Undefined 1 (2016) 1–5
IOS Press



# RDF Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications

Mohamed Ben Ellefi [a], Zohra Bellahsene [a], John G. Breslin [b], Elena Demidova [c], Stefan Dietze [c], Julian Szymański [d] and Konstantin Todorov [a]

[a] *LIRMM, University of Montpellier and CNRS, Montpellier, France*
*E-mail: {benellefi, bella, todorov}@lirmm.fr*
[b] *Insight Centre for Data Analytics, NUI Galway, University Road, Galway, Ireland*
*E-mail: john.breslin@nuigalway.ie*
[c] *L3S Research Center, Appelstr. 9a, 30167 Hannover, Germany*
*E-mail: {demidova, dietze}@L3S.de*
[d] *Gdańsk University of Technology, Poland*
*E-mail: julian.szymanski@eti.pg.gda.pl*

**Abstract.** The Web of Data, and in particular Linked Data, has seen tremendous growth over the past years. However, reuse and take-up of these rich data sources is often limited and focused on a few well-known and established RDF datasets. This can be partially attributed to the lack of reliable and up-to-date information about the characteristics of available datasets. While RDF datasets vary heavily with respect to the features related to quality, provenance, interlinking, licenses, statistics and dynamics, reliable information about such features is essential to enable dataset discovery and selection in tasks such as entity linking, distributed query, search or question answering. Even though there exists a wealth of works contributing to the task of dataset profiling in general, these works are spread across a wide range of communities. In this survey, we provide a first comprehensive overview of the RDF dataset profiling features, methods, tools and vocabularies. We organize these building blocks of dataset profiling in a taxonomy and illustrate the links between the dataset profiling and feature extraction approaches and several application domains. This survey is aimed towards data practitioners, data providers and scientists, spanning a large range of communities and drawing from different fields such as dataset profiling, assessment, summarization and characterization. Ultimately, this work is intended to facilitate the reader to identify the relevant features for building a dataset profile for intended applications together with the methods and tools capable of extracting these features from the datasets as well as vocabularies to describe the extracted features and make them available.

Keywords: Linked Data assessment, RDF dataset profiling, Dataset features, Dataset profiling vocabularies

## 1. Introduction

The Web of Data, and in particular Linked Data [8], has seen tremendous growth over the past number of years, leading up to the availability of a large amount of RDF datasets[1] on the Web, where a recent crawl[2] of linked datasets retrieved over 1000 datasets alone, including over 8 million explicit resources and an estimated 100 billion triples. RDF datasets and their inherent subgraphs vary heavily with respect to their size, topic and domain coverage, resource types and schemas as well as their dynamics and currency.

Given this scale, the discovery of suitable RDF datasets, which satisfy specific criteria, has become a challenging problem for a variety of applications including *entity linking*, *entity retrieval*, *distributed search*, *query federation*, and *question answering*, just to name a few. This prevalent problem is underlined

[1] For readability, we use the terms "RDF dataset" and "dataset" interchangeably within this survey.

[2] `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state`

by the strong bias towards using established and well-known reference knowledge graphs such as DBpedia [3], YAGO [71] or Wikidata[3], although there exists a long tail of potentially suitable domain-specific yet under-recognized datasets.

In the context of this survey, *an RDF dataset* is defined in accordance with the dataset definition in the Vocabulary of Interlinked Datasets (VoID)[4], namely: "*A Dataset* is a set of RDF triples that are published, maintained or aggregated by a single provider"[5]. According to VoID, a dataset represents a meaningful collection of triples as envisioned by its provider. *An RDF Dataset Profile* is a formal representation of a set of dataset characteristics (features). It describes the dataset and aids dataset discovery, recommendation and comparison with regard to the represented features. *A Dataset Profile Feature* is a characteristic describing a certain attribute of the dataset. For instance, "dataset conciseness" is a dataset profile feature providing information on the degree of redundancy of the information contained in the dataset. A dataset profile is extensible with respect to the features it contains. Usually, the relevant feature set is application-oriented and depends on the envisaged application scenarios.

A number of popular dataset registries have emerged, which tackle the problem of dataset discovery through the curation of lightweight dataset descriptions, often also exposing structured metadata according to the state-of-the-art vocabularies such as DCAT[6] or VoID. Popular examples include DataHub[7] or DataCite[8], while the LinkedUp Catalog[9] (for education) represents a domain-specific example. However, while such metadata is usually edited and curated manually, it is often sparse, not in sync with the constant evolution of the actual datasets, and prone to errors.

On the one hand, as the Web of Data as a whole is evolving along with the constant evolution of individual datasets, manual assessment and representation of a large variety of dataset features is neither feasible nor sustainable. On the other hand, a wide variety of competing as well as complementary approaches exist, aimed at automatic assessment and description of arbitrary datasets. This body of work is spanning several research communities and includes works in fields such as *dataset characterisation*, *data summarisation*, *dataset assessment* or *dataset profiling*. While the problem of dataset profiling is of particular importance in the context of the Web of Data, it has been identified and approached already in other related fields, such as general database and data management research. Emerging from the aforementioned works, a wealth of tools, methods, vocabularies and applications for assessing, describing and profiling datasets has become available throughout the past few years, where a comprehensive overview and classification is still missing. Myriads of terms and notions do co-exist, whereas a clear distinction, classification and comparison is still required. Only recently, some first efforts have been made to bring together such disparate yet closely related fields, e.g. in [21].

The aim of this survey is to provide researchers, dataset providers and application developers with an overview of *dataset profiling* and closely related approaches, including *dataset profile features*, *feature extraction methods and tools*, *vocabularies*, and example *applications* to encourage experimentation and facilitate the broader use of RDF datasets. Being the first comprehensive study in this area, we provide a thorough analysis and definition of related terms and typical dataset profile features. Furthermore, we provide a systematic study of the available methods and tools for assessing and profiling structured datasets, and survey state-of-the-art vocabularies for representing dataset profiles.

While some of the discussed works are dedicated to profiling RDF datasets in particular, works of relevance from other related fields are also discussed. In this survey we address domain-agnostic dataset profiling approaches (e.g., the Linked Data Observatory [26]) as described in Section 4, and RDF-based vocabularies for representing resource metadata, such as general metadata, quality, provenance, links, licensing, statistics and dynamics, which are applicable to datasets as a particular kind of resource on the Web as described in Section 5. It should be noted that domain-specific vocabularies (e.g., Medical Subject Headings (MESH)[10] or Systematized Nomenclature of Medicine (SNOMED)[11]) are out of the scope of this survey, even though they can be useful in formalizing domain-specific aspects of a dataset description.

[3] https://www.wikidata.org
[4] http://vocab.deri.ie/void
[5] http://www.w3.org/TR/void/#dataset
[6] http://www.w3.org/TR/vocab-dcat/
[7] http://www.datahub.io
[8] https://www.datacite.org/
[9] http://data.linkededucation.org/linkedup/catalog/
[10] https://www.nlm.nih.gov/mesh
[11] http://www.snomed.org

In summary, in this survey we provide the following contributions:

- A taxonomy of dataset profile features, including "general", "qualitative", "provenance", "links", "licensing", "statistical" and "dynamics" feature categories;
- A systematic overview of dataset profile feature extraction approaches and tools and their discussion in the context of the taxonomy;
- An overview and a classification of available vocabularies for representing dataset features and profiles according to the taxonomy;
- An illustration of the use of dataset profiles in several application scenarios.

The remainder of the survey is organized as follows: In Section 2, we present the methodology adopted to collect and organise the publications included in this survey. Next, we provide a comprehensive set of commonly investigated dataset features (Section 3), based on the existing literature and organize these features into a taxonomy. Then, we provide an overview of the existing approaches and tools for the automatic extraction of dataset profile features (Section 4), followed by an overview of existing RDF vocabularies for the representation of certain dataset profiles and features (Section 5). Where feasible, we also provide suggestions on vocabulary use and offer vocabulary recommendations suitable for representing particular dataset profile features. We conclude by exemplifying subsets of features considered relevant in selected application scenarios in Section 6, and have a final discussion in Section 7.

## 2. Survey Procedure

In this section, we present the methodology adopted to select the publications discussed in this survey. The stages of the survey process are depicted in Fig. 1 and described in the following.

### 2.1. Terminology, Taxonomy and Search Process

As a starting point, we identified a basic terminology of dataset profile features, from which we extracted keywords that were potentially relevant for the scope of this survey, such as profiling, dynamicity, quality, index, etc. These keywords were defined and embedded into a taxonomy, which guided the overall study. The taxonomy was iteratively refined throughout the process. During the review process, we updated the taxonomy and consequently further modified the keywords by both including or excluding relevant features. The extracted keywords from the taxonomy were used individually and in combination to query several online databases and search engines (Fig. 1). For example, we used keywords and multiword expressions to build the following combinations: $\{$Semantic Web, Linked Data, Linked Open Data (LOD), ...$\}$ $\times$ $\{$profiling, dynamicity, quality, index, ...$\}$.

### 2.2. Literature Review

Each category of the resulting taxonomy covers a range of works in the Semantic Web and related fields and would potentially deserve a dedicated survey. In this survey, we provide a pivotal guide for readers to obtain a global view on the various dataset profile features illustrated by examples. For this purpose, we focused our review on key approaches established in each category of the taxonomy, while providing examples for: *(i)* The identification of the feature extraction methods and tools (Section 4); *(ii)* The identification of vocabularies for dataset profile representation (Section 5); and *(iii)* The illustration of some application-driven profiles (Section 6).

### 2.3. Overview of Selected Publications

By applying the selection and review procedure described in Section 2.1, we obtained a list of $85$ publications ranging from $1993$ to $2017$ with about $78\%$ of publications originating from $[2009 - 2016]$ as depicted in Fig. 2 and $7$ W3C recommendations. The publications considered include $22$ journal and $1$ magazine articles, $40$ conference papers, $19$ workshop papers, $1$ book and $2$ PhD theses as listed below.

#### 2.3.1. Journals

1. Semantic Web Journal (SWJ) [65,76,85].
2. Information Processing and Management (IPM) [5].
3. ACM Computing Surveys (CSUR) [9,14].
4. Journal of Web Semantics (JWS) [7,11,25,43, 48].
5. Australasian Medical Journal (AMJ) [50].
6. Transactions of the Association for Computational Linguistics (TACL) [54].
7. International Journal on Semantic Web and Information Systems (IJSWIS) [8,21,62].
8. Journal of Information Systems (JIS) [82,68].

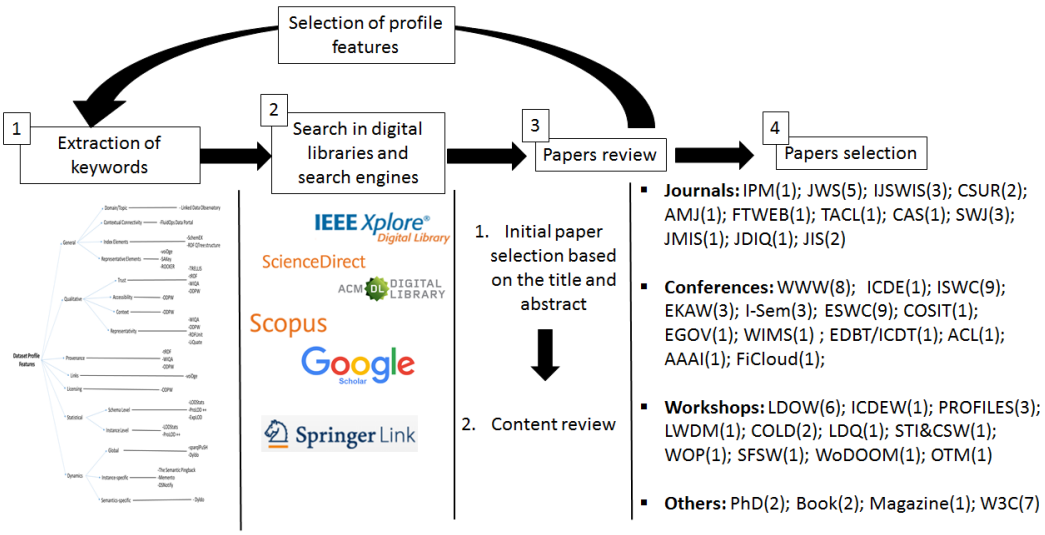


Fig. 1. Survey methodology workflow.

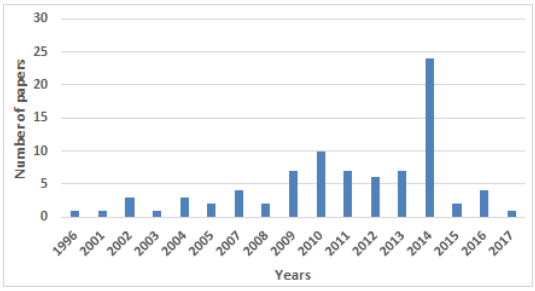


Fig. 2. Referenced papers per year.

9. Foundations and Trends in Web Science (FTWEB) [53].
10. Journal of Management Information Systems (JMIS) [80].
11. Cybernetics and Systems (CAS) [75].
12. Journal of Data and Information Quality (JDIQ) [56].

*2.3.2. Conference Proceedings*

1. International Semantic Web Conference (ISWC) [3,27,34,39,42,57,69,73,83].
2. International World Wide Web Conference (WWW) [10,13,15,36,49,64,70,71].
3. International Conference on Knowledge Engineering and Knowledge Management (EKAW) [4,29,31].
4. IEEE International Conference on Data Engineering (ICDE) [1].
5. I-Semantics (I-Sem) [16,18,84].
6. Extended Semantic Web Conference (ESWC) [23,26,33,41,45,47,67,79,81].
7. Conference on Spatial Information Theory (COSIT) [44].
8. International Conference on eDemocracy and eGovernment (EGOV) [51].
9. Association for the Advancement of Artificial Intelligence (AAAI) [72].
10. IEEE International Conference on Future Internet of Things and Cloud (FiCloud) [77].
11. Joint International Conference on Extending Database Technology and International Conference on Database Theory (EDBT/ICDT) [52].
12. Annual Meeting of the Association for Computational Linguistics (ACL) [66].
13. International Conference on Web Intelligence, Mining and Semantics (WIMS) [58].

*2.3.3. Workshop Proceedings*

1. Workshop on Linked Data on the Web (LDOW) at WWW [2,17,19,20,38,46].
2. International Conference on Data Engineering Workshops (ICDEW) [12].
3. Workshop on PROFIling & fEderated Search for Linked Data (PROFILES) at ESWC [24,28,78].
4. International Workshop on Consuming Linked Data (COLD) [22,32].
5. International Workshop on Linked Web Data Management (LWDM) [30].
6. Workshop on Linked Data Quality (LDQ) at I-Semantics [35].
7. STI Berlin & CSW PhD Workshop (STI&CSW) [37].

8. Workshop on Ontology and Semantic Web Patterns (WOP) [40].
9. Workshop on Scripting and Development for the Semantic Web (SFSW) [60].
10. International Workshop on Debugging Ontologies and Ontology Mappings (WoDOOM) at ESWC [61].
11. On the Move to Meaningful Internet Systems Workshops (OTM) [74].

