RDF dataset profiling - a survey of features, methods, vocabularies and applications

Mohamed Ben Ellefi, Zohra Bellahsene, John Breslin, Elena Demidova, Stefan Dietze, Julian Szymański, Konstantin Todorov

Undefined 1 (2016) 1–5
IOS Press

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

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] *ENG-3047, Engineering NUI Galway, Galway City, Ireland*
E-mail: breslin@ieee.org
[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, coverage, dynamics and currency, reliable information about such features is essential to enable dataset discovery in tasks such as entity linking, distributed query, search or question answering. Even though there exists a wealth of works contributing to the problem of dataset profiling in general, these works are spread across a wide range of communities. In this survey, we provide a first comprehensive survey of the RDF dataset profile features, methods, tools and vocabularies. We organize these building blocks of dataset profiling in a taxonomy and emphasize the links between the dataset profiling and feature extraction approaches and several application domains. The 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 and locate the relevant features for building a dataset profile for intended applications together with the tools capable of extracting these features from the data.

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

1. Introduction

The Web of Data, and in particular Linked Data [10], has seen tremendous growth over the past 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 [66]. RDF datasets and their inherent subgraphs vary heavily with respect to their size, topic and domain coverage, the resource types and schemas as well as the dynamics and currency.

To this extent, 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* and *query federation*, just to name a few examples. This prevalent problem is underlined by the strong bias towards using established and well-known reference

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

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

knowledge graphs such as DBpedia [4], YAGO [71] or Wikidata[3], although there exists a long tail of potentially suitable domain-specific yet under-recognized datasets.

We begin by providing definitions of several central concepts of our study. In this survey, *an RDF dataset* is defined in accordance with the dataset definition in the VoID Vocabulary[4] stating: "A dataset is a set of RDF triples that are published, maintained or aggregated by a single provider"[5]. According to VoID, this definition reflects the social dimension, such that a dataset represents a meaningful collection of triples as envisioned by its provider, such that this dataset would benefit from descriptive metadata.

A Dataset Profile Feature is a metadata element describing a certain attribute of the dataset. For instance, "dataset dynamicity" is a dataset profile feature providing information on the temporal variation of the dataset. Descriptive metadata consisting of a collection of dataset profile features constitute a dataset profile. A dataset profile is a substantial building block in facilitating effective application-oriented dataset discovery and usage.

An RDF Dataset Profile is a formal representation of a set of dataset profile features.

A dataset profile characterizes the dataset and aids dataset discovery, recommendation and comparison with regard to the represented characteristics. 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 light-weight 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] 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.

[3] https://www.wikidata.org
[4] http://vocab.deri.ie/void
[5] See: 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/

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 this problem is of particular importance in the context of Linked Data, it has been identified and approached already in 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 of datasets has become available throughout the past years, where a comprehensive overview and classification is still missing. A myriad of terms and notions does co-exist, whereas a clear distinction, classification and comparison is still required. Only recently, first efforts [24] have been made to bring together such disparate yet closely related fields.

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 *applications* to encourage experimentation and facilitate 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 profiling 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 structured dataset profiles. While some of the discussed works are dedicated to profiling of graph-based RDF datasets in particular, works of relevance from other related fields are also discussed. It should be noted that 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.

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

- a taxonomy of dataset profile features, including semantic, qualitative, statistical and temporal feature categories;

- a systematic overview of dataset profile feature extraction approaches and tools discussed in the context of our dataset profile feature taxonomy;
- an overview and a classification of available vocabularies for representing dataset features and profiles;
- 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 adopted methodology 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 in the field of dataset profiling and organize these features in a taxonomy. Then, we provide an overview of the existing approaches and tools for the automatic extraction of dataset profile features (Section 4). Following that we provide an overview of the existing RDF vocabularies for the representation of certain dataset profiles and features (Section 5). Where feasible, we also provide suggestions on the vocabulary use and offer vocabulary recommendations suitable for representing particular dataset profile features. Then, we close the circle by exemplifying subsets of features that are considered relevant in selected application scenarios in Section 6. Finally, we provide a conclusion in Section 7.

2. Survey Procedure

In this section, we present the procedure that we adopted to retrieve and filter journal articles and conference papers for this survey. The stages of the survey process are depicted in Fig. 1 and described in the following.

2.1. Terminology and Taxonomy

We began by identifying a basic terminology of dataset profile features from which we extracted potential terms that were most relevant for this systematic review, such as: profiling, dynamicity, quality, index, etc. Terms 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.

2.2. Digital Libraries (/Search Engines) Search

The extracted terms from the taxonomy were used individually and in combination to query different online databases and several search engines (cf. Fig. 1). For example, we used keywords and multiword expressions to build the following combinations: {Semantic Web, Linked Data, Linked Open Data (LOD), etc.} $\times$ {profiling, dynamicity, quality, index, etc.}.

2.3. Literature Review

Each category of the dataset profile features taxonomy covers a large range of works in the Semantic Web field and can be surveyed in a separate paper. In this article, 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 the existent surveys in each category of the dataset profile taxonomy, while providing some examples for: *(i)* the identification of the feature extraction methods (cf. Section 4), *(ii)* the identification of vocabularies for datset profiles representation (cf. Section 5), and *(iii)* the identification of some application-driven profiles (cf. Section 6). Of all the criteria considered, this was the one that produced the sharpest cut down on the number of the articles to be reviewed in detail.

2.4. Paper Selection and Review

By applying a careful review and paper comparison, we obtained a final list of 85 papers to be included in this survey ranging from 1996 to 2016 with about 70% of articles originating from $[2010-2016]$. The selected works are retrieved from different journals, conferences and workshops, mainly in the Semantic Web field as follows:

Journals

- Semantic Web Journal (SWJ)
- Information Processing and Management (IPM)
- ACM Computing Surveys (CSUR)
- Journal of Web Semantics (JWS)
- Australasian medical journal (AMJ)
- FnT Technology, Information and Operations Management (FnT)
- Transactions of the Association for Computational Linguistics (TACL)
- International Journal on Semantic Web and Information

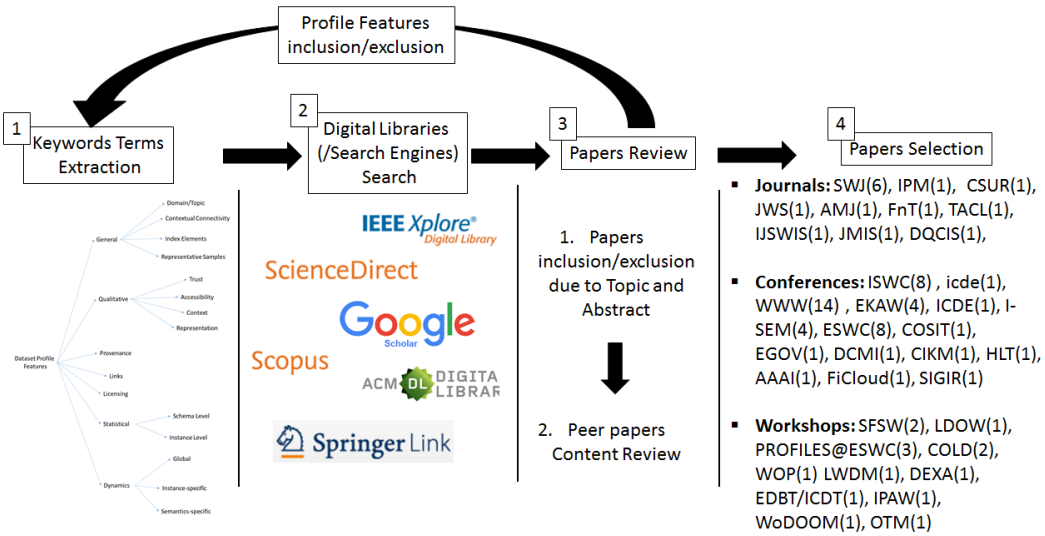

Fig. 1. Survey methodology workflow.

– Systems (IJSWIS)
– Journal of Management Information Systems (JMIS)
– Data Quality in Cooperative Information
– Systems (DQCIS)

Conferences

– International Semantic Web Conference (ISWC)
– International World Wide Web Conference(WWW)
– International Conference on Knowledge Engineering and Knowledge Management (EKAW)
– IEEE International Conference on Data Engineering (ICDE)
– I-Semantics (I-Sem)
– European Semantic Web Conference (ESWC)
– Conference on Spatial Information Theory (COSIT)
– International Conference on eDemocracy and eGovernment (EGOV)
– Dublin Core Metadata Initiative Conference (DCMI)
– Conference on Information and Knowledge Management (CIKM)
– Human Language Technology Conference (HLT)
– Association for the Advancement of Artificial Intelligence (AAAI)
– The IEEE International Conference on Future Internet of Things and Cloud (FiCloud)
– The International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)

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

3. Dataset Profiling Features and Taxonomy

This section provides an inventory of dataset features of relevance to dataset profiling. Features are derived from the literature, in particular, from available dataset profiling methods and vocabularies presented in the following sections. Identified features are clustered and arranged in a feature taxonomy, which provides a categorisation system for the purpose of this survey. We would like to highlight that this taxonomy is extensible and provides one of several feasible ways to categorise profiling features.

