An Extensible Framework to Validate 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 usually provided manually thus giving little or bad quality insights. To address this issue, we propose a scalable automatic approach for extracting, validating and generating descriptive linked dataset profiles. This approach applies several techniques to check the validity of the attached metadata as well as providing descriptive and statistical information 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 bad 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) [5]. 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 [11]. This success lies in the cooperation between data publishers and consumers. Consumers 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 in size, used languages and freshness, finding useful datasets without prior knowledge is increasingly

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

complicated. This can be clearly noticed in the LOD Cloud [4] as few datasets like DBPedia [6], Freebase [10] and YAGO [38] 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. However 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 [30].

Data profiling is the process of creating descriptive information about the data examined. It is a cardinal activity when facing an unfamiliar dataset [33].It helps in assessing the importance of the dataset, improving users' ability to search and reuse part of the dataset and detect irregularities to improve its quality. Data profiling typically includes several tasks:

- **Metadata profiling**: Provides general information on the dataset (dataset description, release and update dates, etc.), legal information (license information, openness, etc.), practical information (access points, data dumps, etc.), etc.
- **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 knowledge on the dataset content and structure. This can be in form of tags and categories used to facilitate search and reuse.

In this work, we address the above mentioned challenges of automatic validation 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 validates 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 them as additional 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 validate and generate dataset profiles. Section 4 shows the results of running this tool on some of the most prominent data portals and analyzes them. Finally, Section 5 presents the conclusion and future work.

2 Related Work

There exists a considerable amount of tools that tackle specific profiling tasks. However, to the best of our knowledge, this is the first effort towards extensible automatic assessment and generation of descriptive dataset profiles. For this paper, we will focus on Linked Data profiling tasks. However, one of the advantages of this framework is the ability to easily configure additional profiling tasks accommodating different data types e.g. relational.

Metadata profiling: Data Catalog Vocabulary (DCAT) [19] and the Vocabulary of Interlinked Datasets (VoID) [15] are concerned with metadata about RDF datasets. There exists several tools aiming at exposing dataset metadata using these vocabularies. In [8] authors generate VoID descriptions limited to a subset of properties that can be automatically deduced from resources within the dataset. However, it still provides data consumers with interesting insights. Quality Assessment of Data Sources (Flemming's Data Quality Assessment Tool)[9] provides basic metadata assessment as it calculates data quality scores based on manual user input. The user should assign weights to the predefined quality metrics and answer a series of questions regarding the dataset. These include, for example, the use of obsolete classes and properties by defining the number of described entities that are assigned disjoint classes, the usage of stable URIs and whether the publisher provides a mailing list for the dataset. The ODI certificate[10] on the other hand provides a description of the published data quality in plain English. It aspires to act as a mark of approval that helps publishers understand how to publish good open data and users how to use it. It gives publishers the ability to provide assurance and support on their data while encouraging further improvements through an ascending scale. ODI comes as an online and free questionnaire for data publishers focusing on certain characteristics about their data. Although these approaches try to perform metadata profiling, they are either incomplete or manual. In our framework, we propose a more automatized and complete approach.

[9] http://linkeddata.informatik.hu-berlin.de/LDSrcAss/datenquelle.php
[10] https://certificates.theodi.org/

Statistical profiling: Calculating statistical information on datasets is vital to applications dealing with query optimization and answering, data cleansing, schema induction and data mining [26] [21] [30]. Semantic sitemaps [14] and RDFStats [31] where one of the first to deal with RDF data statistics and summaries. ExpLOD [27] creates statistics on the interlinking between datasets based on `owl:sameAs` links. In [33] the author introduces a tool that induces the actual schema of the data and gather corresponding statistics accordingly. LODStats [3] is a stream-based approach that calculates more general dataset statistics. ProLOD++ [1] is a Web-based tool that allows LOD analysis via automatically computed hierarchical clustering [9]. Aether [35] generates VoID statistical descriptions of RDF datasets. It also provides a Web interface to view and compare VoID descriptions. LODOP [4] is a MapReduce framework to compute, optimize and benchmark dataset profiles. The main target for this framework is to optimize the runtime costs for Linked Data profiling. In [29] authors calculate certain statistical information for the purpose of observing the dynamic changes in datasets.