*2.3.4. Magazines*

1. Communications of the ACM [63].

*2.3.5. Books*

1. Felix Naumann. *Quality-driven Query Answering for Integrated Information Systems*. Springer-Verlag. 2002. [55].

*2.3.6. PhD Theses*

1. Christian Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems*. 2007. [6].
2. Fabrizio Orlandi. *Profiling User Interests on the Social Semantic Web*. 2014. [59].

*2.3.7. W3C Recommendations*

1. OWL Web Ontology Language Reference: `https://www.w3.org/TR/owl-ref`.
2. RDF Schema 1.1: `https://www.w3.org/TR/rdf-schema`.
3. Data Catalog Vocabulary (DCAT): `http://www.w3.org/TR/vocab-dcat`.
4. SKOS Simple Knowledge Organization System Reference: `https://www.w3.org/TR/2009/REC-skos-reference-20090818`.
5. The RDF Data Cube Vocabulary: `http://www.w3.org/TR/vocab-data-cube`.
6. PROV-O: The PROV Ontology: `https://www.w3.org/TR/prov-o`.
7. PROV-DM: The PROV Data Model: `https://www.w3.org/TR/prov-dm`.

Overall, in this survey, we aim to give the reader a bird's-eye view of the RDF datasets profiling problem (whether or not referred to explicitly by using this term), while providing some examples of a worm's-eye view, especially in terms of feature extraction methods, vocabularies for dataset profile representations and application-driven profiles.

## 3. Dataset Profiling: Features and Taxonomy

This section provides an inventory of dataset features of relevance to dataset profiling. The features identified in the literature are grouped in an extensible feature taxonomy, which provides a categorization system for the purpose of this survey. This taxonomy reflects the authors' consensus and provides one way (of several feasible ones) to structure the profiling features.

In particular, we propose to organise the features into the following top-level categories: "General", "Qualitative", "Provenance", "Links", "Licensing", "Statistical" and "Dynamics". This feature categorization guides the categorization of profiling tools and vocabularies in the subsequent sections.

Fig. 3 depicts the resulting taxonomy including references to instances of feature extraction systems (discussed in more detail following the taxonomy structure in Section 4). Although we do not discuss the measurements for the different dataset features in detail within this survey, they partially follow from the definition of a particular feature (e.g., in case of statistical features) or have been extensively discussed in the literature (e.g., qualitative features in [85]).

### *3.1. General Features*

General features are dataset profile features carrying high-level semantic information (e.g., domain and topic of the dataset) that do not fit into any of the more specific categories defined in this survey.

1. **Domain/Topic** A domain refers to the field of knowledge that the dataset pertains to (e.g., music, people). It captures the topics covered by a dataset (e.g., life sciences or media). Topics can be either represented through literals, that is, sets of words, or by structured topic references, such as entities or categories, where [26] provides an example of using DBpedia categories as topic indicators.
2. **Contextual Connectivity** This feature describes the dataset in the context of other datasets. Here, we identify two members of this group as stated in [78]:
   (a) ***connectivity properties***: the set of entities shared with other datasets.
   (b) ***domain/topical overlap with other datasets***: the overlap of the domains or topics covered by a dataset and other datasets. This overlap can provide important information, especially with regard to user queries and can be expressed, for instance, by the presence of shared topics or entities between two datasets.

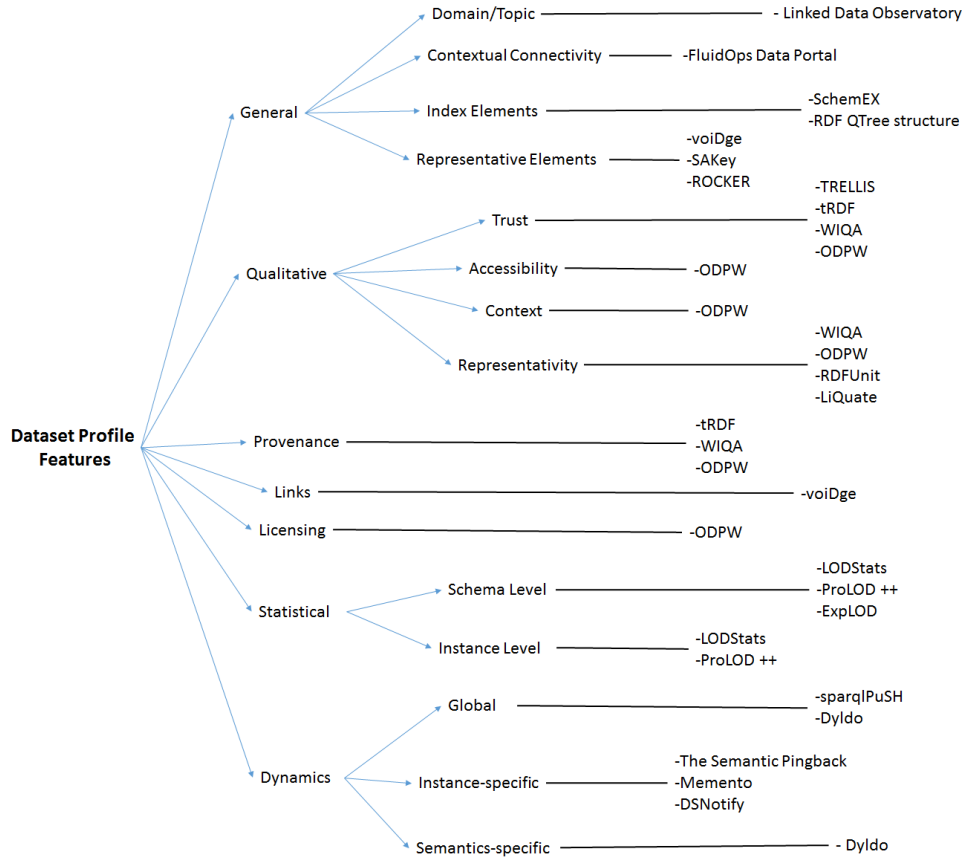


Fig. 3. A taxonomy including dataset profile features organized into General, Qualitative, Provenance, Links, Licensing, Statistical and Dynamics categories (blue arrows) as well as links to the corresponding feature extraction systems (black lines).

3. **Index Elements** Index models have been introduced in order to retrieve information from the LOD graph. An index is defined as a set of key elements (e.g., types), which are used to look up and retrieve RDF data instances. A dataset, therefore, can be inversely described by the set of index elements that are pointing to it in a given index or a set of indices. In that sense, a set of index elements is viewed as a descriptive general dataset feature. These elements can be defined at the schema level (e.g., [48]) or at the instance level (e.g., [36]).
4. **Representative Elements** This group of features is also found both at the schema and at the instance level. On the one hand, representative schema elements can be understood as *(i)* the set of most descriptive types (schema concepts) [23], or *(ii)* the set of schema properties that can be used as keys (almost keys) in instance identification. On the other hand, representative instances are understood as a data sample that accurately portrays the whole dataset [24].

### 3.2. *Qualitative Features*

The study of data quality has a strong and ongoing tradition in the computer science community at large, and in particular, with respect to Web data and data reuse. According to [80], data quality is generally conceived as *fitness for use*, i.e. the capability of data to respond to the demands of a specific user given a specific use case. Data quality has multiple dimensions, and many of them cannot be evaluated in a task-independent manner.

In the context of Linked Data, Bizer et al. [7] classified data quality metrics into three groups according to the type of information that is used as a quality dimension: (i) Content-based metrics – analyzing the information content or comparing information to related information; (ii) Context-based metrics – employing

meta-information about the information content and the circumstances in which information was claimed; and (iii) Rating-based metrics – relying on explicit ratings about the information itself, information sources, or information providers. Zaveri et al. [85] identified further dimensions and reorganized the quality dimension into four groups: *(i) Accessibility; (ii) Intrinsic; (iii) Contextual; and (iv) Representational*. Another approach to assess metadata quality can be found in [56], monitoring the quality of 259 Open Data portals (as of May 2017) classified in five quality dimensions: *existence, conformance, retrievability, accuracy* and *openness*.

In this work, we collected commonly used qualitative features and re-arranged them into the following categories: *(1) Trust; (2) Accessibility; (3) Representativity; and (4) Context/Task Specificity*.

1. **Trust** Data trustworthiness, which is particularly important when dealing with Web Data, can be expressed by the following features.
   (a) ***verifiability:*** the "degree and ease, with which the information can be checked for correctness", according to [6].
   (b) ***believability:*** the "degree, to which the information is accepted to be correct, true, real and credible" [63]. This can be verified by the presence of the provider/contributor in a list of trusted providers.
   (c) ***reputation:*** a judgement made by a user to determine the integrity of a source [85]. The following two aspects are to take into consideration:
      i. ***reputation of the data publisher:*** a score coming from a survey in a community that determines the reputation of a source; and
      ii. ***reputation of the dataset:*** scoring the dataset on the basis of its Web references.
2. **Accessibility** This family of features refers to various aspects regarding the process of accessing data.
   (a) ***availability:*** the extent, to which information is available and easily accessible or retrievable [6].
   (b) ***security:*** refers to the degree to which information is passed securely from users to the information source and back [85].
   (c) ***performance:*** the response time in query execution [85].
   (d) ***versatility of access:*** a measure of the provision of alternative access methods to a dataset [85].
3. **Representativity** The features included in this group provide information in terms of noisiness, redundancy or missing information in a given dataset.
   (a) ***completeness:*** the degree, to which all required information regarding schema, properties and interlinking is present in a given dataset [85]. In the Linked Data context, the following sub-features are defined in [6]:
      i. ***schema (ontology) completeness*** – refers to the degree to which the classes and properties of a schema are represented in the dataset.
      ii. ***property completeness*** – refers to the degree to which values are missing for a specific property.
      iii. ***population completeness*** – the percentage of all real-world objects of a particular type that are represented in the dataset.
      iv. ***interlinking completeness*** – refers to the degree to which links are missing in a dataset.
   (b) ***understandability:*** refers to expression, or, as defined by [63], the extent to which data is easily comprehended.
   (c) ***accuracy / correctness:*** the equivalence between an instance value in a dataset and the actual real-world value corresponding to that instance.
   (d) ***conciseness:*** the degree of redundancy of the information contained in a dataset.
   (e) ***consistency:*** the presence of contradictory information.
   (f) ***versatility:*** whether data is available in different serialization formats, or in different formal and/or natural languages.
4. **Context/Task specificity** This category is comprised of features that refer to the data quality with respect to a specific task.
   (a) ***relevance:*** the degree to which the data needed for a specific task is appropriate (applicable and helpful) [63], or the importance of data to a given user query [6].

(b) ***sufficiency:*** the availability of a sufficient amount of data for a particular task (the expression *"amount-of-data"* is used in [6]).
(c) ***timeliness:*** — inspired by a definition from [85], the timeliness feature refers to the availability of timely information in a dataset with regard to a given application.

### 3.3. Provenance Features

A variety of definitions have been given for provenance over the past number of years. One very pragmatic definition comes from the W3C Provenance Working Group[12], especially when thought of in the context of the Web: *"Provenance is defined as a record that describes the people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing."* Provenance can pertain to any resource found on the Web — documents, data, or datasets — but also to the real-world objects described by web resources. It can be seen as the piece of contextual metadata that provides indicators about timeliness, currency and update cycles of datasets – important characteristics that allow us to understand the origins of data, to trace errors and, ultimately, to establish trust.

### 3.4. Links Features

"Links" here is understood as the number of datasets with which a dataset is interlinked, or as the number of triples in a dataset, in which the subject and the object refer to different datasets. Two datasets can be linked through: *(i)* explicit links when they have linked instances, for example when sharing instances by using *owl:sameAs*[13] [43]; and *(ii)* implicit links when sharing topic profiles or context profiles, where explicit links like *rdfs:seeAlso*[14] can also be used [78]. Note that this category of features can be seen as a type of statistical feature (described below). However, it is assigned a separate category in our taxonomy in order to reflect its importance. This category of features covers both schema-level and instance-level representations of links in a dataset profile.

[12] http://www.w3.org/TR/2013/REC-prov-dm-20130430/
[13] https://www.w3.org/TR/owl-ref/#sameAs-def
[14] https://www.w3.org/TR/rdf-schema/#ch_seealso

### 3.5. Licensing Features

Here, we adopt the recommendation of Heath et al. [42]: "in order to enable information consumers to use your data under clear legal terms, each RDF document should contain a license, under which the content can be used". In other words, the type of license, under which a dataset is published, indicates whether reproduction, distribution, modification or redistribution are permitted. This can have a direct impact on data quality, both in terms of trust and accessibility. Hence the availability of license information is important in both human-readable and machine-readable profiles (i.e. including the description in a license vocabulary, Section 5.5).

### 3.6. Statistical Features

This group of features comprises a set of statistical features, such as size, coverage, average number of triples, property co-occurrence and others [4,28].

1. **Schema-level** With respect to schema, we can compute statistical features such as *class / properties usage count*, *class / properties usage per subject and per object* or *class / properties hierarchy depth*.
2. **Instance-level** Features at the instance level are computed according to the data instances only, e.g.: URI usage per subject (/object), triples having a resource (/blanks) as subject (/object), triples with literals, min(/max/avg.) per data type (integer/float/time, etc.), number of internal and external links, number of ingoing (/outgoing) links per instance, number of used languages per literal, classes distribution as subject (/object) per property, property co-occurrence.

### 3.7. Dynamics Features

This class of features concerns the dynamicity of a dataset. In principle, every dataset feature can be dynamic, i.e. changing over time (think, for example, of data quality). Inversely, the dynamics of a dataset can be seen as a separate feature describing data (take the example of the dynamics of data quality). For that reason, this family of features is seen as transversal (spanning over the groups of features described above). Käfer et al. [46] provide a study of LOD dynamicity use cases, based on which we identify the following sub-categories:

1. **Global**
   (a) ***lifespan:*** measured on an entire dataset or parts of it.
   (b) ***stability:*** an aggregation measure of the dynamics of all dataset features.
   (c) ***update history:*** a feature with multiple dimensions regarding the dataset update behavior, divided into:
      i. ***frequency of change:*** the frequency of updating a dataset, regardless of the kind of update.
      ii. ***change patterns:*** the existence and kinds of categories of updates, or change behavior.
      iii. ***degree of change:*** to what extent the performed updates impact the overall state of the dataset.
      iv. ***change triggers:*** the cause or origin of the update as well as the propagation effect reinforced by the links.
2. **Instance-specific**
   (a) ***growth rate:*** the level of growth of a dataset in terms of data instances.
   (b) ***stability of URIs:*** the level of constancy of URIs (knowing that an URI can be, for example, modified or removed).
   (c) ***stability of links:*** the level of broken links between resources. A link between a resource and a target URI is considered as broken if the target URI changes [64]. This implies that, while the stability of URIs is rated with respect to the source dataset, the stability of links/backlinks is rated with respect to the stability of the linked URIs in linked datasets.
3. **Semantics-specific** [22,33]
   (a) ***structural changes:*** this feature evaluates the degree of modification in the internal or external structure of a dataset.
   (b) ***domain-dependent changes:*** this feature reflects the dynamics across different domains that impacts data.
   (c) ***vocabulary-dependent changes:*** this is a measure of the dynamics of vocabulary usage.
   (d) ***vocabulary changes:*** this is a measure of the impact of a change in a vocabulary on the dataset that uses it.
   (e) ***stability of index models:*** this features describes the level of change in the original data after data indexing.