In particular, based on an extensive literature overview, we propose to organise features into seven categories: *general*, *quality*, *provenance*, *licensing*, *links*, *statistical* and *dynamics*. This categorisation mirrors the profiling vocabularies distribution, as described in Section 5.

Fig. 2 depicts the resulting taxonomy including references to instances of feature extraction systems. 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 [86]).

3.1. General Features

General features are dataset profile features carrying high-level semantic information that do not fit to any of the more specific categories defined in this survey.

1. **Domain/Topic** A domain refers to the field of life or knowledge that the dataset treats (e.g., music, people). It describes and englobes the topics covered by a dataset (e.g., life sciences or media), understood as more granular, structured metadata descriptions of a dataset, as the one found in [29].
2. **Contextual Connectivity**
   We identify two members of this group:
   (a) ***connectivity properties***, meaning the set of entities shared with other datasets, and
   (b) ***domain/topical overlap with other datasets***. Important information, especially with regard to user queries, can be made available by the overlap of the domains or topics covered by a dataset and other datasets. This overlap can be expressed, for instance, by the presence of shared topics between two datasets [79].
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 lookup and retrieve RDF data items. 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., [49]) or at the instance level (e.g., [39]).
4. **Representative Samples** This group of features is found at the schema and at the instance level. On the one hand, representative schema elements can be understood as: *(i)* the most descriptive set of types (schema concepts) [27], or *(ii)* the set of schema properties that can be used as a keys (almost keys) in instance identification. On the other hand, representative instances are understood as a group of selected data that accurately portrays the whole dataset [26].

3.2. Qualitative Features

The study of data quality has a strong and on-going tradition in the computer science community at large and particularly in the Web Data domain. According to [81], 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.* [9] classified the 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 compare information with 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 information itself, information sources, or information providers. Zaveri *et al.* [86] identified further dimensions and reorganized the quality dimension into four groups: *(i) Accessibility; (ii) Intrinsic; (iii) Contextual; and (iv) Representational*. Yet another approach of metadata quality assessment can be found in [77] that monitors and assesses the quality of 82 active Open Data portals classified in six groups: *retrievability, usage, completeness, accuracy, openness and contactability*.

In this work, we collected commonly used quality features and re-ordered them in a manner that matches the global dataset profile features taxonomy that we introduce giving rise to the following groups of quality features: *(1) Trust; (2) Accessibility; (3) Context; (4) Degree of connectivity; and (5) Representation*.

1. **Trust** Trust is a major concern when dealing with Linked Data. Data trustworthiness can be expressed by the following features.
   (a) ***verifiability:*** the “degree and ease with which the information can be checked for correctness", according to [8].
   (b) ***believability:*** the “degree to which the information is accepted to be correct, true, real and credible" [64]. 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 [86]. 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 the references to it on the Web.

2. **Accessibility** This family of features regards various aspects of the process of accessing the data.
   (a) ***availability:*** an extent to which information is available and easily accessible or retrievable [8].
   (b) ***security:*** refers to the degree to which information is passed securely from users to the information source and back [86].
   (c) ***performance:*** the response time in query execution [86].
   (d) ***versatility of access:*** a measure of the provision of alternative access methods to a dataset [86].
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 [86]. In the Linked Data context, the following sub-features are defined in [8]:
      i. ***schema completeness (/ontology completeness)*** – the degree to which the classes and properties of a schema are represented in the dataset.
      ii. ***property completeness*** – measure of the missing values 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 [64], 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 comprises features that tell something about 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) [64], or the importance of data to the user query [8].
   (b) ***sufficiency:*** the availability of enough data for a particular task ([8] uses the term “amount-of-data").
   (c) ***timeliness:*** the availability of timely information in a dataset with regard to a given application.

3.3. Statistical Features

This group of features comprises a set of statistical features, such as size and coverage or average number of triples, property co-occurrence, etc.

1. **Schema-level** According to the 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, i.e. *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.*

3.4. 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 (take for example data quality). Inversely, the dynamics of a dataset can be seen as a feature of, for example, quality. For that

reason, this family of features is seen as transversal (spanning over the three groups of features described above).

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 to 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 stability of URIs i.e. a URI can be moved, modified or removed.
   (c) ***stability of links:*** the level of broken links between resources, i.e. a link is considered as broken if the a target URI changes [65]. Whereas 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 other linked datasets.
3. **Semantics-specific** [36] [25]
   (a) ***structural changes:*** evaluation of the degree of change in the structure (internal or external) of a dataset.
   (b) ***domain-dependent changes:*** this feature reflects the dynamics across different domains that impacts the data.
   (c) ***vocabulary-dependent changes:*** a measure of the dynamics of vocabulary usage.
   (d) ***vocabulary changes:*** a measure of the impact of a change in a vocabulary to the dataset that uses it.
   (e) ***stability of index models:*** the level of change in the original data after the data has been indexed.

3.5. Orthogonal Features

Here, we draw the reader's attention to the fact that some quality features may be orthogonal in the distribution of profiles features, notably to general categories. As orthogonal profile features we consider licensing, provenance and links, described as follows:

1. **Licensing** Here, we adopt the recommendation of Heath *et al.* [45]; "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, redistribution are permitted. This can have a direct impact on data quality, both in terms of trust and accessibility. Hence, the importance of the existence of license in both human-readable and machine-readable profiles (i.e, including the description in a license vocabulary cf. Section 5.7).
2. **Provenance** the contextual metadata that provides indicators about timelines, currency and update cycles of datasets, which are necessary to know the origin of data, trace errors and notably establish trust. Hence, the provenance is a profile feature used to determine the believability of a dataset. An example use case scenario is to determine some trust score for SPARQL query results in a data sharing triple-store with different provenance.
3. **Links** Links here is understood as the number of datasets, with which a dataset is interlinked, or as the number of triples in which either the subject or the object come from another dataset. Two datasets can be linked through: (i) explicit links when they have linked instances, for example using *owl:sameAs*[10] when sharing identical instances, and (ii) implicit links when sharing topic profiles or context profiles, where explicit links

[10] `http://www.w3.org/2002/07/owl#sameAs`

like *rdfs:seeAlso*[11] can be also used. Dataset links feature covers both schema-level and instance-level representation of links in a dataset profile.

4. Dataset Profiling and Feature Extraction Methods

The field of dataset profiling and features extraction comprises a broad range of tools that is too large to cover here. In this section, we provide examples of relevant dataset profiling approaches for each category of features, as introduced in the previous section. An overview of the dataset profile features categories and the corresponding extraction approaches is given in Fig. 2 and described below in detail.

4.1. Semantic Features Extraction

Semantic features presented in Section 3.1 include domain/topic, context, index elements and representative schema/instances. In the following we present a selection of tools that support feature extraction in this category.

FluidOps Data Portal [79] is a framework for source contextualization. It allows the users to explore the space of a given source, i.e. search and discover data sources of by topics in `http://data.fluidops.net/resource/Topics`. Here, the contextualization engine favors the discovery of relevant sources during exploration. For this, entities are extracted/clustered to give 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 tool enables the retrieval of the "Context" features.

Linked Data Observatory [29] 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 allows to find datasets providing data for a given set of topics or to discover datasets covering similar fields. This Structured Dataset Topic Profiles are represented in RDF using the VoID vocabulary in tandem with the Vocabulary of Links (VoL) (the vocabularies will be reviewed in Section 5 in more detail). The Linked Data Observatory allows the extraction of the "Domain/Topic" dataset profile features.

voiDge is a tool that automatically generates VoID descriptions for large datasets. This tool allows users to compute various VoID information and statistics on dumps of LOD as illustrated in [14]. Additionally, the tool identifies (sub)datasets and annotates the derived subsets according to the VoID specification. The voiDge describes the "Schema/Instances" dataset profile features.