Topical Profiling: Topical and categorical information facilitates dataset search and reuse. Topical profiling focuses on content-wise analysis at the instances and ontological levels. GERBIL [40] is a general entity annotation framework that provides machine processable output allowing efficient querying. In addition, there exist several entity annotation tools and frameworks [13] but none of those systems are designed specifically for dataset annotation. In [23], authors created a semantic portal to manually annotate and publish metadata about both LOD and non-RDF datasets. In [30], authors automatically assigned Freebase domains to extracted instance labels of some of the LOD Cloud datasets. The goal was to provide automatic domain identification, thus enabling improving datasets clustering and categorization. In [7], authors extracted dataset topics by exploiting the graph structure and ontological information, thus removing the dependency on textual labels. In [20] authors generate VoID and VoL descriptions via a processing pipeline that extracts dataset topic models ranked on graphical models of selected DBpedia categories.

Although the above mentioned tools are able to provide general, statistical and topical information about a dataset, there exist no approach that is able to combine them into a unified profile presented to consumers. Dataset search can be done without relying on attached metadata (tags and categories). For example, there exist several approaches to create LOD indexes. In [2], authors used VoID descriptions to optimize query processing by determining relevant queryable datasets. In [24], authors created an approximate index structure (QTree) and an algorithm for answering conjunctive queries over Linked Data. SchemEX [28] is a stream-based approach leveraging type and property information of RDF instances to create schema-level indexes.
Semantic search engines like Sindice [17], Swoogle [18] and Watson [16] help in entities lookup but are not designed specifically for dataset search. In [36],

authors utilized the sig.ma index [39] to identify appropriate data sources for interlinking.

3 Profiling Data Portals

In this section, we provide an overview of the processing steps for validating and generating dataset profiles. Figure 1 shows the main steps which are the following: (i) Data portal identification; (ii) metadata extraction; (iii) instance and resource extraction; (iv) profile validation (v) profiler and report generation.

Our framework is built as a Command Line Interface (CLI) application using Node.js. Related functions are encapsulated into modules that can be easily plugged in/out the processing pipeline. The various steps are explained in details below.

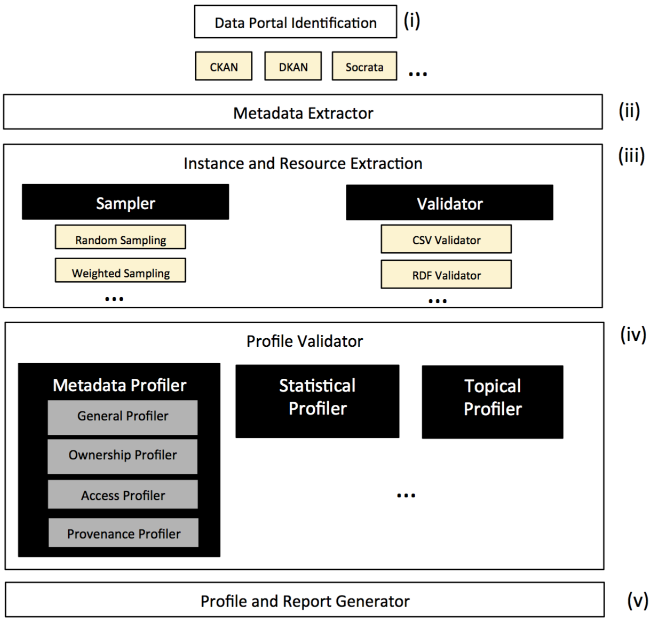

Fig. 1. Processing pipeline for validating and generating dataset profiles

3.1 Data Portal Identification

Data portals can be considered as data access points providing tools to facilitate data publishing, sharing, searching and visualization. CKAN[11] is the world's leading open-source data portal platform powering websites like the DataHub, Europe's Public Data and the U.S Government's open data. Modeled on CKAN, DKAN[12] is a standalone Drupal distribution that is used in various public data portals as well. Socrata[13] helps public sector organizations improve data-driven decision making by providing a set of solutions including an open data portal. In addition to these tradition data portals, there is a set of tools that allow exposing data directly as RESTful APIs like Datatank[14], Database-to-API[15] and CSV-to-API[16].