## 4. Dataset Profiling and Feature Extraction Methods & Tools

The field of dataset profiling and feature extraction is comprised of a broad range of tools, and is much too extensive to fully cover here. Therefore, we provide examples of relevant dataset profiling approaches for each category of features, as introduced in the previous section (Fig. 3). We describe these approaches according to their respective categories below.

### *4.1. General Features*

General features, presented in Section 3.1, include "Domain/Topic", "Contextual Connectivity", "Index Elements" and "Representative Elements". In the following, we present a selection of tools that support feature extraction in this category.

**FluidOps Data Portal** [78] is a framework for source contextualization. It allows users to explore the space of a given source, i.e. to search and discover data sources by topics[15]. Here, the contextualization engine favors the discovery of relevant sources during exploration. For this, entities are extracted and clustered, providing for every source a ranked list of contextualization sources. This approach is based on well-known data mining strategies and does not require schema information or data adhering to a particular form. The FluidOps Data Portal enables the retrieval of "Contextual Connectivity" features.

**Linked Data Observatory** [26] provides an explorative way to browse and search through existing datasets in the LOD Cloud according to the topics they cover. By deploying entity recognition, sampling and ranking techniques, the Linked Data Observatory creates structured dataset topic profiles, allowing one to find datasets providing data for a given set of topics or to discover datasets covering similar fields. These profiles are represented in RDF using the VoID vocabulary in tandem with the Vocabulary of Links (VoL) (see Section 5 for more detail). The Linked Data Observatory allows the extraction of "Domain/Topic" features.

**voiDge** is a tool that automatically generates VoID descriptions for large datasets. This tool allows users

[15] `http://data.fluidops.net/resource/Topics`

to compute various types of VoID information and statistics on dumps of LOD as illustrated in [11]. Additionally, the tool identifies (sub)datasets and annotates the derived subsets according to the VoID specification. Here, we are particularly interested in the attribute "void:exampleResource", which names representative resources within the dataset, i.e. the predicate/object combinations that make good entry points. Hence, voiDge supports the generation of "Representative Elements" dataset profile features, and in particular the set of representative instances.

**Key discovery approaches** aim at selecting the smallest set of relevant predicates representing an RDF dataset within the context of link discovery. In other words, a key represents a set of schema properties that uniquely identifies every instance of a given schema concept. We cite two main key discovery approaches: ***(i)*** *SAKey* [73] — an approach to discover *almost keys* in datasets where erroneous data or duplicates exist. *SAKey* is an extension of *KD2R* [74], which aims to derive exact composite keys from a set of non-keys discovered on RDF data sources; and ***(ii)*** *ROCKER* [70] — a key discovery approach that uses a refinement operator. Reportedly, *ROCKER* is more suited to large scale data than *SAKey*. Keys can be seen as a "Representative Elements" dataset profile feature, and in particular as representative sets of schema properties.

**RDF QTree Structure** [36] is an approximate multidimensional indexing scheme storing descriptions of the content of RDF data sources. A QTree is a combination of histograms and an R-tree multidimensional structure. The method identifies relevant RDF data sources for a given query that incorporates **instance-level** information by adding triples to the corresponding buckets in the QTree. The QTree structure allows the extraction of the "Index Elements" dataset profile feature.

**SchemEX** [48] is a stream-based indexing and schema extraction approach over Linked Data. The schema extraction abstracts RDF instances to RDF schema concepts that represent instances with the same properties. The index contains each schema concept that maps to the data sources containing instances with the corresponding properties. While SchemEX provides a different index structure than that of QTree, both indexing tools relate to the "Index Elements" dataset profile feature from the general features category of our taxonomy.

### *4.2. Qualitative Features*

As discussed in Section 3, in this survey we focus on selected groups of qualitative features such as trust, accessibility, representativity, and context, which are most relevant in the context of dataset profiling. In the following, we discuss a selection of relevant tools for these groups. Note that a broader overview of the quality assessment approaches in the context of Linked Data in general is provided by Zaveri et al. [85], who conducted an extensive survey of 21 works.

**TRELLIS** [31] is an interactive environment that examines the degree of trust of datasets based on user annotations. The user can provide TRELLIS with a semantic markup of annotations through interactions with the ACE tool[16] [10]. The tool allows several users to add and store their observations and viewpoints. The annotations made by the users with ACE can be used in TRELLIS to detect conflicting information or handle incomplete information. Trellis provides a description for the "Trust" profile feature.

**tRDF** [37] is a framework that provides tools to represent, determine, and manage trust values that represent the trustworthiness of RDF statements and RDF graphs. It contains a query engine for tSPARQL, a trust-aware query language. tSPARQL extends the RDF query language SPARQL with two clauses: the TRUST AS clause and the ENSURE TRUST clause. Trust values are based on subjective perceptions about the query object. While TRELLIS is based on users' annotations, tRDF extracts the "Trust" feature by allowing users to query the dataset and access the trust values associated with the query results in a declarative manner.

**WIQA** [7] is a generic qualitative platform allowing one to evaluate the trust of a dataset using a wide range of different filtering policies based on quality indicators like provenance information, ratings, and background information about information providers. This framework is composed of two components: a Named Graph Store for representing information together with quality related meta-information, and an engine, which enables applications to filter information and to retrieve explanations about filtering decisions. WIQA policies are expressed using the WIQA-PL syntax, which is based on the SPARQL query language. WIQA can provide descriptions related to the "Trust" and "Representativity" dataset profile features.

[16] Annotation Canonicalization through Expression synthesis.

**RDFUnit** [49] is a framework for data quality assessment that tests RDF knowledge bases by using Data Quality Test Patterns (DQTP). These patterns can have different forms, e.g.: ***(i)*** "a resource of a specific type should have a certain property", ***(ii)*** "a literal value should contain at most one literal for a certain language". The user can select and instantiate existing DQTPs. If an adequate test pattern for a given dataset is not available, the user has to write their own DQTPs, which can then become part of a central library to facilitate later re-use. RDFUnit provides "Representativity" dataset profile features in the form of DQTPs.

**Open Data Portal Watch** (ODPW) [77] is a publicly available dashboard component that displays quality metrics for web-based data portal platforms using various views and charts. For an automatic crawl of data portal metadata, ODPW maps the heterogeneous models of data portals to the Data Catalog Vocabulary (DCAT, Section 5). The framework uses the Data Quality Vocabulary (DQV, Section 5) to make the quality measures of the ODPW framework available as RDF and to link the assessments to dataset descriptions. ODPW covers all qualitative features introduced in this survey, as well as features from other categories (such as "Licensing", and "Provenance"), as we will see in subsequent sections.

**LiQuate** [67] is a tool allowing data quality to be assessed with respect to both link completeness, and ambiguities among labels and links. The quality evaluation relies on queries to a Bayesian Network that models RDF data and dependencies among properties. This allows one to estimate the probability that different resources have redundant labels or that a link between two resources is missing. LiQuate enables the retrieval of the "Representativity" dataset profile features, i.e. the interlinking completeness feature.

### *4.3. Provenance Features*

**tRDF** [37] is a framework that also allows one to generate a provenance model for RDF datasets by using the Provenance Vocabulary (Section 5.3). The query engine of tRDF allows the retrieval of provenance metadata for linked datasets through their SPARQL endpoints. Furthermore, outdated data objects are filtered out from the query results by assessing the provenance metadata timeliness.

**WIQA** [7], already seen in the qualitative features category, also contains a provenance profiling component, which allows the description of provenance metadata for named graphs using the Semantic Web Publishing (SWP) vocabulary (Section 5.3). The framework assumes that for each named graph the provenance profile should be published by a specific metadata provider at a certain point in time. Furthermore, the WIQA browser allows the storage of data together with provenance metadata as a set of named graphs.

**ODPW**, another cross-category framework, allows users to evaluate the trustfulness of data by enabling data traceability and by keeping track of dataset provenance. The framework uses the provenance ontology PROV-O (Section 5.3) to annotate weekly generated snapshots of portals. The use of PROV-O allows the changes within different dataset versions to be tagged over time.

### *4.4. Links Features*

**voiDge**, presented above within the general features group, also allows one to automatically generate descriptions of links for LOD datasets based on the VoID vocabulary. In particular, voiDge provides the "void:Linkset" attribute dedicated to cross-dataset triples (i.e. triples, in which the subject belongs to a different dataset than the one of the object). Furthermore, the tool distinguishes between two categories of linksets: (i) Crisp Linksets, which is the implementation of VoID linksets in a reflexive and non-symmetric way; and (ii) Fuzzy Linksets, where two resources are declared as similar to a certain degree ($k$-similar) if they share a common set of attributes ($k$ of their predicate/object combinations are exact matches).

### *4.5. Licensing Features*

**ODPW**, in addition to its other functionalities discussed above, also provides a search service[17] that retrieves the licenses of a given resource URI in Open Data portals. Licences are assessed by an openness indicator providing information regarding their conformance to the Open Definition[18]. A license specification for a dataset is considered as open only if it matches a license in the Open Definition list. ODPW maintains an up-to-date list of all available licenses via a weekly crawl of the Open Data portals.

[17] http://data.wu.ac.at/portalwatch/licensesearch

[18] http://licenses.opendefinition.org/licenses/groups/all.json

### 4.6. Statistical Features

In this survey we consider statistical feature extraction approaches at both schema and instance level, as defined in Section 3.6.

**LODStats** [4] is a statement-stream-based tool and framework for gathering comprehensive statistics about datasets adhering to RDF. The tool calculates 32 different statistical criteria on LOD such as those covered by VoID. It computes descriptive statistics such as the frequencies of property usage and datatype usage, the average length of literals, or the number of namespaces appearing at the subject URI position. It is available for integration with the CKAN[19] metadata repository, either as a patch or as an external web application using CKAN's API. LODStats provides methods to generate both "Schema-level" and "Instance-level" statistical profile features.

**ExpLOD** [47] creates usage summaries from RDF graphs including metadata about the structure of an RDF graph, such as the sets of instantiated RDF classes of a resource or the sets of used properties. This structural information is aggregated with statistics such as the number of instances per class or the number of properties used. ExpLOD provides descriptions about the "Schema-level" statistical features of a dataset.

**ProLOD++** [1], an enhanced version of **ProLOD** [12], is an interactive web-based user interface, which allows one to visualize a dataset as a cluster tree and explore it by selecting clusters for further statistical data extraction. In addition to mining and cleansing options, the tool is able to generate dataset profiling features related to key analysis, predicate and value distribution, string pattern analysis, link analysis and data type analysis. Hence, ProLOD++ allows arbitrary LOD datasets to be profiled in terms of the "Schema-level" and "Instance-level" statistical profile features.

### 4.7. Dynamics Features

**sparqlPuSH** [60] is an interface that can be plugged into any SPARQL endpoint and that broadcasts notifications to clients interested in what is happening in the store using the PubSubHubbub[20] protocol[21] i.e. $SPARQL + pubsubhubbub = sparqlPuSH$. Practically, this means that one can be notified in real-time of any change happening in a SPARQL endpoint. A resource can ping a PubSubHubbub hub when it changes, then the notifications will be broadcast to interested parties. $sparqlPuSH$ consists of two steps: *(i)* register the SPARQL queries related to the updates that must be monitored in an RDF store, and *(ii)* broadcast changes when data mapped to these queries are updated in the store. Thus, $sparqlPuSH$ extracts "Global" dataset profile features from the "Dynamics" category.

**The Semantic Pingback** [76] is a mechanism that allows users and publishers of RDF content, weblog entries and scientific articles to obtain immediate feedback when other people establish a reference to them or their work, thus facilitating social interactions. It also allows backlinks to be published automatically from the original WebID profile (or other content, e.g., status messages) to comments or references of the WebID (or other content) elsewhere on the Web, thus ensuring timeliness and coherence of datasets. It is based on the advertisement of a lightweight RPC (Remote Procedure Call) service. This system is particularly useful for detecting the stability of links/backlinks. This mechanism provides feedback about the "Instance-specific" features of a dataset profile in the "Dynamics" category.

**Memento** [19] is a protocol-based "time travel" tool that can be used to access archived representations of a resource identified by a given URI. The current representation of a resource is named the *Original Resource*, whereas resources that provide prior representations are named *Mementos*. This system provides relationships like the *first-memento, last-memento, next-memento* and *prev-memento*, available in both HTML and RDF/XML. These relationships are particularly useful for the extraction of the "Instance-specific" features and in particular of the "Growth Rate" feature.

**DSNotify** [64] is a link monitoring and maintenance framework, which attenuates the problem of broken links due to URI instability. When remote resources are created, removed, changed, updated or moved, the

[19] http://ckan.org/

[20] PubSubHubbub is a decentralized real-time web protocol that delivers data to subscribers when they become available. Parties (servers) speaking the PubSubHubbub protocol can get near-instant notifications when a topic (resource URL) they are interested in is updated.

[21] https://en.wikipedia.org/wiki/PubSubHubbub

| Method Name | H/M | Accessibility | Home Page |
|---|---|---|---|
| FluidOps Data Portal | H | O.S. | http://data.fluidops.net |
| Linked Data Observatory | H/M | O.S./Online | http://data-observatory.org/lod-profiles/profile-explorer/ |
| voiDge | H/M | O.S. | https://hpi.de/naumann/projects/btc/btc-2010 |
| SAKey | H | O.S. | https://www.lri.fr/sakey |
| ROCKER | H/M | O.S./Online | http://rocker.aksw.org/ |
| RDF QTree Structure | H/M | – | (⋆)http://swse.deri.org/index.lighttpd.html |
| SchemEX | H/M | – | – |
| TRELLIS | H | O.S. | http://www.isi.edu/ikcap/trellis |
| tRDF | H/M | O.S./Online | http://trdf.sourceforge.net |
| WIQA | H/M | O.S. | http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa |
| LiQuate | H | Online | http://liquate.ldc.usb.ve |
| RDFUnit | H/M | O.S./Online | http://rdfunit.aksw.org |
| ODPW | H | Online | http://data.wu.ac.at/portalwatch |
| LODStats | H/M | O.S./Online | http://aksw.org/Projects/LODStats.html |
| ProLOD++ | H | Online | https://www.hpi.uni-potsdam.de/naumann/sites/prolod++ |
| PubSubHubbub | M | O.S. | https://github.com/pubsubhubbub/ |
| sparqlPuSH | H/M | O.S. | https://code.google.com/archive/p/sparqlpush/ |
| The Semantic Pingback | M | O.S. | https://aksw.github.io/SemanticPingback/ |
| Memento | H/M | – | http://mementoarchive.lanl.gov/ |
| Dyldo | H | Online | http://swse.deri.org/dyldo |
| DSNotify | M | O.S. | http://www.cibiv.at/~niko/dsnotify |

Table 1

Dataset profile feature extraction methods: Homepages (checked on February 2017); Accessibility that can be Open Source (O.S.) or Online (via SPARQL Endpoint or via HTTP API, etc.); and Human readability (H) vs. Machine readability (M).

system revises links to these resources accordingly. This system can easily be extended by implementing custom crawlers, feature extractors, and comparison heuristics. DSNotify relates to the "Instance-specific" features in the "Dynamics" category.