The keys discovery approaches aim at selecting the smallest set of relevant predicates representing the 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 keys 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. ***(ii)*** *ROCKER* – [70] key discovery approach that uses a refinement operator. This operator is able to detect sets of properties that describe any instance of a given class in a unique manner. Reportedly, *ROCKER* is more suited to large scale data than *SAKey*. Keys can be seen as a "Representative Schema/Instances" dataset profile feature.

RDF QTree structure [39] is an approximate multidimensional indexing structure to store 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 [49] 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 is each schema concept that maps to the data sources containing instances with the corresponding properties. While SchemEX provides different index structure than the QTree index, both indexing tools involve the "Index Elements" dataset profile feature in the category Semantic Features.

[11] `https://www.w3.org/TR/rdf-schema/#ch_seealso`

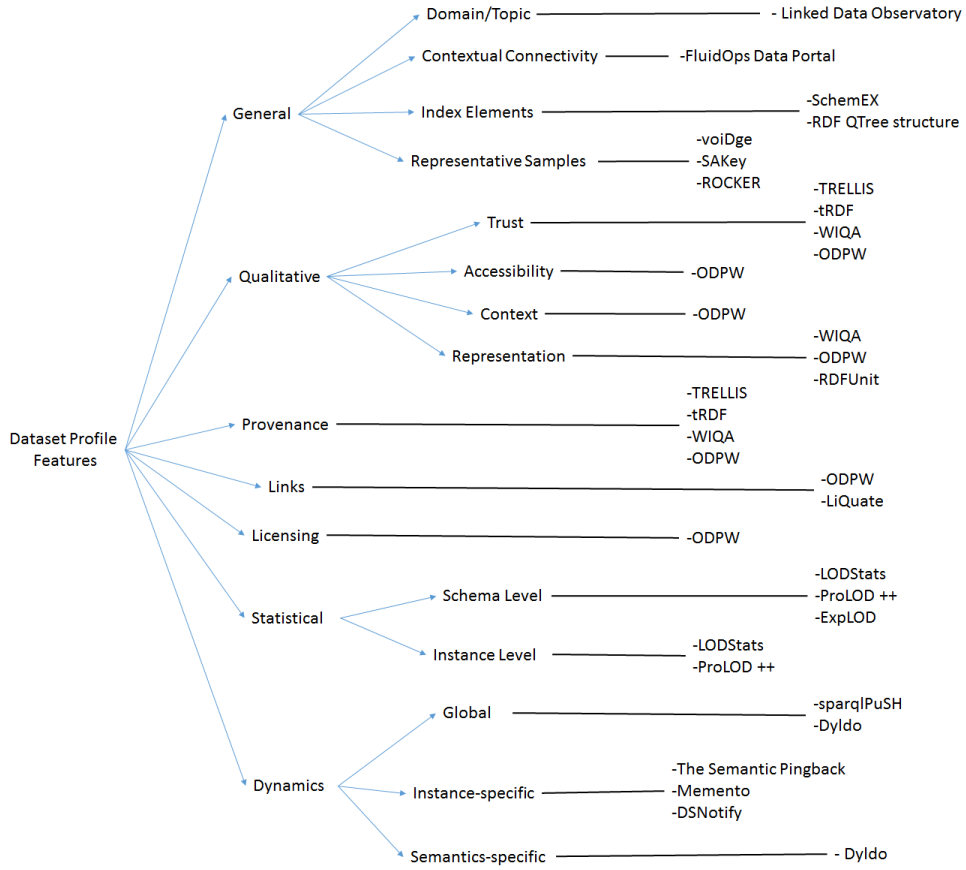

Fig. 2. A taxonomy including dataset profile features organized in general, qualitative, statistical and dynamics categories as well as links to the corresponding feature extraction systems.

4.2. *Quality Features Extraction*

As discussed in Section 3, in this survey we focus on selected groups of quality features such as trust, accessibility, context and representation, 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 assessement approaches in the context of Linked Data in general is provided by Zaveri *et al.* [86], who conducted an extensive survey of 21 works.

TRELLIS [34] is an interactive environment that examines the degree of trust of datasets based on user annotations. The user can provide Trellis with semantic markup of annotations through the interaction with the ACE tool[12] [12]. 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 description for the "Trust" feature in a dataset profile.

tRDF [40] 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* is an extension of the RDF query language SPARQL in two clauses: TRUST AS clause and the ENSURE TRUST clause. The trust values are based on subjective perceptions about the query object. While TRELLIS is based on users annotation, tRDF extracts the "Trust" feature by allowing users to query the dataset and access the trust values associated to the query solutions in a declarative manner.

WIQA [9] is a set of components to evaluate the trust of a dataset using a wide range of different filtering policies based on quality indicators like prove-

[12] Annotation Canonicalization through Expression synthesis.

nance 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 is a generic qualitative tool which can provide description about the "Trust", "Provenance" and the "Representations" dataset profile features.

LiQuate [68] is a tool to assess the quality related to both incompleteness of links, and ambiguities among labels and links. This quality evaluation is based on queries to a Bayesian Network that models RDF data and dependencies among properties. LiQuate enables the retrieval of the "Links" dataset profile features.

RDFUnit [50] is a framework for the data quality that tests RDF knowledge based on Data Quality Test Pattern, DQTP. A pattern can be: ***(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 the adequate test pattern for a given dataset is not available, the user has to write his own DQTPs, which can then become part of a central library to facilitate later re-use. RDFUnit provides "Representations" dataset profile features in form of DQTPs.

Open Data Portal Watch (ODPW) [77] is a publicly available dashboard component that displays quality metrics for different data portals using various views and charts. These quality metrics are grouped in six dimensions which are retrievability, usage, completeness, accuracy, openness and contactability. The openness indicator provide information to which licenses and file formats conform to the open definition. Furthermore, the watch provides a search service that retrieve the licenses for a given resource URI. ODPW involves all the quality dataset profile features besides the orthogonal features "Links", "Licensing" and the "Provenance".

4.3. Statistical Features Extraction

Statistical features discussed in Section 3.3 comprise schema-level and instance-level statistics.

LODStats [5] is a statement-stream-based tool and framework for gathering comprehensive statistics about datasets adhering RDF. The tool calculates 32 different statistical criteria on LOD such as those covered by the VoID Vocabulary. 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 CKAN[13] metadata repository, either as a patch or as an external web application using CKAN's API. LODStats provides descriptions for "schema-level" and "instance-level" statistical dataset profile features.

ExpLOD [48] 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 structure information is aggregated with statistics like the number of instances per class or the number of property used. ExpLOD provides description about the "schema-level" statistical features for a given dataset.

ProLOD++ [2] is an interactive user interface, which is divided into a cluster tree view and a detailed view. The cluster view enables users to explore the cluster tree and to select a cluster for further investigation for statistics. ProLOD ++ is an extension of *ProLOD* [15], which generates basic statistics. In addition to the mining and the cleansing tasks, the tool generates dataset profiling features related to key analysis, predicate and value distribution, string pattern analysis, link analysis and data type analysis. Hence, ProLOD ++ is a web-based tool, which allows to profile arbitrary LOD datasets in terms of "schema-level" and "instance-level" dataset profile features.

4.4. Temporal Features Extraction

sparqlPuSH [61] is an interface that can be plugged in any SPARQL endpoint and that broadcasts notifications to clients interested in what is happening in the store using the PubSubHubbub[14] protocol [30] 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 broadcasted to interested parties. sparqlPuSH consists in two steps:

[13] http://ckan.org/

[14] 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're interested in is updated.

(i) register the SPARQL queries related to the updates that must be monitored in an RDF store, *(ii)* broadcast changes when data mapped to these queries are updated in the store. sparqlPuSH extracts "global" dataset profile features in the temporal dataset profile category.

The Semantic Pingback [76] is a mechanism that allows users and publishers of RDF content, of weblog entries or of a scientific article to obtain immediate feedback when other people establish a reference to them or their work, thus facilitating social interactions. It also allows to publish backlinks 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 facilitating 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 "instance-specific" features of a dataset profile.

Memento [23] is a protocol-based time travel that can be used to access archived representations of a resources 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*. These relationships are particulary useful for the extraction of the "instance-specific" features and in particular of the "growth rate" feature. Mementos are available both in HTML and RDF/XML.

DSNotify [65] is a link monitoring and maintenance framework, which attenuates the problem of broken links due to the URI instability. When remote resources are created, removed, changed, updated or moved, the 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 dataset temporal profiling category.

The Dynamic Linked Data Observatory (Dyldo) [47], is a framework to achieve a comprehensive overview of how LOD changes and evolves on the Web. It is an observatory of the dynamicity on the Web of Data over time. The observatory provides weekly crawls of LOD data sources starting from $02/11/2008$ and contains 550K RDF/XML documents with a total of 3.3M unique subjects with 2.8M locally defined entities. The system examines, firstly, the usage of Etag and Last-Modified HTTP header fields, followed by an analysis of the various dynamic aspects of a dataset (change frequency, change volume, etc). Dyldo provides temporal dataset profile features in terms of both "global" and "semantics-specific" features.