Identifying the software powering data portals is a vital first step to understand the API calls structure. Web scraping is a technique for extracting data from Web pages. We rely on several scraping techniques in the identification process which includes a combination of the following:

– **URL inspection**: Check the existence of certain URL patterns. Various CKAN based portals are hosted on subdomains of the `http://ckan.net`. For example, CKAN Brazil (`http://br.ckan.net`).
– **Meta tags inspection**: The `<meta>` tag provides metadata about the HTML document. They are typically used to specify page description, keywords, author, etc. Inspecting the `content` attribute can indicate the type of the data portal. We use CSS selectors to check the existence of these meta tags. An example of a query selector is `meta[content*="ckan]` (all meta tags with the attribute content containing the string *CKAN*). This selector can identify CKAN portals whereas the `meta[content*="Drupal"]` can identify DKAN portals.
– **Document Object Model (DOM) inspection**: Similar to the meta tags inspection, we check the existence of certain DOM elements or properties. For example, CKAN powered portals will have DOM elements with class names like `ckan-icon` or `ckan-footer-logo`. A CSS selector like `.ckan-icon` will be able to check if a DOM element with the class name `ckan-icon` exists.
The list of elements and properties to inspect is stored in a separate configurable object for each portal. This allows the addition and removal of elements as deemed necessary.

The identification process for each portal can be easily customized by overriding the default function. Moreover, adding or removing steps from the identification process can be easily configured.

[11] http://ckan.org
[12] http://drupal.org/project/dkan
[13] http://www.socrata.com
[14] http://thedatatank.com
[15] https://github.com/project-open-data/db-to-api
[16] https://github.com/project-open-data/csv-to-api

After those preliminary checks, we try to query one of the portal's API endpoints. For example, DataHub is identified as CKAN, so we will query the API endpoint on `http://datahub.io/api/action/package_list`. A successful request will list the names of the site's datasets, whereas a failing request will signal a possible failure of the identification process.

3.2 Metadata Extraction

Data portals expose a set of information about each dataset as metadata. The model used varies across portals. However, a standard model should contain information about the dataset's title, description, maintainer email, update and creation date, etc. Building a standard metadata model is not the scope of this paper, and since we focus on CKAN-based portals, thus we validate the extracted metadata against the CKAN standard model[17]. The standard CKAN model contains the following information:

General information: General information about the dataset. e.g. title, description, ID, etc. This general information is manually filled by the dataset owner. It requires manual correction, thus any missing values like the title or description will not be automatically corrected but will appear in the final report.

Access information: Information about the access and usage of the dataset. e.g. license information, license URL, dataset URL, etc. Despite the legal issues surrounding Linked Data licenses [37], it is still considered a gold mine for organizations who are trying to leverage external data sources in order to produce more informed business decisions [11]. In [25] the authors see the potential economic effect unfolding in education, transportation, consumer products, electricity, oil and gas, health care and consumer finance. They estimate the potential annual value enabled by Open Data in these domains to be 3 trillion US Dollars across seven domains. As a result, validating license related information is vital. However, from our experiments, we found out that datasets' license information is noisy. The license names if found are not standardized. For example, Creative Commons CCZero can be also CC0 or CCZero. Moreover,the license URI if found and if de-referenceable can point to different reference knowledge bases e.g. `http://opendefinition.org`. To overcome this issue, we have manually created a mapping file standardizing the set of possible license names and the reference knowledge base.

Ownership information: Information about the ownership of the dataset. e.g. organization details, maintainer details, author, etc.
The existence of this information is important to identify the authority on which the generated report and the newly corrected profile will be sent to.

[17] http://demo.ckan.org/api/3/action/package_show?id=adur_district_spending

Provenance information: Temporal and historical information on the dataset and its resources. For example, creation and update dates, version information, version, etc. Most of this information can be automatically filled and tracked.

After identifying the underlying portal software, we perform iterative queries to the API in order to fetch datasets metadata and persist them in a file-based cache system. Depending on the portal software we can issue specific extraction jobs. For example, in CKAN based portals, we are able to crawl and extract the metadata of a specific dataset, all the datasets in a specific group e.g. LOD Cloud or all the datasets in the portal.
The caching process can replicate the portal's structure reflecting the various groups or hierarchies defined. Overwriting or disabling caching can be easily done by overloading the call to the extractor.