**The Dynamic Linked Data Observatory** (Dyldo) [45] is a framework aiming at the provision of a comprehensive overview of how LOD changes and evolves on the Web over time. The observatory provides weekly crawls of LOD data sources starting from $02/11/2008$, and contains $550$K RDF/XML documents having a total of $3.3$M unique subjects with $2.8$M locally defined entities. The system first examines the usage of Etag and Last-Modified HTTP header fields, and then analyzes the various dynamic aspects of a dataset (changes in frequency, volume, etc.). Dyldo provides "Dynamics"-related dataset profile features including both "Global" and "Semantics-specific" ones.

### *4.8. A Note on Dataset Profiling Methods*

We would like to highlight several issues regarding the dataset profile extraction methods that we observed in the survey process. First, for "General" features, these typically require domain knowledge with respect to the content of the dataset. As a best practice, we recommend that the general category should be provided by the data domain experts (e.g., data providers or maintainers) to ensure a high quality profile. Second, profile features like "Provenance" and "License" are meant to be augmented manually by the data provider and cannot be derived automatically. Third, we consider that "Qualitative", "Links", "Statistical" and "Dynamics" profile features would in general require less domain expertise and can be extracted automatically by applications in many cases. Furthermore, we observe an obvious need for more general profile extraction tools, notably for the "Do-

main/Topic" and "Contextual Connectivity" general features, where only a few automatic extraction approaches exist.

Regarding the dynamics aspects of a dataset, in order to ensure that profiles are not out of date, they need to be regenerated periodically and regularly, according to the dataset dynamicity. Dataset versioning and archiving also requires versioning and archiving of the corresponding profiles in order to ensure coherence between the dataset snapshots and their profile versions.

Finally, we emphasize the fact that RDF dataset profiles need to provide representations for both human and machine readability. Hence, in Table 1, we provide an overview of the dataset profiling methods with respect to their representation formats and we verify for each method if the extracted profile features are designed for humans or machines (or both). In addition, the table provides links to the webpages for each method.

## 5. Vocabularies for Representation of Dataset Profiles and Features

This section introduces vocabularies for the representation of dataset profiles, structured according to the feature categories introduced in Section 3. These vocabularies range from general dataset metadata to vocabularies dedicated to one or more of the features introduced in Section 3. Note that general-purpose vocabularies such as Dublin Core[22] often provide useful terms also for dataset-specific metadata. Even though an exhaustive discussion of such broad vocabularies is not within the scope of this survey, we discuss their use to model specific aspects of datasets, such as provenance or licensing.

### 5.1. General Dataset Metadata

A range of vocabularies exist, which can be used to provide more general metadata about datasets or ontologies, where Dublin Core is an obvious candidate to represent metadata about any resource, including datasets.

While the Ontology Metadata Vocabulary (OMV) [40] is aimed at providing descriptive information about ontologies — specifically their creators, contributors, reviewers, and creation/modification dates — here we focus specifically on dataset metadata vocabularies.

[22] http://dublincore.org/documents/dces/

| Category | Datasets (Percent) |
|---|---|
| Social Web | 6 (1.16) |
| Government | 75 (40.32) |
| Publications | 14 (13.46) |
| Life Sciences | 29 (32.58) |
| User-Gen. Content | 6 (10.91) |
| Cross Domain | 5 (11.36) |
| Media | 2 (5.41) |
| Geographic | 15 (36.59) |
| Total | 140 (13.46) |

Table 2
Adoption of VoID across LOD Datasets per Category[23].

The Vocabulary of Interlinked Datasets (VoID) [2] provides a core vocabulary for describing datasets and their links. The schema[24] includes the classes *Dataset*, *DatasetDescription*, *LinkSet*, *TechnicalFeature*. The authors distinguish *dataset* from *RDF graph*, where *dataset* refers to a "meaningful collection of triples, that deal with a certain topic, [that] originate from a certain source or process, are hosted on a certain server, or are aggregated by a certain custodian." A *LinkSet* is defined as a set of triples, where the subject and object are in different datasets/namespaces. The VoID guidelines recommend additional vocabularies for general metadata such as DC Terms[25] and FOAF[26]. VoID is already widely used in the Web of Data, as documented by Table 2, depicting the use of VoID descriptions among the 1,014 datasets and per category in the current inventory of the Web of Data[27].

The Data Catalog vocabulary (DCAT)[28] follows a similar rationale and has been created based on a survey of government data catalogues [51] and is partly derived from Dublin Core. Key classes include *Catalog*, *Dataset*, *CatalogRecord* where the latter has a similar scope as the VoID DatasetDescription, i.e. it is making the useful distinction between dataset metadata and metadata of the dataset description (the record) itself. Additional classes include *Distribution* — i.e. the instantiation of a particular dataset in a specific access format (e.g., an RDF dump or a SPARQL

[23] Taken from the 30/08/2014 snapshot of the LOD cloud available at http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state

[24] http://vocab.deri.ie/void

[25] http://dublincore.org/documents/dcmi-terms

[26] http://xmlns.com/foaf/spec

[27] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/

[28] http://www.w3.org/TR/vocab-dcat/

endpoint). For the categorisation of datasets, the *dcterms:subject* predicate and controlled SKOS vocabularies are recommended.

### 5.2. Dataset Quality

Early works by Supelar et al. in [72] define a set of knowledge quality features applicable for knowledge graphs, respectively ontologies, and a corresponding ontology. Their features are classified into *quantifiable* and *non-quantifiable* characteristics and include characteristics such as usability, availability, accuracy, or complexity. The suggested ontology, however, only includes a higher level taxonomy, but neither a fully fledged vocabulary for annotation nor a specific set of metrics to quantify the quantifiable metrics.

Fürber et al. [30] describe the DQM Ontology[29], a general vocabulary for representing data quality features, to some extent also covering statistical information, such as notions of property completeness or property uniqueness. Key concepts include:

- Data Quality Assessment as an abstract container of scores and metrics describing class / property quality aspects.
- Completeness, derived into: Property Completeness as a measure of the degree to which properties are consistently populated; and Population Completeness as the degree to which all objects of a certain reference are represented in a specific class.
- Accuracy as a notion representing the degree to which a statement captures the intended semantics and syntax (subtypes are Syntactic Accuracy and Semantic Accuracy).
- Uniqueness of properties and entities is introduced to capture the existence of duplicates.
- Timeliness captures the recency of a specific statement/entity.

The authors also introduce a preliminary classification for data quality problems.

In addition, the Web Information Quality Assessment (WIQA) Framework[30] describes some early work to filter content according to quality features, also introducing WIQA-PL, a vocabulary for modeling content access policies. However, the work appears to be deprecated and not maintained.

Another one worth mentioning is the work in [29], where the authors use the SPARQL Inferencing Notation (SPIN) — a vocabulary that allows the representation of SPARQL queries — to represent data quality rules.

Additionally, the Dataset Quality vocabulary (daQ)[31] [20] and the Data Quality Vocabulary (DQV)[32] provide complementary terms for annotating DCAT dataset descriptions with quality aspects and metrics. While both vocabularies provide a general framework for annotating quality information and metadata about associated metrics, DQV is not specifically tailored for Linked Data quality features but is for any type of dataset, and only describes a framework, yet not particular metrics and measures [65].

Finally, while provenance information often provides indicators about timeliness, currency and update cycles of datasets, Section 5.3 introduces additional vocabularies of relevance.

### 5.3. Data and Dataset Provenance

A provenance record is essentially a record of metadata that details the entities and processes that were involved in creating, modifying and delivering a resource, be it physical or digital. Such records include details about when an item was created, what were the original sources of information used in its creation, what kind of evolution has the resource undergone (e.g., what were the other entities or processes that may have modified the resulting piece of information). Moreau [53] states that "the provenance of a piece of data is the process that led to that piece of data".

This section describes some of the main provenance models used on the Web, some of which have specific applicability in terms of whole datasets.

1. **voidp** [58] builds on and extends the aforementioned VoID linked dataset ontology to describe the provenance relationships of data across linked datasets. Publishers can use a lightweight set of classes and properties to describe the provenance information of data within their linked datasets using voidp. This enables users to find the right data for their tasks based not

[29] http://semwebquality.org/dqm-vocabulary/v1/dqm

[30] http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/#wiqapl

[31] http://purl.org/eis/vocab/daq

[32] https://www.w3.org/TR/vocab-dqv/

only on the types of data being sought but also on the origins of that data, e.g., "given a set of attributes and data authorship conditions, which available resources match a desired set of criteria and where can these resources be found?"

2. From the perspective of archiving and long-term preservation of data, the **Data Dictionary for Preservation Metadata (PREMIS)**[33] set of terms can be used to describe the provenance of archived, digital objects (e.g., files, bitstreams, aggregations and datasets), and therefore has applicability in our scenario. It does not provide provenance information for the descriptive metadata for those objects, and therefore one of the other vocabularies can be used for this.
3. Inspired by the notion of changesets in code or document revisions, the **Changeset vocabulary**[34] consists of a set of terms that can be used to describe changes in the description of a resource. The primary concept is that of a ChangeSet which defines the delta (changes) between versions of a resource description.
4. The **Proof Markup Language (PML)** is used for defining and exchanging proof explanations created by various intelligent systems, including web services, machine learning components, rule engines, theorem provers and task processors. It provides terms for annotating "IdentifiedThings" such as name, description, create date and time, authors, owners, etc. IdentifiedThings are the entities used or processed in an intelligent system, of which a dataset could be one.
5. The **Semantic Web Publishing vocabulary (SWP)** by [15] makes it possible "to represent the attitude of a legal person to an RDF graph. SWP supports two attitudes: claiming the graph is true and quoting the graph without a comment on its truth. These commitments towards the truth can be used to derive a data publisher's or a data creating entity's relation to provided or created artifacts. Furthermore, the SWP allows to describe digests and digital signatures of RDF graphs and to represent public keys."
6. The **Provenance Vocabulary**[35] was developed to describe provenance of Linked Data on the Web. It is defined as an OWL ontology and it is partitioned into a core ontology and supplementary modules.
7. The **Open Provenance Model (OPM)** is used to describe provenance histories in terms of the processes, artifacts, and agents involved in the creation and modification of a resource. The OPM model was the primary outcome of a series of Provenance Challenge workshops, and is one to which many other provenance vocabularies are mapped to. In fact, it was taken as the basis for the development of PROV-O, described below. Two variants exist, the OPM Vocabulary (OPMV)[36] as a lightweight vocabulary, and the OPM Ontology (OPMO)[37] using more advanced OWL constructs.
8. The **PROV Ontology (PROV-O)**[38] was published as a W3C Recommendation in 2013 by the W3C Provenance Working Group to be a new standard ontology for representing provenance. This is part of a larger *PROV* Family of Documents [52] created to support "the widespread publication and use of provenance information of Web documents, data, and resources" – including a Data Model (PROV-DM)[39] and an ontology (PROV-O) – for provenance interchange on the Web. PROV defines a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or any artifact in the world[40].

As well as the above vocabularies that are specifically designed to facilitate provenance and related primitives, there are a number of commonly-used vocabularies and de-facto standards on the Web that also contain terms of relevance to provenance derivation and definition. These include Dublin Core (DC), Friend-of-a-Friend (FOAF), and Semantically Interlinked Online Communities (SIOC). Some of these terms were highlighted by [38], and we outline these and others below. Since a dataset can be identified by a resource, we can use many of the properties described below with full datasets as well as individual resources or pieces of data in those datasets.

[33] http://bit.ly/premisOntology
[34] http://purl.org/vocab/changeset
[35] http://trdf.sourceforge.net/provenance/ns.html
[36] http://purl.org/net/opmv/ns#
[37] http://openprovenance.org/model/opmo
[38] http://www.w3.org/TR/prov-o/
[39] https://www.w3.org/TR/prov-dm/
[40] http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

– **Dublin Core**: *dcterms:contributor* and *dcterms:creator* can be used in analyses of the activity of a user in the data creation process, although the type of the user and their role may need to be further specified using other vocabularies. In our case, it could also be used to identify the creator of an entire dataset. *dc:source* describes the source from which a resource or dataset is derived, and therefore has usefulness as a provenance element. *dcterms:created* and *dcterms:modified* can be used to define both the creation of a resource or dataset and the modification of that resource or dataset respectively. *dcterms:publisher* can be used to define the provider of a particular resource or dataset, although as [38] points out the type of publisher is left ambiguous. Finally, Dublin Core also defines a *dcterms:provenance* term which can link a resource to a set of provenance change statements.
– **Friend-of-a-Friend**: *foaf:made* and its inverse functional property (IFP) *foaf:maker* can be used to link a resource or dataset to the *foaf:Agent* (person or machine) who created it. In addition, the *foaf:account* property can be used to link to a *foaf:Agent* to a *foaf:OnlineAccount* or *sioc:UserAccount*, which in turn can be identified as the means of creation for a resource or dataset (see below).
– **Semantically Interlinked Online Communities**: As with Dublin Core, the properties *sioc:has_creator*, *sioc:has_modifier* (and their IFPs *sioc:creator_of* and *sioc:modifier_of* respectively) can be used to refer to a resource's creators and modifiers (identified by *sioc:UserAccount*s). *sioc:has_owner* and its IFP *sioc:owner_of* indicates ownership. *sioc:ip_address* can be used to link the created data and creator if specified to a URL. Also, *sioc:last_activity_date* can be used to reference the last activity associated with a resource. A *sioc:sibling* can be used to define a new resource (or perhaps a dataset) that is very similar to but differs in some small manner from another one. Finally, *sioc:earlier_version*, *sioc:later_version*, *sioc:next_version* and *sioc:previous_version* can be used to connect versioned artifacts together as one would find in a provenance graph.
– In addition to the "SIOC Core" ontology terms, there are also SIOC modules which can be used in provenance descriptions for datasets. The most relevant is the **SIOC Actions** [16] module, which was designed to represent how users in a community are manipulating the various digital artifacts that constitute the application supporting that community. The main terms in SIOC Actions are *sioca:Action*, *sioca:DigitalArtifact*, *sioca:modifies*, *sioca:creates*, *sioca:deletes*, *sioca:uses*, *sioca:object*, *sioca:product*, *sioca:source* and *sioca:byproduct*. These have been aligned to OPM and PROV-O in recent work by [59].

### 5.4. Dataset Links

Links as important features of Linked Data datasets are represented through a variety of means, covering both schema-level and entity-level links. VoID, for instance, includes specific linksets which can be instantiated to define metadata about a dataset's links. *SKOS*[41], the Simple Knowledge Organization System, on the other hand provides a formal vocabulary for defining taxonomic and mapping relations among both concepts and entities and is a well-used means to describe links between concepts and entities across datasets. By providing an established vocabulary for less strict relations, for instance,*broader* or *narrower*, respectively *broaderMatch* and *narrowerMatch*, it enables the representation of taxonomic relationships as well as the alignment of different schemas and knowledge bases, i.e. datasets.