4.5. A Note on Dataset Profiling Methods

Here, we discuss several issues regarding dataset profile extraction methods that we observed in the survey process. We begin by the most sensitive profile representations, the semantic features, which typically require domain knowledge with respect to the content of the dataset. As best practice, we recommend that the semantic category should be provided by the data domain experts (e.g. data providers or maintainers) to ensure high quality of the semantic profile. On the other hand, we consider that qualitative, statistical and temporal 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 semantic profile extraction tools, notably for the "domain/topic" and "context" features, where only few approaches allow automatic extraction of such profiles features.

Further on the dynamic aspect, in order to facilitate up-to-date dataset profiles, these profiles need to be regenerated periodically, based on the dataset dynamicity. The dataset versioning/archiving also requires versioning/archiving of the corresponding dataset profiles in order to ensure coherence between the dataset snapshots and their profile versions.

Finally, we stress the fact that RDF dataset profiles need to provide representations for both human and machine reading. Hence, in Table 1, we provide an overview of the dataset profiling methods including representation formats. In other words, we check for each method if the extracted profile features are designed for human reading or machine reading. In addition the table provides links to the homepages of each extraction method.

5. Vocabularies for Representation of Dataset Profiles and Features

This section introduces vocabularies for representation of dataset profiles, ranging from general dataset metadata to vocabularies dedicated to one or more of the features introduced in Section 3. Note that general-

Method Name	H/M	Accessibility	Home Page
FluidOps Data Porta	H	O.S.	http://data.fluidops.net
Linked Data Observatory	H/M	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.	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.	http://trdf.sourceforge.net/tsparql
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.	http://rdfunit.aksw.org
ODPW	H	Online	http://data.wu.ac.at/portalwatch
LODStats	H/M	Online	http://stats.lod2.eu/
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 features extraction methods: Homepages (⋆ means that the homepage was not available at the time of access); Accessibility that can be Open Source (O.S.) or Online (via SPARQL ENDOINT or via HTTP API, etc.); and Human readability (H) *vs.* machine readability (M).

purpose vocabularies such as Dublin Core[15] 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.

5.1. General Dataset Metadata Vocabularies

A range of vocabularies exist which can be used to provide more general metadata of datasets or ontologies. While the Ontology Metadata Vocabulary (OMV) [43] 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.

The Vocabulary of Interlinked Datasets (VoID) [3] provides a core vocabulary for describing datasets and their links. The schema[16] includes the classes *Dataset*, *DatasetDescription*, *LinkSet*, *TechnicalFeature*. The authors distinct *dataset* from *RDF graph*, where *dataset* refers to "meaningful collection of triples, that deal with a certain topic, 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 subject and object are in different datasets/namespaces. The VoID guidelines recommend additional vocabularies (DC-

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

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

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 (Source: http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/).

Terms, FOAF for general metadata and SCOVO - the Statistical Core Vocabulary[17] for statistical information. VoID is already widely used in the Web of Data, as documented by Table 2, depicting the use of VoID descriptions among the 1014 datasets and per category in the current inventory of the Web of Data[18].

The Data Catalog Vocabulary (DCAT)[19] follows a similar rationale and has been created based on a survey of government data catalogues [53]. 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 particular dataset in a specific access format (e.g., an RDF dump or a SPARQL endpoint). For categorisation of datasets, the *dcterms:subject* predicate and controlled SKOS vocabularies are recommended.

5.2. 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 dataset's links. *SKOS*[20], 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)[21], which provides a general vocabulary to describe metadata about links or linksets, within or accross 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)[22] 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 or restrictions to constrain classes/entities. EDOAL in that sense provides the means to on-the-fly interpretation of mapping statements as part of data integration scenarios. On the other hand, in contrast to VoL, there are no means for representation of provenance of mapping statements. Next to being more comprehensive and expressive than SKOS or VoL, another major difference seems to be that 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.

5.3. 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

[17] http://purl.org/NET/scovo
[18] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/
[19] http://www.w3.org/TR/vocab-dcat/
[20] https://www.w3.org/TR/2009/REC-skos-reference-20090818/
[21] http://data.linkededucation.org/vol/index.htm
[22] http://alignapi.gforge.inria.fr/edoal.html

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.* [33] describe the DQM Ontology[23], 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 Assessement 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.

In addition, the authors introduce a preliminary classification for data quality problems.

In addition, the WIQA - Web Information Quality Assessment Framework[24] describe some early work to filter content according to quality features, also introduce WIQA-PL, a vocabulary for modeling content access policies. However, the work appears to be deprecated and not maintained.

Also worth to mention is the work in [32], where authors use the SPARQL Inferencing Notation (SPIN) - a vocabulary that allows the representation of SPARQL queries - to represent data quality rules.

In addition, the Dataset Quality Vocabulary (daQ)[25] and the Data Quality Vocabulary (DQV)[26] 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, several concerns about practical issues are raised as part of the DQV working draft documentation.

Finally, while provenance information often provides indicators about timelines, currency and update cycles of datasets, Section 5.6 introduces additional vocabularies of relevance.

5.4. Dataset Dynamics & Evolution

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

The Dataset Dynamics group[27] for instance lists a number of vocabularies for representing dataset changeset and updates. The *Talis Changeset vocabulary*[28] provides some early, yet discontinued work on representing changeset and specific characteristics, and has a similar approach as the Delta vocabulary[29]. The *Triplify Update vocabulary*[30] 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.* [38] 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 (cf. 5.6). While RMO implements baseline revision management notions for data graphs, it is of lesser relevance for the purpose of this section.

A more abstract approach is offered by the *Dataset Dynamics (DaDy) Vocabulary*[31], 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

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

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

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

[26] https://www.w3.org/TR/2015/WD-vocab-dqv-20150625/

[27] http://www.w3.org/wiki/DatasetDynamics

[28] http://vocab.org/changeset/schema.html

[29] http://www.w3.org/2004/delta

[30] http://triplify.org/vocabulary/update

[31] http://vocab.deri.ie/dady

dady:UpdateDynamics. The latter captures information about the update regularity and frequency.

For capturing specific features and observation related to dynamics and evolution, beyond the ones covered by the vocabulary above, in particular the vocabularies mentioned in the following section, aimed at representing statistical dataset features, which may or may not be related to dynamics.

5.5. 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[32], SDMX[33] or SCOVO[34].

The VoID guidelines, for instance, recommend the use of SCOVO to share statistical dataset features [3]. Authors foresee, on the one hand, statistics concerning the whole dataset or linkset, such as triple count, and attributing statistics to a source, to capture where a statistical datum stems from. Inline with some of the authors' concerns about the adequacy of SCOVO, it has been superseded by the Data Cube Vocabulary in the more recent past.

The RDF Data Cube vocabulary[35], currently a W3C Editors Draft developed by the Government Linked Data Working Group[36] 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)[37] 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[38] 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.

SCOVO[39], also described by Hausenblas *et al.* [44], 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 [21], both approaches are superseded by the Data Cube vocabulary, which represents the state of the art in representing statistical data on the Web.

Auer *et al.* present LODStats [6], a framework for dataset analytics, which introduces a set of 32 statistical features and uses the most recommended combination of VoID and the DataCube 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 the two state of the art vocabularies for general dataset metadata (VoID), respectively statistical data (Data Cube), it turns out to be the most future-proof approach to capture statistical dataset metadata.

5.6. Data and Dataset Provenance

A variety of definitions have been given for provenance over the past number of years. One very pragmatic definition comes from the Provenance Working Group[40] of the W3C, 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."* On the Web, provenance can pertain to any resource found on the Web - documents, data, or datasets - but it can also be found in a resource that is used to describe the provenance of an object in the real world.

The main aim of the Provenance Working Group was to create standards that could be used to define and work with provenance data. A document from its previous incarnation as an Incubator Group states

[32] http://www.w3.org/TR/vocab-data-cube/
[33] http://sdmx.org
[34] http://vocab.deri.ie/scovo
[35] https://dvcs.w3.org/hg/gld/raw-file/default/data-cube/index.html
[36] http://www.w3.org/2011/gld/
[37] http://sdmx.org/
[38] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html
[39] http://vocab.deri.ie/scovo
[40] http://www.w3.org/TR/2013/REC-prov-dm-20130430/

the difficulties involved in such standardisation efforts: *"provenance is too broad a term for it to be possible to have one, universal definition - like other related terms such as "process", "accountability", "causality" or "identity", we can argue about their meanings forever (and philosophers have indeed debated concepts such as identity or causality for thousands of years without converging)"*.[41]

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 [57]. 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). A provenance process is defined by Moreau [56] as "the provenance of a piece of data is the process that led to that piece of data".

