the web. Benefiting from this mine of data requires the existence of descriptive information about each dataset in the accompanying metadata. Such meta information is currently very limited to few Data Portals where they are provided manually thus giving little or bad quality insights. To address this issue, we propose a scalable automatic approach for extracting and generating descriptive linked dataset meta information. This approach apply several techniques to check the validity of the attached metadata of a certain dataset as well as a whole Data Portal. Using our framework on prominent Data Portals shows that the general state of the Linked Open Data needs attention as most of datasets suffer from ad quality metadata and lack additional informative metrics.

**Keywords:** Linked Data, Dataset Profile, Metadata, Data Quality

## 1 Introduction

In the last few years the Semantic Web gained a momentum supported by the introduction of many related initiatives like the Linked Open Data (LOD)[1]. From 12 datasets cataloged in 2007, the Linked Open Data has grown to almost 1000 datasets containing almost 82 billion triples[3]. Data is being published by both public and private sectors and covers a diverse set of domains from life sciences to military.

The Linked Open Data is a gold mine for organizations and individuals who are trying to leverage external data sources in order to produce more informed business decisions [4]. This success lies in the cooperation between data publishers and consumers. Users are empowered to find, share and combine information in their applications easily. However, the heterogeneous nature of data sources reflects directly on the data quality as these sources often contain inconsistent as well as misinterpreted and incomplete information and meta information. Accompanied with the significant variation of size, used languages and freshness, finding useful datasets without prior knowledge is increasingly complicated. This

[3] http://datahub.io/dataset?tags=lod

can be clearly noticed in the LOD Cloud [4] as few datasets like DBPedia[2], Freebase[3] and YAGO[7] are favored over hidden gems that may include domain specific knowledge more suitable for the tasks on hand.

The main entry point for discovering and identifying needed datasets is through public Data Portals like DataHub[5] and Europe's Public Data[6] or private ones like Quandl[7] and Engima[8]. Private portals harness manually curated data from various sources and expose them to users either freely or through paid plans. The data available is of higher quality but lesser quantity compared to what is available in public portals. Similarly in some public Data Portals, administrators manually review datasets information and attach suitable meta information. This information is mainly in form of predefined tags such as *media, geography, life sciences, etc.* for organization and clustering purposes. The increasing number of available datasets makes the review and curation process unsustainable even when outsourced to communities. Furthermore, the diversity of those datasets makes it hard to classify them with a fixed number of predefined tags that can be subjectively assigned without capturing the essence and breadth of the dataset[5].

*Data profiling* is the process of creating descriptive dataset metadata. It is a cardinal activity when facing an unfamiliar dataset[6].It helps in assessing the importance of the dataset, improve users' ability to search and reuse part of the dataset and detect irregularities to improve its quality. Data profiling includes several tasks:

- **Metadata profiling**: Provide several information about the dataset. This can include general information (dataset description, release and update dates, etc.), legal information (license information, openness, etc.), practical information (access points, data dumps, etc.) and so on.
- **Statistical profiling**: Provides statistical information about data types and patterns in the dataset. i.e. properties distribution, number of entities and RDF triples, etc.
- **Topical profiling**: Provides descriptive and reliable knowledge on the dataset's content and structure. This can be in form of tags and categories used to facilitate search and reuse of existing datasets.

In this work, we address the above mentioned challenges of automatic assessment and generation of descriptive datasets profiles. This paper proposes an extensible framework consisting of a processing pipeline that combines techniques for Data Portals identification, datasets crawling and a set of pluggable modules combining several profiling tasks. The framework assesses the provided dataset metadata against an aggregated standard set of information. Metadata fields are automatically corrected when possible i.e. missing license link. Moreover, a report is created with issues that cannot be automatically fixed and is

[4] http://lod-cloud.net
[5] http://datahub.io
[6] http://publicdata.eu
[7] https://quandl.com/
[8] http://enigma.io/

sent to the dataset's maintainer via e-mail. There exist various statistical and topical profiling tools for both relational and Linked Data. The architecture of the framework allows to easily add several profiling tasks. For this paper, we will focus on Linked Data profiling tools as we will present our findings on the overall state of Linked Data in some of the prominent Data Portals.

The remainder of the paper is structured as follows. Section 2 reviews related literature. Section 3 describes the framework's architecture to assess and generate dataset profiles. Section 4 shows the results of running this tool on some of the most prominent Data Portals and their discussion. Finally, Section 5 presents the conclusion and future work.

## References

1. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
2. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 7(3):154–165, Sept. 2009.
3. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, 2008.
4. D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, pages 1–17, 2011.
5. S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain. Automatic domain identification for linked open data. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, volume 1, pages 205–212, Nov 2013.
6. H. Li. Data profiling for semantic web data. In F. Wang, J. Lei, Z. Gong, and X. Luo, editors, *Web Information Systems and Mining*, volume 7529 of *Lecture Notes in Computer Science*, pages 472–479. Springer Berlin Heidelberg, 2012.
7. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, 2007.