A more specific approach is followed by the Vocabulary of Links (VoL)[42], which provides a general vocabulary to describe metadata about links or linksets, within or across specific datasets. VoL was designed specifically to represent additional metadata about computed links which cannot be expressed with default RDF(S) expressions and enable a qualification of a link or linkset. This includes, for instance, the description of linking scores or linking provenance, for instance, through a specific linking method.

The Expressive and Declarative Ontology Alignment Language (EDOAL)[43] enables the representation of correspondences between entities and concepts in different ontologies beyond mere mapping relationships (equivalence, subsumption). For these reasons, EDOAL introduces formalisms for representing transformations, constructions of complex classes/entities,

[41] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[42] http://data.linkededucation.org/vol/index.htm
[43] http://alignapi.gforge.inria.fr/edoal.html

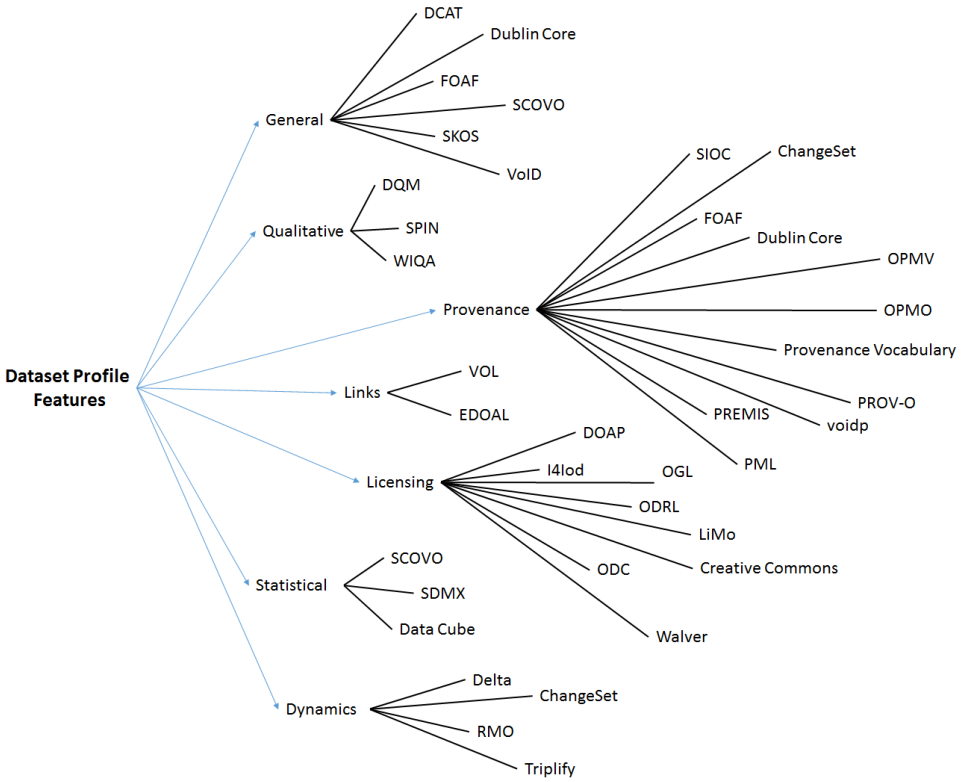


Fig. 4. Overview of relevant vocabularies as classified by type of dataset profile features. The figure is based on January 2017 statistics.

or restrictions to constrain classes/entities. EDOAL in that sense provides the means to provide on-the-fly interpretation of mapping statements as a part of data integration scenarios. On the other hand, in contrast to VoL, there are no means for representing the provenance of mapping statements.

### 5.5. Dataset Licensing

This section examines vocabularies available to assist with representing licenses of data and datasets. These include RDF versions of common licensing frameworks and alignments of multiple licensing frameworks into a combined vocabulary. In addition to the dedicated licensing vocabularies stated below, general resource metadata vocabularies provide basic features to indicate licensing information. This includes the *DCMI Metadata Terms*[44], featuring dedicated *license* and *rights* properties. These enable the association of arbitrary resources with a particular *LicenseDocument* or *RightsDocument*, however, without providing dedicated vocabularies for representing license information.

- **Creative Commons (CC)**[45] is a framework that allows users to define the rights regarding how others can reuse the content that the users themselves have published. It provides various licenses to define if and how people can reuse content that has been published, if they can modify it, and if it may be used for commercial purposes. Creative Commons also allows licensing information to be expressed in RDF using the ccREL (REL, or rights expression language) vocabulary[46]. Many datasets in the LOD cloud are already licensed under Creative Commons, as we will see later.
- The **Open Data Commons (ODC)** license[47] was originally released by Talis in 2008 as a means to tackle the issue of Creative Commons licenses being applied to non-creative resources such as data

[44] http://dublincore.org/documents/dcmi-terms

[45] https://creativecommons.org/licenses/

[46] https://wiki.creativecommons.org/wiki/CC_REL

[47] http://opendatacommons.org/licenses/

| Vocabulary Name | Type | Triples Feb. '15 | Datasets Feb. '15 | Triples Jan. '17 | Datasets Jan. '17 |
|---|---|---|---|---|---|
| Dublin Core | General, Provenance | 21,397,721 | 154 | 20,056,611 | **213** |
| FOAF | General, Provenance | 3,689,178 | 117 | 3,399,261 | **190** |
| SKOS | General | 10,581,530 | 67 | 5,606,905 | **108** |
| VoID | General | 9,754 | 41 | 987 | **53** |
| voidp | Provenance | 172 | 21 | 173 | 16 |
| PROV-O | Provenance | 482 | 20 | **577** | 17 |
| SIOC | Provenance | 148 | 16 | **6,255** | **45** |
| DOAP | Licensing | 306 | 14 | 53 | 7 |
| Creative Commons | Licensing | 16,525 | 12 | 83 | **21** |
| Provenance Vocabulary | Provenance | 84 | 12 | 61 | 2 |
| Data Cube | Statistical | 581,381 | 10 | 101,757 | **75** |
| SCOVO | General, Statistical | 408 | 9 | 399 | 1 |
| PML | Provenance | 259 | 8 | 0 | 0 |
| OPMO | Provenance | 63 | 8 | 4 | 1 |
| SDMX | Statistical | 285,904 | 6 | 90,586 | **11** |
| OPMV | Provenance | 4 | 2 | 1 | 1 |
| DCAT | General | 8 | 1 | **2,010** | **3** |
| Waiver | Licensing | 1 | 1 | 0 | 0 |
| Delta | Dynamics | 0 | 0 | 0 | 0 |
| RMO | Dynamics | 0 | 0 | 0 | 0 |
| Triplify | Dynamics | 0 | 0 | 0 | 0 |
| ChangeSet | Dynamics, Provenance | 0 | 0 | 0 | 0 |
| VoL | General | 0 | 0 | 0 | 0 |
| l4lod | Licensing | 0 | 0 | 0 | 0 |
| LiMo | Licensing | 0 | 0 | 0 | 0 |
| ODC | Licensing | 0 | 0 | 0 | 0 |
| ODRL | Licensing | 0 | 0 | 0 | 0 |
| OGL | Licensing | 0 | 0 | 0 | 0 |
| PREMIS | Provenance | 0 | 0 | 0 | 0 |
| DQM | Quality | 0 | 0 | 0 | 0 |
| SPIN | Quality | 0 | 0 | 0 | 0 |
| WIQA | Quality | 0 | 0 | 0 | 0 |

Table 3

Overall usage and dataset counts for the aforementioned vocabularies, sorted by number of datasets in February 2015. Those numbers in **boldface increased in 2017. Statistics were re-checked in January 2017.**

and datasets. The ODC "Public Domain Dedication and License" was a fusion of ideas from their earlier Talis Community License and related efforts such as the provision of scientific datasets using Science Commons.

– The **Open Digital Rights Language (ODRL)** vocabulary[48] enables the fine-grained specification of licensing terms (rights, policies, etc.) in a machine-readable format. Developed by the W3C ODRL Community Group, ODRL 2.0[49] uses RDF or JSON, evolving from an earlier XML-based REL version[50].

– **Open Government License (OGL)**[51] is a license produced specifically for Crown copyright works published by the UK government and other public sector bodies. It is aligned to both CC and

[48] http://www.w3.org/community/odrl/two/model/

[49] http://w3.org/ns/odrl/2/

[50] http://www.w3.org/TR/odrl/

[51] http://www.nationalarchives.gov.uk/doc/open-government-licence/

ODC. One of the dataset projects using OGL is the data.gov.uk service.

- The **License Model (LiMo)**[52] is an ontology for open data and dataset licensing. It links to terms from Dublin Core, VoID, CC and PROV-O, and also defines legal terms, conditions of use and distribution, and other rights. One of the main terms is *limo:LicenseModel* which is equivalent to the *cc:License* concept from Creative Commons.
- **Description of a Project (DOAP)**[53] is an RDF vocabulary that provides a common metadata modelling scheme for describing projects creating software applications, in order to provide a unified way to represent a software project no matter the source. The main class is *Project* which has properties such as its licence, the project's maintainers, the URL for subversion access, etc. Many of the concepts in DOAP could also be reapplied to datasets since they share many of the same properties.
- **Licenses for Linked Open Data (l4lod)**[54] was introduced in [34] to provide an alignment with many of the licensing vocabularies we have just described. It can be used to express a machine-readable composite license for a dataset. l4lod is composed of three deontic components (obligations, permissions and prohibitions) that can be used to reconcile a set of licenses that are associated with heterogeneous datasets whose information items have been returned together for consumption (e.g., via a single SPARQL query).

### 5.6. Statistical Dataset Metadata

A range of vocabularies exist, which partially support the representation of dataset statistics and can be used in conjunction with general dataset metadata vocabularies such as VoID or DCAT. These include, for instance, the RDF Data Cube vocabulary[55], SDMX[56] or SCOVO[57].

The VoID guidelines, for instance, recommend the use of SCOVO to share statistical dataset features [2]. There can be statistics concerning the whole dataset or linkset, such as triple count, and others attributing statistics to a source, to capture where a statistical datum stems from. SCOVO, also described by Hausenblas et al. [41], is an earlier, native RDF vocabulary for statistical data, consisting of three main classes, *Dataset*, *Dimension*, and *Item*. While there exist efforts to merge SCOVO and SDMX-RDF [17], both approaches are superseded by the Data Cube vocabulary, which represents the state of the art in representing statistical data on the Web.

The RDF Data Cube vocabulary[58], currently a W3C Editors Draft developed by the Government Linked Data Working Group[59] is an RDF vocabulary for representing multi-dimensional so-called *data cubes* in RDF. The Data Cube vocabulary describes general statistical notions, such as *dimensions* or *observations*, and as such, can be perceived as a meta-level vocabulary for representing any statistical notion.

While the Data Cube vocabulary builds on SKOS, its Data Cubes approach originates from and is compatible with the cube structure underlying the SDMX (Statistical Data and Metadata eXchange) information model. The latter is an ISO standard, describing an information model for exchanging statistical data and metadata which has been serialised into XML, EDI and recently, RDF. SDMX-RDF[60] can be seen as a natural predecessor of the Data Cube vocabulary which is not a one-to-one representation of SDMX but uses an SDMX subset, plus additional elements, to provide a vocabulary tailored to represent data published as RDF on the Web.

Auer et al. present LODStats [4], a framework for dataset analytics, which introduces a set of 32 statistical features and uses the most recommended combination of VoID and the Data Cube vocabulary. Links between the Data Cube class *qb:Observation* and the *void:Dataset* class are represented using a native property (*void-ext:observation*). While VoID already represents properties for several statistically described objects (triples, classes, distinctSubjects, etc.), additional features were represented using *void:classPartition* and *void:propertyPartition*. While this approach combines two state of the art vocabularies for general dataset metadata (VoID) and statistical data (Data

[52] http://purl.org/LiMo/0.1
[53] http://usefulinc.com/ns/doap
[54] http://ns.inria.fr/l4lod/
[55] http://www.w3.org/TR/vocab-data-cube
[56] http://sdmx.org
[57] http://vocab.deri.ie/scovo
[58] https://dvcs.w3.org/hg/gld/raw-file/default/data-cube/index.html
[59] http://www.w3.org/2011/gld/
[60] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html

Cube), it turns out to be quite a future-proof approach to capture statistical dataset metadata.

### 5.7. Dataset Dynamics

While there does exist a wealth of methods for assessing characteristics related to the dynamics and evolution of datasets, as illustrated in earlier sections of this survey, most vocabularies in this area are dedicated to representing the actual evolution of a dataset, rather than higher level observations about dynamics.

The Dataset Dynamics group[61] for instance lists a number of vocabularies for representing dataset changesets and updates. The *Talis Changeset vocabulary*[62] provides some early, yet discontinued work on representing changesets and specific characteristics, and has a similar approach as the Delta vocabulary[63]. The *Triplify Update vocabulary*[64] provides a very simple RDF schema for capturing dataset updates where each *Update* or *UpdateSet* is annotated with provenance information about the updater and the time stamp.

In a similar direction is the recent work of Graube et al. [35] on *R43ples*, a revision management approach for RDF datasets using named graphs for capturing revisions and SPARQL for manipulation of the latter. Authors introduce the so-called Revision Management Ontology (RMO) based on PROV-O (Section 5.3). While RMO implements baseline revision management notions for data graphs, it is of lesser relevance for the purposes of this section.

A more abstract approach is offered by the *Dataset Dynamics (DaDy) vocabulary*[65], which allows the representation of more abstract dynamics-related observations for a specific dataset. It is specifically foreseen to be used in conjunction with VoID, where a *void:Dataset* is annotated with instantiations of *dady:UpdateDynamics*. The latter captures information about the update regularity and frequency.

For capturing specific features and observations related to dynamics and evolution, beyond the ones covered by the vocabularies above, some of the vocabularies mentioned in Section 5.6 (for representing statistical dataset features) may also be closely related to dynamics.

[61] http://www.w3.org/wiki/DatasetDynamics
[62] http://vocab.org/changeset/schema.html
[63] http://www.w3.org/2004/delta
[64] http://triplify.org/vocabulary/update
[65] http://vocab.deri.ie/dady

### 5.8. Observations on Vocabulary Adoption

We use the LOD2 Stats service[66] to give us some context as to how often terms from these vocabularies are being used and within how many datasets. These statistics are shown in Table 3, where the type refers to the vocabulary type as per the headings above.[67]

While our quantitative assessment indicates mere usage of a particular vocabulary, it does not provide any insights into the way a vocabulary has been used in particular scenarios. In addition, it is worth noting that particular features, for instance, license information, are often represented through a variety of means, which may not be captured by the vocabularies identified here [69] [43].

While we were unable to filter the instances of dataset profiling-specific terms from our suggested vocabularies while examining their usage statistics in LOD2, we can gain some insight into which ones may be more widely adopted by looking at the existing overall statistics and dataset usages, especially over time, i.e., from 2015 to 2017, we can see which vocabularies are consistently being used and are growing in usage. It is reasonable to assume that users will be more willing to adopt terms from widely-used vocabularies for representing dataset profiles, as long as they are fit for purpose. We note that for many of the vocabularies that have changing numbers of datasets and triples over time, there can be somewhat conflicting numbers (e.g. for SKOS, VoID, etc. where the number of datasets increases but the number of triples decreases, sometimes by an order of magnitude). We consider that this can be explained by the removal of a particular dataset/website that has a particularly high number of triples of a particular type, or by the adoption of a new vocabulary/removal of a particular vocabulary for a set of triples on a website.