We will now describe some of the main provenance models used on the Web, some of which have specific applicability in terms of whole datasets.

1. **voidp** 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 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)**[42] 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**[43] 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 [19] 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**[44] 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)[45] as a lightweight vocabulary, and the

[41] http://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

[42] http://bit.ly/premisOntology

[43] http://purl.org/vocab/changeset

[44] http://trdf.sourceforge.net/provenance/ns.html

[45] http://purl.org/net/opmv/ns#

OPM Ontology (OPMO)[46] using more advanced OWL constructs.

8. The **PROV Ontology (PROV-O)**[47] 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 [55] created to support "the widespread publication and use of provenance information of Web documents, data, and resources" – including a Data Model (PROV-DM) [57] and an Ontology (PROV-O) [52] – 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.[48]

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 [41], 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.

– **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 [41] 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 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 who has control over a resource or dataset. *sioc:ip_address* can be used to link the created data and creator if specified to an Internet address. Also, *sioc:last_activity_date* can be used to reference the last activity associated with a resource, although this may still be interpreted in different ways (modified, read, etc.). As with *dc:source*, 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 some SIOC modules which can be used in provenance descriptions for datasets. The most relevant is probably the **SIOC Actions** [20] 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:byproduct*, *sioca:creates*, *sioca:deletes*, *sioca:modifies*, *sioca:object*, *sioca:product*, *sioca:source* and *sioca:uses*. These have been aligned to OPM and PROV-O in recent work by [60].

[46] http://openprovenance.org/model/opmo
[47] http://www.w3.org/TR/prov-o/
[48] http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

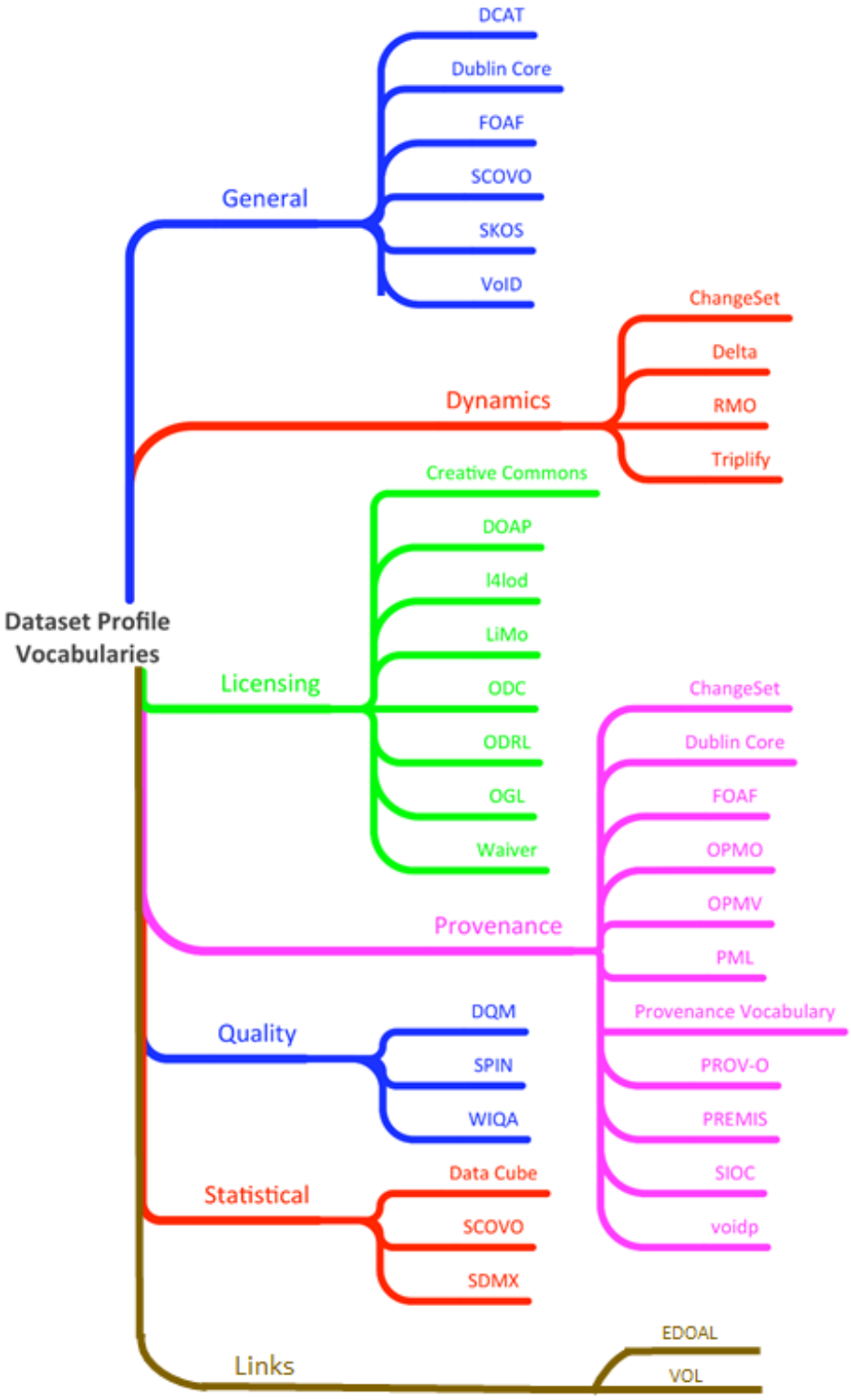

Fig. 3. Overview of relevant vocabularies as classified by type of dataset profile features.

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
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
PROV-O	Provenance	4,537	1	577	**17**
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.**

5.7. Dataset Licensing

We will now examine what vocabularies are available to assist with licensing of data and datasets. These include RDF versions of common licensing frameworks and alignments of multiple licensing frameworks into a combined vocabulary.

– **Creative Commons (CC)**[49] 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

[49] http://creativecommons.org/licenses/by/3.0/

to be expressed in RDF using the ccREL (REL, or rights expression language) vocabulary. Many datasets in the LOD cloud are already licensed under Creative Commons, as we will see later.

- The **Open Data Commons (ODC)** license[50] 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 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[51] 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[52] uses RDF or JSON, evolving from an earlier XML-based REL version[53].
- **Open Government License (OGL)**[54] 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 ODC. One of the dataset projects using OGL is the data.gov.uk service.
- The **License Model (LiMo)**[55] 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)**[56] 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 re-applied to datasets since they share many of the same properties.
- **Licenses for Linked Open Data (l4lod)**[57] was introduced in [37] 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).

5.8. Observations

We use the LOD2 Stats service[58] 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.[59] 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 2016, 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.

251 datasets use RDF syntax, 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 somewhere around 21% (52) and 10% (25) of datasets respectively.

Another observation is that none of the quality or dynamics and evolution vocabularies appear in LOD2 Stats. That points to a significant underutilization

[50] http://opendatacommons.org/licenses/
[51] http://www.w3.org/community/odrl/two/model/
[52] http://w3.org/ns/odrl/2/
[53] http://www.w3.org/TR/odrl/
[54] http://www.nationalarchives.gov.uk/doc/open-government-licence/
[55] http://purl.org/LiMo/0.1
[56] http://usefulinc.com/ns/doap
[57] http://ns.inria.fr/l4lod/
[58] http://stats.lod2.eu/ as accessed on 2nd February 2015 and re-checked again on 19 January 2017
[59] 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).

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) containing Creative Commons metadata. However, according to work by [37], 95% of the datasets in the LOD cloud[60] 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 applications in many domains, including, 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.

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 [22], Illinois Wikifier [67] as well as Babelfy [58].

Example features for data linking applications: Data linking applications typically use the semantic features discussed in Section 3.1 such as topics, domains, languages (versatility) and location coverage, as well as representative parts of schema/instances, and specifically the key candidates extracted with the key discovery approaches described in Section 4.1.

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 [85].

A number of recent works focus on statistical methods for: (1) outlier detection to detect errors in numerical values [31], [63], [82]; (2) automatic prediction of missing types of instances [63]; and (3) the identification of incorrect links between datasets [62]. 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., [16], [84]). Thereby some works go beyond error detection and attempt to automatically determine correct data values in case of inconsistencies [16]. 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 [31] 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 3.3) 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 [16] 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.,

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

answering queries over distributed data [11]) rely on the schema information. Recently, approaches aimed at the automatic inference of missing schema information have been developed (e.g., [63], [49]).