3.3 Instance and Resource Extraction

From the extracted metadata we are able to identify all the resources associated with that dataset. They can have various types like a SPARQL endpoint, API, file, visualization ,etc. However, before extracting the resource instance(s) we perform the following steps:

- **Resource metadata validation and enrichment**: Check the resource attached metadata values. Similar to the dataset metadata, each resource should include information about its mimetype, name, description, format, valid de-referenceable URL, size, type and provenance. The validation process automatically fills up various missing information when possible, like the mimetype and size. However, missing fields like name and description that needs manual input are marked as missing and will appear in the generated summary report.
- **Format validation**: Validate specific resource formats against a linter or a validator. For example, node-csv[18] for CSV files and n3[19] to validate N3 and Turtle RDF serializations.

Considering that certain dataset contains large amounts of resources and the limited computation power of some machines on which the framework might run on, a sampler module is introduced to execute various sample-based strategies. The following strategies implemented in [20] were found to generate accurate results even with comparably small sample size of 10%.

- **Random Sampling**: Randomly selects resources instances.
- **Weighted Sampling**: Weighs each resources as the ratio of the number of datatype properties used to define a resource over the maximum number of datatype properties over all the datasets resources.

[18] https://github.com/wdavidw/node-csv
[19] https://github.com/RubenVerborgh/N3.js

– **Resource Centrality Sampling**: Weighs each resource as the ration of the number of resource types used to describe a particular resource divided by the total number of resource types in the dataset. This is specific and important to RDF datasets where important concepts tend to be more structured and linked to other concepts.

However, the sampler is not restricted only to these strategies. Strategies like those introduced in [32] can be configured and applied in the processing pipeline.

3.4 Profile Validation

A dataset profile should include descriptive information about the data examined. In our framework, we have identified three main profiling information. However, the extensibility of our framework allows for additional profiling techniques to be plugged in easily. For example, a quality profiling module reflecting the dataset quality. The implemented profiling tasks are:

Metadata profiling
The validation process identifies missing information and the ability to automatically correct them.

Statistical profiling

There exist a set of tools designed specifically to provide statistical information about a dataset (see section 2). Providing comprehensive statistical information about a dataset isn't in the scope of this paper. However, to show the extensibility of our framework we provide a simple RDF statistical profiler module that can be easily extended and configured. The information provided for each class is the number: triples, distinct objects, distinct literals, distinct IRI reference objects, distinct blank nodes objects, distinct subjects, distinct IRI reference subjects and distinct blank nodes subjects.

Topical profiling

Similar to the statistical profiler, a detailed survey of the existing tools can be found in the related work section. However, we implement a very basic topical profiler by applying Named Entity Disambiguation (NED) on the textual description and title of a dataset using DBpedia Spotlight [34].

3.5 Profile and Report Generation

At this point, our framework has successfully validated the dataset profile, highlighting the missing information and presenting them in a human readable report. The report can be automatically sent to the dataset maintainer email if exists in the metadata. The generated report looks like:

```
relationships_as_object field exists but there is no value defined'
private field exists but there is no value defined'
```

```
creator_user_id field is missing'
relationships_as_subject field exists but there is no value defined'
isopen field exists but there is no value defined
```

Listing 1.1. Excerpt of a generated validation report

In addition to the generated report, the enhanced profiles are represented in RDF using the VoID vocabulary and are publicly available according to linked data principles[20].

4 Results and Evaluation

5 Conclusions and Future Work

References

1. Z. Abedjan, T. Gruetze, A. Jentzsch, and F. Naumann. Profiling and mining rdf data with prolod++. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 1198–1201, March 2014.
2. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets - on the design and usage of void, the 'vocabulary of interlinked datasets'. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*, Madrid, Spain.
3. S. Auer, J. Demter, M. Martin, and J. Lehmann. Lodstats — an extensible framework for high-performance dataset analytics. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management*, EKAW'12, pages 353–362, 2012.
4. A. J. Benedikt Forchhammer and F. Naumann. Lodop - multi-query optimization for linked data profiling queries. In *In Proceedings of the International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES) in conjunction with ESWC.*, Heraklion, Greece, 0 2014.
5. 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.
6. 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.
7. C. Böhm, G. Kasneci, and F. Naumann. Latent topics in graph-structured data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 2663–2666, 2012.
8. C. BöHm, J. Lorey, and F. Naumann. Creating void descriptions for web-scale data. *Web Semant.*, 9(3):339–345, Sept. 2011.
9. C. Bohm, F. Naumann, Z. Abedjan, D. Fenz, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with prolod. In *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on*, pages 175–178, March 2010.
10. 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.