251 datasets used RDF syntax in 2015 (increasing to 1,718 in 2017), giving us an overall total. From the data in Table 3, we observe that general metadata about the datasets is readily provided, but that more specific information on provenance and statistics using specialised vocabularies is only available in some-

[66] http://stats.lod2.eu/ as accessed on 2nd February 2015 and re-checked again on 19 January 2017

[67] Where multiple entries exist for a vocabulary on LOD2 Stats, we use the numbers from the largest entry rather than adding usage figures together, as modules in a vocabulary may be used together in the same dataset, e.g., DC Terms and DC Elements, or SDMX Dimension and SDMX Measure).

where around 22% (55) and 10% (25) of datasets in 2015 respectively (in 2017, these numbers reduced to 2% (36) and 5% (87) for provenance and statistics respectively).

Another observation is that none of the quality or dynamics and evolution vocabularies appear in LOD2 Stats. That points to a significant under-utilization of terms relating to dataset quality, the evolution of a dataset, or the dynamics involved in a changing dataset. The assumption is that dataset creators are more interested in providing the datasets themselves without giving assurances to others who may want to use them about their quality or how they have changed over time.

It does not seem from Table 3 that many datasets are explicitly licensed via some machine-readable form, with just 5% (12) in 2015 and 1% (21) in 2017 containing Creative Commons metadata. However, according to work by [34], 95% of the datasets in the LOD cloud[68] did indeed express licensing information via the *dcterms:license* or the *dcterms:rights* properties of Dublin Core (albeit in human-readable format). Creative Commons represented 51% of all licenses in their analysis, followed by Open Data Commons at 18%. This points to the need for more explicit license definitions in datasets, with a link to the license type and conditions and not just a simple text string in an attribute field.

## 6. Application-Driven Dataset Profiles

Dataset profiles are highly important for a wide variety of cross-domain applications, for example, data linking and curation, schema inference, federated query and search, as well as question answering. In this section, we highlight important applications from these domains that use dataset profiles along with their relevant profile features. Some of these applications can use, verify and update dataset profile features (e.g., including statistical characteristics of datasets) and may in turn generate additional statistics that can become part of the dataset profile. The list of the applications and relevant features presented in this section aims to illustrate the use of dataset profiles by state-of-the-art tools and is not exhaustive.

[68] http://lod-cloud.net/

### 6.1. Data Linking Applications

Data linking applications aim to annotate, disambiguate and interlink entities and events in text using Natural Language Processing (NLP) techniques and external sources including Linked Data. In this context, popular services include DBpedia Spotlight [18], Illinois Wikifier [66] as well as Babelfy [54].

*Example features for data linking applications:* Data linking applications typically use the general features discussed in Section 3.1 such as topics, domains, as well as representative schema elements and instances.

### 6.2. Data Curation, Cleansing and Maintenance

As linked datasets are often generated from semi-structured or unstructured sources using automated extraction approaches, these datasets vary heavily with respect to quality, currentness and completeness of the contained information [84].

A number of recent works focus on statistical methods for: (1) outlier detection to detect errors in numerical values [27,62,81]; (2) automatic prediction of missing types of instances [62]; and (3) the identification of incorrect links between datasets [61]. A further line of research in Linked Data quality is related to the discovery of errors in the data based on existing interlinkings (e.g., [13,83]). Thereby some works go beyond error detection and attempt to automatically determine correct data values in case of inconsistencies [13]. As mentioned above, additional statistics generated by these approaches that can become part of the dataset profile.

*Example features for error detection in numerical values:* In [27] the authors detect errors in numerical values using outlier detection. To identify the properties to which numerical outlier detection can be applied, the following statistical characteristics (discussed in Section 3.6) are used: (1) total number of instances, (2) names of the properties used in the dataset, (3) frequency of usage with numerical values in the object position for each property, and (4) total number of distinct numerical values for each property.

*Example features for conflict resolution in multilingual DBpedia:* The features used in conflict resolution in [13] include provenance metadata at the statement, property and author levels. The temporal dataset profile includes in particular: (1) Recency of the specific statement (measured using the time of the last edit); (2) Overall editing frequency of the property in the

dataset; and (3) The overall number of edits performed by the specific editor.

### 6.3. Schema Inference

Many existing Linked Data sources do not explicitly specify schemas, or only provide incomplete specifications. However, many real-world applications (e.g., answering queries over distributed data [9]) rely on the schema information. Recently, approaches aimed at the automatic inference of missing schema information have been developed (e.g., [62,48]).

*Example features for type inference:* Statistical characteristics of datasets (see Section 3.6) play an important role in type inference applications. For example, in [62] statistics on the completeness of type statements as well as property-specific type distributions are required (i.e. the types of resources appearing in subject and object positions of each property including their frequencies).

### 6.4. Distributed Query Applications

The Linked Data Cloud can be queried either through direct HTTP URI lookups or using distributed SPARQL endpoints [36] that can include full-text search extensions (see e.g., [57]). Also combinations of both query paradigms are possible [39]. Typically, the first step of query answering over distributed data is the generation of ordered query plans against the mediated schema on a number of data sources [82]; In this step, dataset profiling plays an important role.

In order to guide distributed query processing, existing applications rely on indexes of varying granularity including *Schema-level Indexes* and *Data Summaries*. *Schema-level Indexes* contain information about properties and classes occurring at certain sources. *Data Summaries* use a combined description of instance- and schema-level elements to summarise the content of data sources [36]. The majority of existing federated query approaches for LOD (e.g., [39,36,79,32]) aim to optimize efficient query processing and do not (yet) take the quality parameters of LOD sources into account. Therefore, existing *Data Summaries* mostly contain frequencies and interlinking statistics of varying granularity.

*Example features for efficient and quality-aware query applications:* The majority of existing query applications rely on general and statistical characteristics (see Sections 3.1 and 3.6) at the schema-level, i.e. properties and classes occurring at certain sources for effective query interpretation. In addition, applications that optimize for efficient query processing require data-level statistics (including frequency and interlinking) either on triple level or for each subject, object and predicate individually [36]. Finally, quality-aware query applications also take into account qualitative characteristics (see Section 3.2) (e.g., completeness and accuracy) at different granularity levels. This includes overall data source statistics [55], as well as property-specific [68] and type-specific statistics [82].

### 6.5. Information Retrieval (IR) Applications

In IR, Linked Data is mostly used in the context of semantic search, a typical demonstration of which can be found in [25]. The majority of semantic search applications are domain-oriented; a large number of practical cases have been shown for repositories related to biomedical sciences. For example, the concept-based search mechanism [50] allows biologists to describe the topics of interest in a search more specifically and retrieve information with higher precision (in comparison to the usage of keywords only). It should be stressed here that concept-based search requires linking to high-quality external resources (such as, e.g., Unified Medical Language System – UMLS), which involves features related to trust, especially verifiability and believability.

Datasets providing semantic features enable us to go beyond the standard bag of words representation [75]. A wide range of methods based on linking to external, domain-oriented resources has been proposed, e.g., [66]. They also employ statistical features extracted from large-scale text corpora and allow one to expand the user queries to increase recall [5]. In addition, geographical and temporal contexts play an increasingly important role in IR applications. These contexts enable the retrieval of information that is relevant with respect to the spatial [44] and temporal [14] dimensions of the query.

*Example features for Information Retrieval applications:* IR involves qualitative profile features related to trust (i.e., verifiability and believability) and the accessibility of data. In addition, to facilitate semantic search, IR implies general profile features like topical domains and context.

### 6.6. Discussion

Overall, we observe that although existing applications make use of the whole spectrum of the dataset

profile feature categories, including general, qualitative, statistical and dynamics features discussed in this survey, the concrete set of features is application-dependent and the whole set is rarely used within any single application. Whereas some applications rely on existing metadata, many applications compute dataset profile features as part of their own processing pipelines. These applications can thus directly contribute to the dataset profile generation.

## 7. Summary and Conclusions

The availability of dataset profiles has the potential to improve data discovery and reuse on the Web. Remaining challenges and obstacles include the lack of Web-scale adoption of general standards, e.g., for representing profile features, and the lack of automated means for interpreting and using profile information as part of large-scale data reuse scenarios. This survey hence aims at raising the awareness and uptake of profiling techniques and vocabularies.

In this survey, we provided a comprehensive overview of dataset profiling features, methods, tools, vocabularies and applications. Given the complexity of the topic, we first focused on organizing the different dataset profile features in a taxonomy. We then provided a systematic overview of a large set of approaches and tools for assessing and extracting such features from RDF datasets. We reviewed the vocabularies for representing these features, preferably as Linked Data, and finally we discussed several prominent applications of dataset profiles.

Wherever feasible, we also provided insights into the adoption and impact of the discussed works; for instance, based on the profile extraction tools distribution in the provided taxonomy, we proposed that certain profile features, notably in the general category, should be provided by domain experts to ensure high quality profiles. Another observation concerned the vocabulary usage where some features, such as the quality or the dynamicity of vocabularies did not appear in the evaluated statistics. That led us to recommend that dataset providers need to guarantee a high confidence with respect to these profile features in order to ensure better access to their quality and dynamics.

We observed that although existing applications made use of the whole spectrum of the discussed feature categories, including general, qualitative, statistical and temporal features, the concrete set of features was application-dependent and the whole set was rarely used within any single application. Furthermore, we discussed the fact that many applications generated dataset profile features as a part of their own processing pipelines.

Finally, we strongly recommended that dataset profiles should provide representations readable for both humans and machines to open up the Web of Data to a wider variety of users and applications. Given the continuous evolution and expansion of the Web of Data, we believe that the problem of dataset profiling will become an even more prominent one, and corresponding methods will form a crucial building block for enabling the reuse and take-up of datasets beyond established and well-understood knowledge bases and reference graphs.

## 8. Acknowledgements

This paper was partially supported by COST (European Cooperation in Science and Technology) under Action IC1302 (KEYSTONE), Science Foundation Ireland under Grant Number SFI/12/RC/2289 (INSIGHT), the German Federal Ministry of Education and Research (BMBF) under Data4UrbanMobility (02K15A040), the Datalyse project[69] (FSN-AAP Big Data n3), and the European Research Council under ALEXANDRIA (ERC 339233).

## References

[1] Ziawasch Abedjan, Toni Grütze, Anja Jentzsch, and Felix Naumann. Profiling and mining RDF data with prolod++. In *Proceedings of the 30th International Conference on Data Engineering, Chicago, ICDE 2014, IL, USA, March 31 - April 4, 2014*, pages 1198–1201, 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, 2009.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International Semantic Web Conference*, pages 722–735, 2007.

[4] Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. Lodstats - an extensible framework for high-performance dataset analytics. In *Proceedings of the 18th International Conference on Knowledge Engineering and Knowledge Management EKAW 2012, Galway City, Ireland, October 8-12, 2012.*, pages 353–362, 2012.

[69] http://www.datalyse.fr/

[5] Jagdev Bhogal, Andy Macfarlane, and Peter Smith. A review of ontology based query expansion. *Information processing & management*, 43(4):866–886, 2007.

[6] Christian Bizer. *Quality-Driven Information Filtering in the Context of Web-Based Information Systems.* PhD thesis, Freie Universitat, Berlin, March 2007.

[7] Christian Bizer and Richard Cyganiak. Quality-driven information filtering using the WIQA policy framework. *J. Web Sem.*, 7(1):1–10, 2009.

[8] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[9] Jens Bleiholder and Felix Naumann. Data fusion. *ACM Comput. Surv.*, 41(1):1:1–1:41, January 2009.

[10] Jim Blythe and Yolanda Gil. Incremental formalization of document annotations through ontology-based paraphrasing. In *Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004*, pages 455–461, 2004.

[11] Christoph Böhm, Johannes Lorey, and Felix Naumann. Creating void descriptions for web-scale data. *J. Web Sem.*, 9(3):339–345, 2011.

[12] Christoph Böhm, Felix Naumann, Ziawasch Abedjan, Dandy Fenz, Toni Grütze, Daniel Hefenbrock, Matthias Pohl, and David Sonnabend. Profiling linked open data with prolod. In *Workshops Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*, pages 175–178, 2010.

[13] Volha Bryl and Christian Bizer. Learning conflict resolution strategies for cross-language wikipedia data fusion. In *Proceedings of the 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, pages 1129–1134, 2014.

[14] Ricardo Campos, Gaël Dias, Alípio M Jorge, and Adam Jatowt. Survey of temporal information retrieval and related applications. *ACM Computing Surveys (CSUR)*, 47(2):15, 2014.

[15] Jeremy J. Carroll, Christian Bizer, Patrick J. Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web, WWW 2005, Chiba, Japan, May 10-14, 2005*, pages 613–622, 2005.

[16] Pierre-Antoine Champin and Alexandre Passant. SIOC in action representing the dynamics of online communities. In *Proceedings the 6th International Conference on Semantic Systems, I-SEMANTICS 2010, Graz, Austria, September 1-3, 2010*, 2010.

[17] Richard Cyganiak, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. Semantic statistics: Bringing together SDMX and SCOVO. In *Proceedings of the WWW 2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, 2010.

[18] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *Proceedings of the 9th International Conference on Semantic Systems, I-SEMANTICS 2013, Graz, Austria, September 4-6, 2013*, pages 121–124, 2013.

[19] Herbert Van de Sompel, Robert Sanderson, Michael L. Nelson, Lyudmila Balakireva, Harihar Shankar, and Scott Ainsworth. An http-based versioning mechanism for linked data. In *Proceedings of the WWW 2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, 2010.

[20] Jeremy Debattista, Christoph Lange, and Sören Auer. daQ, an Ontology for Dataset Quality Information. In *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014.*, 2014.

[21] Elena Demidova, Stefan Dietze, Julian Szymanski, and John Breslin (Eds.). Special Issue on Dataset Profiling and Federated Search for Linked Data. Editorial. *The International Journal on Semantic Web and Information Systems (IJSWIS)*, 12(3), 2016.

[22] Renata Queiroz Dividino, Ansgar Scherp, Gerd Gröner, and Thomas Grotton. Change-a-lod: Does the schema on the linked data cloud change or not? In *Proceedings of the Fourth International Workshop on Consuming Linked Data, COLD 2013, Sydney, Australia, October 22, 2013*, 2013.

[23] Mohamed Ben Ellefi, Zohra Bellahsene, Stefan Dietze, and Konstantin Todorov. Dataset recommendation for data linking: An intensional approach. In *Proceedings of the 13th International Conference, ESWC 2016, Heraklion, Crete, Greece, May 29 - June 2, 2016, Proceedings*, pages 36–51, 2016.

[24] Mohamed Ben Ellefi, Zohra Bellahsene, François Scharffe, and Konstantin Todorov. Towards semantic dataset profiling. In *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, Crete, Greece, May 26, 2014.*, 2014.

[25] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. Semantically enhanced information retrieval: an ontology-based approach. *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(4):434–452, 2011.

[26] Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. A scalable approach for efficiently generating structured dataset topic profiles. In *The Semantic Web: Trends and Challenges - Proceedings of the 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014.*, pages 519–534, 2014.