Example features for type inference: Statistical characteristics of datasets (see Section 3.3) play an important role in type inference applications. For example, in [63] 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 [39] that can include full-text search extensions (see e.g., [1]). Also combinations of both query paradigms are possible [42]. 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 [83]; 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 [39]. The majority of existing federated query approaches for LOD (e.g., [42], [39], [80], [35]) 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 semantic and statistical characteristics (see Sections 3.1 and 3.3) 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 [39]. 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 [59], as well as property-specific [69] and type-specific statistics [83].

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 [28]. 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 [51] 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., UMLS [13]), 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., [67], [54], [78]. They also employ statistical features extracted from large-scale text corpora [17] and allow one to expand the user queries to increase recall [7]. 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 [46] and temporal [18] 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 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 semantic, qualitative, statistical and temporal 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 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. On the one

hand, these applications can thus directly contribute to the dataset profile generation. On the other hand, the burden to generate dataset profile features for each single application hinders usability of the datasets. Thus 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 and enlarge the number of datasets used in real-world applications.

7. Conclusions

RDF dataset profiling is perceived as a central challenge in enabling and facilitating dataset discovery in application scenarios such as data linking, data curation, distributed query and search, just to name a few. In this survey, we provide a comprehensive overview for 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 propose that certain profiles features, notably in the semantic category, should be provided by the data domain experts to ensure high quality profiles. Another observation concerns the vocabulary usage where some features, such as the quality or the dynamicity of vocabularies do not appear in the evaluated statistics. That leads 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 or how they have changed over time.

We observe that although existing applications make use of the whole spectrum of the discussed feature categories, including semantic, qualitative, statistical and temporal features, the concrete set of features is application-dependent and the whole set is rarely used within any single application. Furthemore, many applications generate dataset profile features as a part of their own processing pipelines, which can be attributed to missing dataset profiles or features in many cases. 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.

Finally, we strongly recommend that dataset profiles 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 assume that the problem of dataset profiling will become an even more prominent one, and corresponding methods will form a crucial building block for enabling reuse and take-up of datasets beyond established and well-understood knowledge bases and reference graphs.

References

[1] Fedsearch: Efficiently combining structured queries and full-text search in a sparql federation. volume 8218 of *Lecture Notes in Computer Science*, pages 427–443. Springer Berlin Heidelberg, 2013.

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

[3] 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.

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pages 722–735, 2007.

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

[6] SÃűren Auer, Jan Demter, Michael Martin, and Jens Lehmann. Lodstats - an extensible framework for high-performance dataset analytics. In Annette ten Teije, Johanna VÃűlker, Siegfried Handschuh, Heiner Stuckenschmidt, Mathieu d'Aquin, Andriy Nikolov, Nathalie Aussenac-Gilles, and Nathalie Hernandez, editors, *EKAW*, volume 7603 of *Lecture Notes in Computer Science*, pages 353–362. Springer, 2012.

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

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

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

[10] 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.

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

[12] 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.

[13] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

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

[15] 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.

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

[17] Paul Buitelaar, Philipp Cimiano, and Bernardo Magnini. *Ontology learning from text: An overview*, volume 123. 2005.

[18] 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.

[19] Jeremy J Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. Named graphs, provenance and trust. In *Proceedings of the 14th international conference on World Wide Web*, pages 613–622. ACM, 2005.

[20] Pierre-Antoine Champin and Alexandre Passant. SIOC in Action - Representing the Dynamics of Online Communities. In *Proceedings of the 6th International Conference on Semantic Systems (I-SEMANTICS 2010)*. ACM, 2010.

[21] Richard Cyganiak, Simon Field, Arofan Gregory, Wolfgang Halb, and Jeni Tennison. Semantic statistics: Bringing together sdmx and scovo. In Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas, editors, *LDOW*, volume 628 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2010.

[22] Joachim Daiber, Max Jakob, Chris Hokamp, and Pablo N. Mendes. Improving efficiency and accuracy in multilingual entity extraction. In *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124, 2013.

[23] 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 WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010*, 2010.

[24] E. Demidova, S. Dietze, J. Szymanski, and J. Breslin, editors. *Proceedings of the 1st International Workshop on Dataset PROFIling & fEderated Search for Linked Data (PROFILES 2014), co-located with the 11th Extended Semantic Web Conference (ESWC 2014), Anissaras, Crete, Greece, 26 May 2014.*, volume 1151. CEUR Workshop Proceedings, 2014.

[25] 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.

[26] 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.

[27] Mohamed Ben Ellefi, Zohra Bellahsene, Konstantin Todorov, and Stefan Dietze. Dataset recommendation for data linking: An intensional approach. In *ESWC: European Semantic Web Conference (ESWC 2016), Crete, GrÃĺce*, number 9678, pages 36–51, 2016.

[28] 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.

[29] 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 - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 519–534, 2014.

[30] Brad Fitzpatrick, Brett Slatkin, and Martin Atkins. Pubsubhubbub core 0.3–working draft. *Project Hosting on Google Code, available at http://pubsubhubbub.googlecode.com/svn/trunk/pubsubhubbub-core-0.3.html*, 2010.

[31] Daniel Fleischhacker, Heiko Paulheim, Volha Bryl, Johanna Völker, and Christian Bizer. Detecting errors in numerical linked data using cross-checked outlier detection. In *Semantic Web Conference (1)*, pages 357–372, 2014.

[32] Christian Fürber and Martin Hepp. Using semantic web resources for data quality management. *Management*, 6317:1–15, 1998.

[33] 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.

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

[35] 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.

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

[37] 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 *Proceedings of the International Semantic Web Conference (ISWC 2013)*, 2013.

[38] Markus Graube, Stephan Hensel, and Leon Urbas. R43ples: Revisions for triples - an approach for version control in the semantic web. In *LDQ@ SEMANTICS*, 2014.

[39] 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.

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

[41] 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.

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

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

[44] Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayers. Scovo: Using statistics on the web of data. In Lora Aroyo, Paolo Traverso, Fabio Ciravegna, Philipp Cimiano, Tom Heath, Eero HyvÃűnen, Riichiro Mizoguchi, Eyal Oren, Marta Sabou, and Elena Paslaru Bontas Simperl, editors, *ESWC*, volume 5554 of *Lecture Notes in Computer Science*, pages 708–722. Springer, 2009.

[45] 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.

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

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

[48] Shahan Khatchadourian and MarianoP. Consens. Explod: Summary-based exploration of interlinking and rdf usage in the linked open data cloud. In *The Semantic Web: Research and Applications*, volume 6089 of *Lecture Notes in Computer Science*, pages 272–287. Springer Berlin Heidelberg, 2010.

[49] 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.

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

[51] 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.

[52] Timothy Lebo, Satya Sahoo, and D McGuinness. PROV-O: The PROV Ontology, 2013.

[53] Fadi Maali, Richard Cyganiak, and Vassilios Peristeras. Enabling interoperability of government data catalogues. In Maria Wimmer, Jean-Loup Chappelet, Marijn Janssen, and Hans Jochen Scholl, editors, *EGOV*, volume 6228 of *Lecture Notes in Computer Science*, pages 339–350. Springer, 2010.

[54] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[55] Paolo Missier, Khalid Belhajjame, and James Cheney. The W3C PROV family of specifications for modelling provenance metadata. In *EDBT/ICDT '13*, pages 773–776, 2013.

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

[57] Luc Moreau and Paolo Missier. PROV-DM: The PROV Data Model, 2013.

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

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

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

[61] 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.

[62] 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.

[63] 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.

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

[65] 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.

[66] Jeffrey Pound, Peter Mika, and Hugo Zaragoza. Ad-hoc object retrieval 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 771–780, 2010.

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

[68] Edna Ruckhaus, Maria-Esther Vidal, Simón Castillo, Oscar Burguillos, and Oriana Baldizan. Analyzing linked data quality with liquate. In *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.

[69] 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.

[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 AAAI Florida AI Research Symposium (FLAIRS-2004), May 1719, 2004*, 2004.

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

[74] Danai Symeonidou, Nathalie Pernelle, and Fatiha Saïs. KD2R: A key discovery method for semantic reference reconciliation. In *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. Proceedings*, 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 *3rd International Conference on Future Internet of Things and Cloud, FiCloud 2015, Rome, Italy, August 24-26, 2015*, pages 404–411, 2015.

[78] Ellen M Voorhees. Using wordnet for text retrieval. *Fellbaum (Fellbaum, 1998)*, pages 285–303, 1998.