[20] SPARQL Endpoint here ?

11. 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.
12. M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. DataID: Towards semantically rich metadata for complex datasets. In *Proceedings of the 10th International Conference on Semantic Systems*, 2014.
13. M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 249–260, 2013.
14. R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic sitemaps: Efficient and flexible access to datasets on the semantic web. In *Proceedings of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC'08, pages 690–704, 2008.
15. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing linked datasets with the VoID vocabulary. W3C note, W3C, Mar. 2011. http://www.w3.org/TR/2011/NOTE-void-20110303/.
16. M. d'Aquin and E. Motta. Watson, more than a semantic web search engine. *Semant. web*, 2(1), Jan. 2011.
17. R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. *The Semantic Web: Research and Applications*, 2010.
18. L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. *CIKM04*, 2004.
19. J. Erickson and F. Maali. Data catalog vocabulary (DCAT). W3C recommendation, W3C, Jan. 2014. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/.
20. B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl. A scalable approach for efficiently generating structured dataset topic profiles. In *The Semantic Web: Trends and Challenges*, volume 8465 of *Lecture Notes in Computer Science*. Springer International Publishing, 2014.
21. M. Frosterus, E. Hyvönen, and J. Laitio. Creating and publishing semantic metadata about linked and open datasets. In D. Wood, editor, *Linking Government Data*, pages 95–112. Springer New York, 2011.
22. M. Frosterus, E. Hyvönen, and J. Laitio. Datafinland - a semantic portal for open and linked datasets. In *ESWC (2)'11*, pages 243–254, 2011.
23. M. Frosterus, E. Hyvönen, and J. Laitio. Datafinland—a semantic portal for open and linked datasets. In *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 243–254. Springer Berlin Heidelberg, 2011.
24. A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data summaries for on-demand queries over linked data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 411–420, 2010.
25. D. F. S. V. K. P. G. James Manyika, Michael Chui and E. A. Doshi. Open data: Unlocking innovation and performance with liquid information. *McKinsey Business Technology Office*, 2013.
26. A. Jentzsch. Profiling the web of data. In *The 2014 International Semantic Web Conference, Doctoral Consortium*, 2014.
27. S. Khatchadourian and M. P. Consens. Explod: Summary-based exploration of interlinking and rdf usage in the linked open data cloud. In *Proceedings of the 7th International Conference on The Semantic Web: Research and Applications - Volume Part II*, ESWC'10, pages 272–287, 2010.

28. M. Konrath, T. Gottron, S. Staab, and A. Scherp. Schemex - efficient construction of a data catalogue by stream-based indexing of linked data. *Web Semant.*, 16:52–58, Nov. 2012.
29. T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing linked data dynamics. In *The Semantic Web: Semantics and Big Data*, volume 7882 of *Lecture Notes in Computer Science*, pages 213–227. 2013.
30. 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.
31. A. Langegger and W. Woss. Rdfstats - an extensible rdf statistics generator and library. In *Proceedings of the 2009 20th International Workshop on Database and Expert Systems Application*, DEXA '09, pages 79–83, 2009.
32. J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, 2006.
33. 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.
34. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, 2011.
35. E. Mäkelä. Aether – generating and viewing extended void statistical descriptions of rdf datasets. In *Proceedings of the ESWC 2014 demo track, Springer-Verlag*, 2014.
36. A. Nikolov, M. d'Aquin, and E. Motta. What should i link to? identifying relevant sources and classes for data linking. In *The Semantic Web*, volume 7185 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012.
37. K. J. Prateek Jain, Pascal Hitzler and C. Venkatramani. There's no money in linked data. 2013.
38. 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.
39. G. Tummarello, S. Danielczyk, R. Cyganiak, R. Delbru, M. Catasta, E. P. Federale, and S. Decker. Sig.ma: Live views on the web of data. In *In Proc. WWW-2010*, pages 1301–1304. ACM Press, 2010.
40. R. Usbeck, M. Röder, A.-C. Ngonga-Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – general entity annotation benchmark framework. In *Submitted to the 24th WWW conference*, 2015.