[27] Daniel Fleischhacker, Heiko Paulheim, Volha Bryl, Johanna Völker, and Christian Bizer. Detecting errors in numerical linked data using cross-checked outlier detection. In *Proceedings of the 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Part I*, pages 357–372, 2014.

[28] Benedikt Forchhammer, Anja Jentzsch, and Felix Naumann. LODOP - multi-query optimization for linked data profiling queries. In *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, PROFILES@ESWC 2014, Anissaras, Crete, Greece, May 26, 2014.*, 2014.

[29] Christian Fürber and Martin Hepp. Using semantic web resources for data quality management. In *Proceedings of the 17th International Conference on Knowledge Engineering and Management by the Masses*, EKAW'10, pages 211–225, 2010.

[30] Christian Fürber and Martin Hepp. Towards a vocabulary for data quality management in semantic web architectures. In *Proceedings of the 1st International Workshop on Linked Web Data Management*, LWDM '11, pages 1–8, New York, NY, USA, 2011. ACM.

[31] Yolanda Gil and Varun Ratnakar. TRELLIS: an interactive tool for capturing information analysis and decision making. In *Proceedings of the 13th International Conference, EKAW 2002, Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web, Siguenza, Spain, October 1-4 2002*, pages 37–42, 2002.

[32] Olaf Görlitz and Steffen Staab. Splendid: Sparql endpoint federation exploiting void descriptions. In *Proceedings of the Second International Workshop on Consuming Linked Data (COLD2011), Bonn, Germany, October 23, 2011*, 2011.

[33] Thomas Gottron and Christian Gottron. Perplexity of index models over evolving linked data. In *The Semantic Web: Trends and Challenges - Proceedings of the 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014.*, pages 161–175, 2014.

[34] Guido Governatori, Antonino Rotolo, Serena Villata, and Fabien Gandon. One license to compose them all - A deontic logic approach to data licensing on the web of data. In *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I*, pages 151–166, 2013.

[35] Markus Graube, Stephan Hensel, and Leon Urbas. R43ples: Revisions for triples - an approach for version control in the semantic web. In *Proceedings of the 1st Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany, September 2nd, 2014.*, 2014.

[36] Andreas Harth, Katja Hose, Marcel Karnstedt, Axel Polleres, Kai-Uwe Sattler, and Jürgen 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, New York, NY, USA, 2010. ACM.

[37] Olaf Hartig. Trustworthiness of data on the web. In *Proceedings of the STI Berlin & CSW PhD Workshop*, 2008.

[38] Olaf Hartig. Provenance information in the web of data. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Kingsley Idehen, editors, *LDOW*, volume 538 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2009.

[39] Olaf Hartig, Christian Bizer, and Johann Christoph Freytag. Executing sparql queries over the web of linked data. In *Proceedings of the 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, pages 293–309, 2009.

[40] Jens Hartmann, York Sure, Peter Haase, Raul Palma, and Mari del Carmen Suárez-Figueroa. OMV – Ontology Metadata Vocabulary. In *Proceedings of the Ontology Patterns for the Semantic Web Workshop*, Galway, Ireland, 2005.

[41] Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. SCOVO: using statistics on the web of data. In *The Semantic Web: Research and Applications, Proceedings of the 6th European Semantic Web Conference, ESWC 2009, Heraklion, Crete, Greece, May 31-June 4, 2009.*, pages 708–722, 2009.

[42] Tom Heath, Michael Hausenblas, Chris Bizer, Richard Cyganiak, and Olaf Hartig. How to publish linked data on the web. In *Tutorial in the 7th International Semantic Web Conference, Karlsruhe, Germany*, 2008.

[43] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. An empirical survey of linked data conformance. *J. Web Sem.*, 14:14–44, 2012.

[44] Christopher B Jones, Harith Alani, and Douglas Tudhope. Geographical information retrieval with ontologies of place. In *Spatial information theory*, pages 322–335. Springer, 2001.

[45] Tobias Käfer, Ahmed Abdelrahman, Jürgen Umbrich, Patrick O'Byrne, and Aidan Hogan. Observing linked data dynamics. In *The Semantic Web: Semantics and Big Data, Proceedings of the 10th International Conference, ESWC 2013, Montpellier, France, May 26-30, 2013.*, pages 213–227, 2013.

[46] Tobias Käfer, Jürgen Umbrich, Aidan Hogan, and Axel Polleres. Dyldo: Towards a dynamic linked data observatory. In *WWW2012 Workshop on Linked Data on the Web, Lyon, France, 16 April, 2012*, 2012.

[47] Shahan Khatchadourian and Mariano P. Consens. Explod: Summary-based exploration of interlinking and RDF usage in the linked open data cloud. In *Proceedings of the 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010, Part II*, pages 272–287. Springer, 2010.

[48] Mathias Konrath, Thomas Gottron, Steffen Staab, and Ansgar Scherp. Schemex - efficient construction of a data catalogue by stream-based indexing of linked data. *J. Web Sem.*, 16:52–58, 2012.

[49] Dimitris Kontokostas, Patrick Westphal, Sören Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen, and Amrapali Zaveri. Test-driven evaluation of linked data quality. In *Proceedings of the 23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 747–758, 2014.

[50] Bevan Koopman, Peter Bruza, Laurianne Sitbon, and Michael Lawley. Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. *The Australasian medical journal*, 5(9):482, 2012.

[51] Fadi Maali, Richard Cyganiak, and Vassilios Peristeras. Enabling interoperability of government data catalogues. In *Electronic Government, 9th IFIP WG 8.5 International Conference, EGOV 2010, Lausanne, Switzerland, August 29 - September 2, 2010. Proceedings*, pages 339–350, 2010.

[52] Paolo Missier, Khalid Belhajjame, and James Cheney. The W3C PROV family of specifications for modelling provenance metadata. In *Joint 2013 EDBT/ICDT Conferences, EDBT '13 Proceedings, Genoa, Italy, March 18-22, 2013*, pages 773–776, 2013.

[53] Luc Moreau. The Foundations for Provenance on the Web. *Foundations and Trends in Web Science*, 2(2-3):99–241, 2010.

[54] Andrea Moro, Alessandro Raganato, and Roberto Navigli. Entity linking meets word sense disambiguation: a unified approach. *TACL*, 2:231–244, 2014.

[55] Felix Naumann. *Quality-driven Query Answering for Integrated Information Systems*. Springer-Verlag, Berlin, Heidelberg, 2002.

[56] Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres. Automated quality assessment of metadata across open data portals. *J. Data and Information Quality*, 8(1):2:1–2:29, October 2016.

[57] Andriy Nikolov, Andreas Schwarte, and Christian Hütter. Fedsearch: Efficiently combining structured queries and full-text search in a SPARQL federation. In *Proceedings of the 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Part I*, pages 427–443, 2013.

[58] Tope Omitola, Landong Zuo, Christopher Gutteridge, Ian C. Millard, Hugh Glaser, Nicholas Gibbins, and Nigel Shadbolt.

Tracing the provenance of linked data using void. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 17:1–17:7, New York, NY, USA, 2011. ACM.

[59] Fabrizio Orlandi. *Profiling user interests on the social semantic web*. PhD thesis, National University of Ireland Galway, 2014.

[60] Alexandre Passant and Pablo N. Mendes. sparqlPuSH: Proactive Notification of Data Updates in RDF Stores Using PubSubHubbub. In *Proceedings of the Sixth Workshop on Scripting and Development for the Semantic Web, Crete, Greece, May 31, 2010*, 2010.

[61] Heiko Paulheim. Identifying wrong links between datasets by multi-dimensional outlier detection. In *Proceedings of the Third International Workshop on Debugging Ontologies and Ontology Mappings, WoDOOM 2014, co-located with 11th Extended Semantic Web Conference (ESWC 2014), Anissaras/Hersonissou, Greece, May 26, 2014.*, pages 27–38, 2014.

[62] Heiko Paulheim and Christian Bizer. Improving the quality of linked data using statistical distributions. *Int. J. Semantic Web Inf. Syst.*, 10(2):63–86, 2014.

[63] Leo Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, 2002.

[64] Niko Popitsch and Bernhard Haslhofer. Dsnotify: handling broken links in the web of data. In *Proceedings of the 19th International Conference on World Wide Web, WWW 2010, Raleigh, North Carolina, USA, April 26-30, 2010*, pages 761–770, 2010.

[65] Filip Radulovic, Nandana Mihindukulasooriya, Raúl García-Castro, and Asunción Gómez-Pérez. A comprehensive quality model for linked data. *Semantic Web Journal*, Preprint:1–22, 2017.

[66] Lev-Arie Ratinov, Dan Roth, Doug Downey, and Mike Anderson. Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 19-24 June, 2011, Portland, Oregon, USA*, pages 1375–1384, 2011.

[67] Edna Ruckhaus, Maria-Esther Vidal, Simón Castillo, Oscar Burguillos, and Oriana Baldizan. Analyzing linked data quality with liquate. In *Proceedings of the Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pages 488–493, 2014.

[68] Monica Scannapieco, Antonino Virgillito, Carlo Marchetti, Massimo Mecella, and Roberto Baldoni. The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems. *Information Systems*, 29(7):551 – 582, 2004. Data Quality in Cooperative Information Systems.

[69] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the linked data best practices in different topical domains. In *Proceedings of the 13th International Semantic Web Conference - Part I*, ISWC '14, pages 245–260, New York, NY, USA, 2014. Springer-Verlag New York, Inc.

[70] Tommaso Soru, Edgard Marx, and Axel-Cyrille Ngonga Ngomo. ROCKER: A refinement operator for key discovery. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 1025–1033, 2015.

[71] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706, 2007.

[72] Kaustubh Supekar, Chintan Patel, and Yugyung Lee. Characterizing quality of knowledge on semantic web. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*, pages 472–478, 2004.

[73] Danai Symeonidou, Vincent Armant, Nathalie Pernelle, and Fatiha Saïs. Sakey: Scalable almost key discovery in RDF data. In *Proceedings of the 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Part I*, pages 33–49, 2014.

[74] Danai Symeonidou, Nathalie Pernelle, and Fatiha Saïs. KD2R: A key discovery method for semantic reference reconciliation. In *Proceedings on the Move to Meaningful Internet Systems: OTM 2011 Workshops - Confederated International Workshops and Posters: EI2N+NSF ICE, ICSP+INBAST, ISDE, ORM, OTMA, SWWS+MONET+SeDeS, and VADER 2011, Hersonissos, Crete, Greece, October 17-21, 2011.*, pages 392–401, 2011.

[75] Julian Szymański. Comparative analysis of text representation methods using classification. *Cybernetics and Systems*, 45(2):180–199, 2014.

[76] Sebastian Tramp, Philipp Frischmuth, Timofey Ermilov, Saeedeh Shekarpour, and Sören Auer. An architecture of a distributed semantic social network. *Semantic Web*, 5(1):77–95, 2014.

[77] Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. Quality assessment and evolution of open data portals. In *Proceedings of the 3rd International Conference on Future Internet of Things and Cloud, FiCloud 2015, Rome, Italy, August 24-26, 2015*, pages 404–411, 2015.

[78] Andreas Wagner, Peter Haase, Achim Rettinger, and Holger Lamm. Entity-based data source contextualization for searching the web of data. In *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data co-located with the 11th Extended Semantic Web Conference, Crete, Greece*, 2014.

[79] Andreas Wagner, Duc Thanh Tran, Günter Ladwig, Andreas Harth, and Rudi Studer. Top-k linked data query processing. In *Proceedings of the 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, pages 56–71, 2012.

[80] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems*, 12(4):5–33, 1996.

[81] Dominik Wienand and Heiko Paulheim. Detecting incorrect numerical data in dbpedia. In *Proceedings of the 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014.*, pages 504–518, 2014.

[82] Naiem K. Yeganeh, Shazia Sadiq, and Mohamed A. Sharaf. A framework for data quality aware query systems. *Inf. Syst.*, 46:24–44, December 2014.

[83] Wancheng Yuan, Elena Demidova, Stefan Dietze, and Xuan Zhou. Analyzing relative incompleteness of movie descriptions in the web of data: A case study. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014.*, pages 197–200, 2014.

[84] Amrapali Zaveri, Dimitris Kontokostas, Mohamed Ahmed Sherif, Lorenz Bühmann, Mohamed Morsey, Sören Auer, and Jens Lehmann. User-driven quality evaluation of dbpedia. In *Proceedings of the 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 97–104, 2013.

[85] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.

RESUBMISSION of **1606-2818**

*"**RDF** Dataset Profiling - a Survey of Features, Methods, Vocabularies and Applications"*

*By authors:*
*Mohamed Ben Ellefi[a], Zohra Bellahsene[a], John G. Breslin[b], Elena Demidova[c], Stefan Dietze[c], Julian Szymanski[d] and Konstantin Todorov[a]*
*[a]{benellefi, bella, todorov}@lirmm.fr*
*[b]{john.breslin@nuigalway.ie}*
*[c]{demidova, dietze}@L3S.de*
*[d]{julian.szymanski@eti.pg.gda.pl}*

*Submitted to "Semantic Web Journal"*
*URL:*
*http://www.semantic-web-journal.net/content/rdf-dataset-profiling-survey-features-methods-applications-and-vocabularies*

We would like to thank the reviewers and the editor for their effort and time and their detailed, insightful and constructive comments. We have revised the survey according to these comments. As a result, all sections have been subjected to structural and/or content changes, summarized as follows.

- The survey methodology in Section 2 has been reworked by extending the bibliographical entries categories, improving presentation and double-checking reference entries numbers per category. Also, a chart for the number of entries per year has been added as represented in Figure 2.
- The feature categories in Section 3 have been revised by providing clearer definitions and renaming certain features in order to avoid ambiguity.
- Sections 3 to 5 have been restructured to follow and reflect the survey's feature and vocabulary taxonomies as represented in Figures 3 and 4, respectively.
- Section 4 has been modified accordingly, all feature extraction frameworks have been double-checked to ensure the correctness of their respective category assignment.
- Section 5 has been revised to ensure compliance with the feature definitions and taxonomy in Section 3, and to improve overall balance between the discussed vocabularies. In addition, new references were added to support particular claims and expand several vocabulary discussions (e.g., DQV). Outdated references were replaced, the vocabulary adoption figures were double-checked and updated where necessary.
- Section 6 has been adjusted to address the review comments, in particular through re-phrasing the parts criticised in the review, and adjusting the references.

- Finally, the paper has been jointly revised by all co-authors in order to improve clarity and accuracy of language, definitions, and terminology, as well as to ensure the overall consistency of the presentation.

Please find below our detailed replies containing precise information on the revisions made according to the specific comments of the reviewers.

---

**Reviewer R1:**
**The revision addresses many of the previous comments. For this survey article the body of knowledge is more or less given now, and the presentation and motivation have improved. No doubt, one could identify many items to improve, but I believe that the text as such could be accepted. There are still a few minor textual aspects to repair, e.g. lines running over the margin, line spacing at the beginning of 3.7.**

**Response:** Thank you for your feedback. We have proof-read and revised the paper throughout and addressed the mentioned issues.