[79] 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, May 26, 2014.*, 2014.

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

[81] 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.

[82] Dominik Wienand and Heiko Paulheim. Detecting incorrect numerical data in dbpedia. In *The Semantic Web: Trends and Challenges - 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, pages 504–518, 2014.

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

[84] 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.

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

[86] 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.

RESUBMIT of 1296-2508

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

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]{breslin@ieee.org}
[c]{demidova, dietze}@L3S.de
[d]{julian.szymanski@eti.pg.gda.pl}

Submitted to "Semantic Web Journal"

We would like to thank the reviewers and the editor for their effort and time and their detailed, insightful and constructive comments. We have completely revised the survey according to these comments. Please find below our replies containing precise details of the revisions made according to the specific comments of the reviewers.

Note that the following structural changes have taken place, leading to changes in sections, tables and figures numbering as follows:

- We added a new Section 2 "Survey Procedure" that describes the procedure adopted to retrieve and filter the papers for this survey.
- We added new Table 1 summarizing profile features extraction methods in Section 4.
- We added a new Section 4.5 "A Note on Profiling Methods" that summarises our observations with respect to the dataset profiling methods.
- We added Figure 3 (Section 4) containing an overview of relevant vocabularies.
- Section 2 "Dataset Features" corresponds to Section 3 in the revised survey version.
- Section 3 "Dataset Profiling and Feature Extraction Methods" corresponds to Section 4 in the revised survey version.
- Section 4 discussing application-driven profiles corresponds to Section 6 in the revised survey version.
- Table 1 on adoption of VoID across linked datasets corresponds to Table 2 in the revised survey version.
- Table 2 on overall vocabulary usage corresponds to Table 3 in the revised survey version.

Reviewer R1:

R1: The paper gives a comprehensive overview on works for profiling linked data sets. While the sheer amount of works in this area collected and organized in the survey is impressive, the paper would benefit from showing the findings and conclusions in a more concise manner.

Response: Thank you for the comment. In the revised version of the survey, we added more discussion and conclusions, in particular in Section 4.5 that summarises our observations with respect to the dataset profiling methods, in Section 5.8 that discusses our observations on vocabulary usage in RDF datasets and in Section 6.6 that discusses dataset profiles in the context of applications. We have also expanded the overall conclusions in Section 7.

R1: First, the title is misleading, as it talks about "Dataset Profiling" in general, but is limited to RDF/Linked Data. I suggest making the title more concise here (instead of increasing the scope of the survey to match the title).

Response: Thank you for the suggestion. According to the suggestion of the reviewer we adjusted the title of the survey to better reflect the focus on the RDF datasets to “RDF Dataset Profiling - a Guide to Features, Methods, Applications and Vocabularies” and also pointed out the focus on the RDF datasets in the abstract and the introduction.

R1: The collection of features and profiling characteristics in section 2 is really comprehensive and interesting. It would be, however, even more concise if measurements for the different characteristics were introduced, where applicable. I would also expect that for some of the characteristics, different measurements are proposed in different works, so that they could be contrasted and discussed.

Response: Thank you for the suggestion. Although we do not discuss the possible measurements for the different characteristics in detail, they often follow from the definition of the feature (e.g. in case of statistical features) or have been extensively discussed in the literature already (e.g. qualitative features in [94]). To address the comment of the reviewer and to further clarify the scope of the survey, we added a remark to the beginning of the corresponding section (Section 3 in the current version of the survey).

R1: The section about tools is also quite impressive, and I was surprised that so many practical profiling tools exist. The authors might consider adding a summary table with the basic characteristics of the tools (open source/commercial, Web/standalone app, etc.).

Response: Thank you for the suggestion. With respect to the reviewer proposition and for more clarity, we added Table 1, which contains a summary of the profile feature extraction methods pointing to the homepages of the tools and containing information with regard to their accessibility (open source, online, via endpoint / API etc.) and human vs. machine readability of the extracted information.

R1: While I appreciate the section about applications, it feels like this selection is a bit arbitrary. Applications that could be listed as well include, e.g., (natural text) question answering, information visualization, or semantic annotation, so it is not clear why the five examples shown in the paper have been chosen. Furthermore, it is not clear why the list of features presented for each application should be exhaustive.

Response: Thank you for the comment. The application examples and the corresponding dataset profile features presented in the survey (Section 6 of the current survey version) are not exhaustive and serve as examples to illustrate how dataset profiles can support certain types of applications and which features do they typically use based on the current literature (whereas the methodology of the literature review is included in the new Section 2). To address the comment, we adjusted the description at the beginning of Section 6 to clarify the scope of the section.

R1: As far as the listing of vocabularies is concerned, I would also like to see a summary table (similar to the tools section), showing the coverage of the vocabularies w.r.t. different groups of features. This would enable the readers to pick and/or combine vocabularies that fit their needs.

Response: Thank you for the suggestion. In order to provide a clear connection between the groups of features and the vocabularies, we replaced the table with a leaf tree (cf. Figure 3) where the leaves represent the individual vocabularies, clustered into the profiling categories identified through our taxonomy earlier.

R1: Finally, I miss some more conclusions. For a survey, it is always interesting to close with a judgement of the findings, and some outlook on a future research agenda. Relevant questions could be: what aspects are currently underrepresented in Linked Data profiling? Which combinations of features would be desirable, but are not supported by any tool? Is there a match or a mismatch between theory, tools, and the requirements of applications that exploit data profiles? In summary, I appreciate the very comprehensive nature of the survey, but I would like to see the findings being presented more concisely in a revised version.

Response: Thank you for the suggestion. To address this comment and provide more discussion and conclusions to the survey, we added Section 4.5 that summarises our observations with respect to the dataset profiling methods, updated Section 5.8 that discusses our observations on vocabulary usage in RDF datasets and added Section 6.6 that discusses dataset profiles in the context of the applications, which we made while reviewing respectively, dataset profiles extraction methods, vocabularies for representation of dataset profiles and features as well as applications. We have also expanded overall conclusions in Section 7.

R1: Minor:
*** in the beginning of section 2, it is mentioned that the notion of an "atomic feature" is introduced, but it only described as a leaf in the hierarchy. It is unclear what those atomic features are in the end, why it is clear that they cannot be further subdivided, and why it is important.**

Response: Thank you for the insightful comment. In the previous version of the survey, "Atomic feature" was used to describe a feature that has no descendants in the hierarchy. In the current version of the survey we revised the description of the features, such that the notion of atomic feature is not used anymore.

*** Please use a blank before a reference (i.e., "reference [1]", not "reference[1]".)**

Response: Thank you, we adjusted the format.

*** p.3: "set of RDF triples shared" - I would rather suspect that two datasets share entities, but it is unlikely that they share exactly the same triples about those entities.**

Response: Thank you, that should indeed be the set of entities shared with other datasets, we adjusted the description.

Reviewer R2:

R2: First, there is a lack of precise framework defining the notions used. The paper actually does not define what is a dataset. This is not very strict, while catalogues and main used vocabularies like DCAT try to make crucial distinctions like dataset vs distribution. "Descriptive metadata, i.e. profiles" on p1 is also disturbing. Profiles are not limited to what is called descriptive metadata for many (e.g., access metadata is not descriptive metadata). In fact for several communities working with descriptive metadata, the notion of "application profile" could conflict with the one of "dataset profile" that is quite different.

Response: Thank you for the comment. As suggested by reviewer, to clarify the notions used in the survey, we added definitions of RDF dataset (in accordance to the VoID definition), RDF dataset profile and RDF dataset profile feature to the introduction of the survey.

R2: The state-of-the-art is very extensive, and constitutes a very useful resource for would-be reader. This is probably indeed the first time this is attempted at such a scale! But there are two problems with it: Even though it is extensive, it is incomplete with respect to dataset profiling. Authors focus on gathering references for dataset (quality) analysis. This is good, but there are important efforts that are not mentioned, about creating frameworks for expressing profiles, especially providing (or re-using) vocabularies. These could have been compared with what the authors proposed. DCAT and VOID are refered but the analysis is very cursory. More domain-specific efforts have been ignored: for example, the Health Care and Life Science community has researched a dataset profile (http://www.w3.org/2001/sw/hcls/notes/hcls-dataset/?). In the geographic domain, the EU initiative GeoDCAT-AP should also be studied. The state-of-the-art on profiling and quality also misses reference to relevant ISO standards: domain-specific as for the geo ones that influence GeoDCAT-AP
or more general like the ISO25000 family (esp 25012).