---

**Reviewer R2:**

**I have read a previously submitted version of this article. The breadth of the work surveyed was impressive but I had rather negative assessment of it. The new version is much better. The updates in the taxonomy, the removal of the authors' RDF vocabulary work, the extra explanations in e.g. section 4 and 6, are helpful.**
**There are some points that must be fixed still, see below. I would conjure the senior authors of the paper to have a triple check before resubmission, as some of the comments (and the most notable ones) apply to the parts that have been most heavily re-worked, thus raising some doubts about the seriousness of the writing process. As a small but revealing example, references [5] and [6] are identical. How come?!?**
**It is only because I have seen how the authors have changed their paper after previous comments that I am ready to trust they can make these final enhancements and produce an acceptable survey paper.**

**Response:** Thank you for the comment. As suggested by this reviewer, all authors, in particular the more senior ones, have jointly revised the paper in a series of proofreads and have addressed the remaining issues. This included in particular improving clarity and accuracy of language and terminology, consistency among sections and and sub-sections (all subsections now follow the order of the taxonomy categories as given in the figure). We have provided

clearer definitions, renamed certain features when needed to remove ambiguity (e.g., “representative samples” is now called “representative elements”). All tools-related paragraphs (Section 4) have been doubled checked to ensure the correctness of their category assignment. Section 3-5 have been fully aligned with the current taxonomy and follow the same structure and organisation. References have been carefully verified and reported in Section 2, which now contains a more detailed list of bibliographical entries categories.

**R2: intro: “the authors are aware that domain-specific approaches to profile and annotate datasets exist. However, to ensure high relevance and applicability, this survey addresses exclusively cross-domain approaches, which are agnostic to the domain of the profiled data.”. I agree with the choice of narrowing the scope of the survey. This answer is generally appropriate to my earlier comment. I am still very surprised that for a survey paper (especially one that has reviewed so many references) no example is given. Readers would surely benefit from a couple of examples, to get the opportunity to realize the difference between what is in focus for this paper and what is not, when the objects seem similar.**

**Response:** Thank you for the comment. To further clarify the scope of the survey, we added examples and modified the corresponding sentence as follows:
“In this survey we address domain-agnostic dataset profiling approaches (e.g., Linked Data Observatory [26]) as described in Section 4 and general vocabularies for representing resource metadata, such as general metadata, quality, provenance, links, licensing, statistics and dynamics, which are applicable to datasets as a particular kind of resource on the Web as described in Section 5. It should be noted that domain-specific vocabularies (e.g., Medical Subject Headings (MESH[1]) or Systematized Nomenclature of Medicine (SNOMED[2])) are out of the scope of this survey even though they can be useful in formalizing domain-specific aspects of a dataset description.”

**R2: the new section on methodology is useful. However it lacks the listing of workshop papers. This is quite surprising! Even more importantly it shows a new problem, especially for a journal like SWJ. The paper includes 86 references: removing [6] (see above) and [66] (a general reference) leads to 84, so the article’s bibliography cannot contain all the references (85) the author claim to have used for the survey. I’m willing to accept that the authors have found references that are not necessarily useful to report in the article, but there should at least be an online annex that gives them all. The list of keywords used to find them could also be good to see, for further assessing the methodology.**

**Response:** Thank you for the comment. As suggested by the reviewer, we have carefully revised the references in the survey. Section 2 now contains a more detailed list of

[1] https://www.nlm.nih.gov/mesh/
[2] http://www.snomed.org/

bibliographical entries and categories, including journal and magazine articles, conference and workshop papers, books, PhD theses and W3C recommendations. The number of papers has been updated accordingly to 85 papers. These papers include 22 journal and 1 magazine articles, 40 conference papers, 19 workshop papers, 1 book, 2 PhD thesis and 7 W3C recommendations, retrieved from the sources listed in Section 2. We also provided examples of keyword queries that have been used to retrieve these references (Section 2.1). Furthermore, a bar chart that depicts the number of referenced papers per year has been added (cf. Figure 2).

**R2: In 4, there is a mismatch between the sections and the main feature categories. In fact this section keeps the structure of the previous version of the paper, without sections corresponding to provenance, licensing and links. And it still used the old 'semantic features' and 'temporal features' terms, which has been replaced by 'general features' and 'dynamics features' in the new version!**

**Response:** Thank you for the observation. In the revised version of the survey we aligned the presentation of the feature categories across the Sections 3-5 to the feature categories in the taxonomy as they appear in Figures 3 and 4. In particular, we added the subsections on provenance, licensing and links to Section 4 and adjusted the terminology in all the sections to match the feature taxonomy. Section 5 has been updated accordingly.

**R2: in the intro of 5, "general-purpose vocabularies such as Dublin Core often provide useful terms also for dataset-specific metadata, but are not discussed in detail here to ensure sufficient focus on vocabularies of more particular relevance for RDF dataset profiling". I am sorry but I can't buy the argument. And in fact the authors don't even buy it, it seems: they end up describing DC in 5.6. And fig 3 mentions DC in the 'General' category. So please refer to DC as early as in 5.1. If only because it DCAT is partly built on it. DC is also worth being mentioned in licensing (dct:License, dc:rights, etc.). And make sure that figure 3 is generally aligned with the content of section 5**

**Response:** Thank you for the comment. We have modified the introduction of Section 5 to better reflect our intentions as follows: "Note that general-purpose vocabularies such as Dublin Core often provide useful terms also for dataset-specific metadata. Even though an exhaustive discussion of such broad vocabularies is not within the scope of this survey, we discuss their use to model specific aspects of datasets, such as provenance or licensing."

In addition, we added a reference to Dublin Core and its influence on DCAT to Section 5.1, discussions of DC/DCTerms to Sections 5.3 (provenance) and 5.5 (licensing).

Finally, we modified Fig.3 to be aligned to the content of Section 5, and vice-versa.

**R2: - EDOAL has been added in 5.2, which is good. The analysis is less good though. In fact this is a wrong sentence: "the typical use case for generating EDOAL statements is the manual formalisation of mapping statements, while less expressive SKOS and VoL**

**statements can be at least partially generated from the output of automated linking and mapping algorithms." EDOAL has been created in the context of the community behind OntologyMatching.org, whose purpose is to evaluate and compare automatic alignment tools. And SKOS happens to be used to represent on the Semantic Web controlled vocabularies that have most often been built manually (even though it can also be used well for representing the result of automatic alignments).**

**Response**: Thank you for the remark. The misleading sentence has been removed in the current revision of the survey.

**R2: daQ has a paper reference, not just a URI. The reference given for DQV made me laugh a bit. I'm really not sure how the authors actually found a working draft from 2015. New versions of DQV have been published until December 2016 (https://www.w3.org/TR/vocab-dqv/). It is no surprise that I have big doubts about the analysis of DQV made by the authors, which is not substantiated anyway ('several concerns about practical issues are raised as part of the DQV working draft documentation.' - which concerns were they?)**

**Response**: Thank you for the comment. We have now strengthened the discussion of DQV by detailing our assessment and supporting it with additional references (Section 5.2). We also replaced the footnote/URL of DQV and added a 2014 reference for daQ [20].

**R2: in general section 5 should really be rationalized wrt. space given to the explanations of the various vocabularies. For example in 5.6 I don't understand why vocabularies coming from other domains like FOAF and SIOC are given as much (or more!) space together than PROV-O. FOAF and SIOC happen to have some elements relevant for provenance, while PROV-O is a quite complex vocabulary which is exclusively devoted to representing provenance facts.**

**- in 5.6 I still don't understand why there is such a long introduction. If the authors want to define what provenance is, this should be done in another, earlier section.**

**Response:** Thank you for the comment. We shortened the introduction of Section 5.3 (previously Section 5.6) and in particular, the descriptions of SIOC and FOAF, in an attempt to balance the content of all subsections. The remaining content reflects exclusively on FOAF/SIOC features for modeling derivations, relations and ownership of resources, i.e. aspects of relevance to provenance.

In addition, we defined provenance earlier in the paper (Section 3.6).

**R2: the authors have added a not on 5.8 trying to motivate that the statistics on general-purpose vocabularies may bring useful insight even if the statistics are on the parts specific to datasets that are used. I agree. However, the problem that I had tried to**

**explain in my earlier review is not "we were unable to filter the instances of dataset profiling-specific terms from our suggested vocabularies while examining their usage statistics in LOD2". My with was rather on filtering on LOD2stats datasets, not vocabularies: I would have liked statistics on datasets that describe datasets (i.e. data catelogues) rather than global statistics for any dataset. In fact I'm worried that LOD2stats has little if no datasets that are about datasets, which would undermine the usefulness of the study.**

**Response:** Thank you for the remark. We agree that there are less datasets about datasets captured in LOD2stats, but some are certainly represented via DCAT (with namespace http://www.w3.org/ns/dcat), VoID, SDMX, SCOVO, and Data Cube, as per the table in the paper.
We do agree that dataset profile descriptions that are available in catalogs may complement the LOD2 statistics. While existing dataset registry platforms such as CKAN are not geared towards natively publishing RDF, the CKAN DCAT extension (https://github.com/ckan/ckanext-dcat) provides an extension for publishing DCAT based profiles serialised into N3, Turtle, JSONLD or XML. However, by default, no additional vocabularies are used.

**R2: In 6.6 I disagree with "Whereas some applications rely on the existing metadata, many applications choose generating dataset profile features as a part of their own processing pipelines. This can be attributed to missing dataset profile features in many cases." Many applications listed in section 6 (especially the data quality assessment tools) are designed to generate dataset profile features. It is their goal. So even if there was profile data pre-existing, they would still compute profile features again!**
**The conclusion has a similar sentence that should be removed.**

**Response:** Thank you for the comment. Indeed, many applications are designed to generate dataset profile features. We removed the last sentence from the text and re-phrased the corresponding part of Section 6.6. to: "Whereas some applications rely on existing metadata, many applications compute dataset profile features as part of their own processing pipelines. These applications can thus directly contribute to the dataset profile generation."

**R2: in 6.6. "we think that availability of dataset profiles including a wide range of features can potentially facilitate a new generation of applications in the distributed LOD settings". This sentence is not really substantiated. Yes, one can say that more data will lead to more applications, but that's not really groundbreaking, when no idea of these new kinds of applications is given. in the conclusion "This leads us to a conclusion that a-priori availability of dataset profiles could facilitate a broader use of profiles and datasets in a variety of application domains." is not very impressive either, especially when "this" (the previous sentence) is very debatable (see previous comment).**

**Response:** Thank you for the comment. We removed the said sentences and adjusted the corresponding statements in the conclusion to the following:
“The availability of dataset profiles has the potential to improve data discovery and reuse on the Web. Remaining challenges and obstacles include the lack of Web-scale adoption of general standards, e.g. for representing profile features, and the lack of automated means for interpreting and using profile information as part of large-scale data reuse scenarios. This survey hence aims at raising awareness and uptake of profiling techniques and vocabularies.”

**R2: Smaller comments:**
**- p3: ‘adopted methodology to’ -> ‘methodology adopted to’**

**Response:** Thank you. We fixed the typo.

**- section 3.3 really needs a reference for the many notions introduced there. There was at least one in the previous paper which seems appropriate ([6]). Or was it not?**

**Response:** Thank you for the observation. We re-added the references to Section 3.3.

**- I don’t understand what makes licensing, provenance and links ‘orthogonal’, even though I’m willing to accept they can be separate (and recommended this, at least for licensing). What does ‘orthogonal in the distribution of profiles’ mean?**

**Response:** Thank you for the comment. In the current version of the survey we removed the term “orthogonal” and the respective description and consider licensing, provenance and links as separate categories.

**- the figures in table 2 should be given a date**

**Response:** Thank you for the comment. We have added a footnote to Table 2 indicating the date of the snapshot.

**- footnote 49 (http://creativecommons.org/licenses/by/3.0/) is not a appropriate reference for the Creative Commons licenses. Please use something else, e.g. http://creativecommons.org/licenses/. And please give a specific reference for CCrel (https://wiki.creativecommons.org/wiki/CC_REL or something like this) as it’s a different vocabulary.**

**Response:** Thank you. We have adjusted the references in the revised version.

**- please make sure that the figures given for 5.8 reflect the latest update (Jan 17). Right now we don’t know.**

**Response:** Thank you for the comment. We have revised the text to reflect the additional data from January 2017.

**- I'm still unsure why one needs so many references in the second paragraph of 6.5.**

**Response:** Thank you for the comment. We reduced the references in 6.5 to the most important ones.

---

**Reviewer R3: The authors have taken considerable efforts to rework this paper. The interpretations and conclusions of the results of the individual sections make the paper much more valuable. There are, however, a few (mainly minor) issues I would like to see addressed.**

**Table 1 lists tools that are either available online or as open source. This indicates that the intersection is empty, i.e., there are no tools that are both available as open source as well as public endpoints. Is that really the case?**

**Response:** Thank you for the observation. We have updated the information in Table 1 to reflect the possibility of intersection.

**Table 3 depicts some interesting trends that require an interpretation. For about half of the vocabularies with non zero values, the trends for triples and datasets are contrary, i.e., there is an increase in triples and a decrease in datasets, or vice versa. The authors should comment on that. PROV-O is a very drastic example, with the number of datasets increasing from 1 to 17, while the number of triples decreases from 4,537 to 577.**

**Response:** Thank you for the comment. We note that for many of the vocabularies that have changing numbers of datasets and triples over time there can be somewhat conflicting numbers (e.g. for SKOS, VoID, etc. where the number of datasets increases but the number of triples decreases). We consider that this can be explained by the removal of a particular dataset/website that has a high number of triples of a particular type, or by the adoption of a new vocabulary/removal of a particular vocabulary for a set of triples on a website. We have added this to the text. With regards to the discrepancies noted in PROV-O, we have corrected one table line as a result (the 2015 data previously in the paper referred to an incorrect namespace also captured in LODstats, namely for http://www.w3.org/ns/prov-o/ instead of

http://www.w3.org/ns/prov). We have also double checked the other vocabularies to ensure that the numbers are correct.

**R3: The analysis of provenance and licensing vocabularies is a bit oversold. In section 5.8, the authors mention this in a few sentences themselves: they do not distinguish between using the vocabulary for provenance/licensing vs. using the vocabulary for something else, hence, the quantitative evaluation depicted in table 3 is a bit shaky. Here, the authors should be more careful in discussing their method of measurement. Furthermore, refinements might be possible here. For example, Hogan et al. [1] discuss an approach for finding license information which goes beyond purely looking at vocabularies (although still a bit hacky). Likewise, in [2], we also applied some further filters that go beyond merely spotting vocabularies.**

**[1] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres and Stefan Decker. "An empirical survey of Linked Data conformance pdf". In the Journal of Web Semantics 14: pp. 14–44, 2012.**
**[2] Schmachtenberg et al.: Adoption of the Linked Data Best Practices in Different Topical Domains, In: ISWC 2014.**

**Response:** Thank you for the comment. We have added the following statement to the discussion of Table 3**: "**While our quantitative assessment in Table 3 indicates mere usage of a particular vocabulary, it does not provide any insights into the way it has been used in particular scenarios. In addition, it is worth noting that particular features, for instance, license information, are often represented through a variety of means, which may not be captured by the vocabularies identified here [Schmachtenberg et al., 2014; Hogan et al., 2012]."