Response: Thank you for the comment. The authors are aware that domain-specific approaches to profile and annotate datasets exist. However, to ensure high relevance and applicability, this survey focuses exclusively on cross-domain approaches, which are agnostic to the domain of the profiled data. We added this statement to the introduction of the survey to clarify the scope of our work.

R2: There are inconsistencies in the way the references are used. First, in section 2 many references are given for different profiles features. Most of them are data analysis papers. For a gathering of features, simple references to a vocabulary or inventory (e.g. a dataset catalogue) that exhibit the features would be enough to give a requirement for the topic. Actually this would be more convincing than a piece of academic work, possibly very technical, which may fall short giving a practical motivation for what it does (as the focus would be on an algorithm or an experiment). For example DCAT has properties for representing the domain/topic of a dataset. There is no need to refer to [59] that extracts such topics. On the other hand, section 3 that gathers methods to extract profile features doesn't refer to these papers that have been cited in section 2, that presents such methods. This is quite a missed opportunity. Furthermore, some of the references in section 2 seem

superfluous for explaining what specific features are: I am not sure one would need both [6] and [35] ([6] looks more relevant) or both [52] and [53] or both [36] and [37].

Response: Thank you for the comment. As suggested by the reviewers, we revised the references in Section 3. In particular, we attempted to be more explicit by citing technical papers in the feature definitions (note that some technical papers are cited explicitly as examples and not as references for the features). We also limited the citations to only the most relevant ones. With respect to the relevant vocabularies mentioned by the reviewer, in the revised version of the survey we provide an overview of vocabularies for representation of dataset profiles and features in Fig. 3 (Section 5) and link them to the taxonomy of dataset profile features introduced in Section 3. (Section numbers refer to the revised version of the survey).

R2: In 2.2 there is a problem with the reference for the classification of quality characteristics, which is presented as coming partly from [97]. This is only an 'under review' paper, without URL nor publication context, so it's unclear what the source is. As a matter of fact the authors mention in [97] have published a very recent paper in this very journal, "Quality Assessment for Linked Data: A Survey: A Systematic Literature Review and Conceptual Framework", so I've used this one. When I've done the comparison, I found many differences, even some features that are classified in different dimensions, which I guess would be found in any recent work of the authors of [97]. For example licensing is in "accessibility" in [97], it is in Trust in the paper. It is possible that [97] (and other references like [92]) has shortcomings. But the choices made here are debatable (I would argue that licensing is orthogonal to data quality). And in any case such deviations compared to the state of the art should be explained. They are currently not even flagged as such, it's very difficult for the reader to guess what is happening in this section. Section 2 actually reads as if the authors propose a new quality framework, which could be interesting but is arguably not what the section had embarked on.

Response: Thank you for the comment. The citation [97] ([86] in the new version of the survey) was updated, as suggested by the reviewer. Regarding the choice of data quality dimensions, we note that there are different patterns for data quality dimensions in the Linked Data field, from which we cited [9], [77] and [85]. In this work, we collected commonly used quality features and we reordered them in a manner that matches the global dataset features profiles taxonomy that we introduce giving rise to the following groups of quality features: (1) Trust, (2) Accessibility, (3) Representativity, and (4) Context (cf. Section 3.2).

R2: In section 3.2 the authors should make more explicit whether they re-use the matter of [97] for a subset of the systems there, if they extend that matter. If that section contains original material, it should be more explicit. If not, then it can be considerably shortened.

Response: Thank you for the comment. The citation [97] ([85] in the new version of the survey), provides a detailed review of 21 general data quality assessment tools. In our survey (Section 4.2 of the revised survey version), we summarized selected methods and techniques for dataset profile extraction with respect to the quality features according to the proposed dataset feature taxonomy. We revised the description in Section 4.2 to make it explicit.

R2: In section 3.3 it's unclear whether all systems mention really "extract" temporal characteristics (as written in the title of the section), or if they just manage them: - Semantic pingback as it is described in the text mostly focuses on cases where a dataset is being re-used in other sources. In principle the pinging doesn't change the content of the original dataset, and thus doesn't facilitate

consistency and timeliness per se. It's possible that the publishers would integrate changes based on the pings, but it's not essential to the general pinging approach.

Response: Thank you for the comment. In the context of the survey, we are interested in the service for dynamic features extraction in Semantic Pingback even if it is considered optional in the general pinging approach. This system is particularly useful for detecting the stability of links/backlinks. We updated the description in the current Section 4.4, to clarify this point.

R2: - Memento is a mechanism to serve different time versions of data. It represents data that can be used to compute temporal characteristics (for example number of versions) but it doesn't extract them by itself.

Response: Thank you for the comment. Here, we are mainly interested in the temporal relationships that the system can provide, i.e., the first-memento, last-memento, next-memento and prev-memento. This system relates to the growth rate dataset profile feature. We updated the description in Section 4.4 to clarify this point.

R2: In section 4.2 and 4.3 tools are presented that compute statistics or make assessments based on them. But these tools don't really motivate the need for statistics to be already expressed in a profile. On the contrary, they compute these statistics or extract features themselves, by querying the data and/or running inferences. It's not very difficult to extract say, the number of properties used in a dataset. And it's more reliable than using already published profile data, which could be outdated - for data assessment tools this is crucial.

Response: Thank you for the comment. The applications for data curation, cleansing, maintenance (described in the section 4.2 of the initial version of the survey) and schema inference (described in the section 4.3 of the initial version of the survey) rely on the statistical characteristics of the datasets. 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. We added a clarification of this point to the beginning of the application section (Section 6 in the revised version of the survey).

R2: Then, the stats on vocabulary usage analysis in section 5 is very promising, but it doesn't look reliable. The data is from early 2015, probably it has changed one year later. There are some finding that are very surprising, such as Creative Commons being used only for 12 datasets. As the authors write it themselves later in the paper, there must be more data out there with CC Licenses. More importantly, it's uncertain whether table 2 really gives the info the authors claim it gives in the paragraph p17-18. The text says that LOD2 is used to find info on how many times a vocabulary is used in datasets. But this doesn't mean that the vocabularies are used in these datasets for dataset profiling (i.e. to describe datasets, e.g. instance of dcat:Dataset). For example Dublin Core and FOAF can be used to described many types of resources that are not datasets. If the authors have indeed filtered in the LOD2 data the statements that are about datasets, this should be explained in more details. Without these details, one will infer that the data is not about datasets, and thus that it's not very informative for section 5 in general.

Response: We thank the reviewer for this observation. The reviewer rightly noted that just because a vocabulary is widely used, it does not mean that it is the best one for dataset profiling. Based on popular terms, however, we can suggest that they are more likely to be used if they are already known, and if they

are also fit for purpose. We have updated the data as per the latest LOD2 statistics (re-checked 19 January 2017). The numbers show consistency in the usage of the main vocabularies and their reliability in terms of being stable/growing number of instances. The updated the corresponding table in the paper (Table 3). We have also added this sentence to the paper to reflect the above: "*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. 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.*"

R2: Finally, the paper completely falls short on presenting the "RDF vocabulary for unambiguously identifying dataset features" that was promised in the abstract. A link is given to http://data.data-observatory.org/vocabs/profiles/ns, but the elements of this vocabulary are not listed and documented. And there's no instruction/example of how to use it. Shall it be combined with existing vocabularies? Is it an alternative to all of them, combined?

Response: Thank you for the observation. In the revised version of the survey we focus on the systematic review of the existing vocabularies, rather than proposing a new vocabulary. A new vocabulary will become a part of our future research.

R2: Minor comments:

R2: - p1: "As the Web of Data is constantly evolving, manual assessment of dataset features is neither feasible nor sustainable." This statements is debatable. Sure, it won't be possible to profile all datasets manually, but one could argue that it would be a feature of good providers that they provide at least some profile metadata.

Response: Thank you for the comment. Indeed, good providers would aim to provide profile metadata. Nevertheless, doing so manually on a regular basis, weil covering a broad range of features would not scale. We adjusted the wording in the paper to clarify this point.

R2: - p5: the difference between stability of URIs and stability of links is quite unclear, as the only definition given to characterize the stability of links refers to stability of URIs.

Response: Thank you for the comment. Whereas 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 other linked datasets. We adjusted the description in Section 3.4 to clarify this point.

R2: - p5: what does "explore the space of a given source, i.e., search and discover data sources of interest." mean?

Response: Thank you for the comment. To clarify this point we re-phrased the corresponding expression in Section 3.4.

R2: - p5: the relation between the references in section 3 and the criteria at 2 is unclear at times. For "selecting the smallest set of relevant predicates representing the dataset in the instance comparison task", do the predicates correspond to a specific criterion in section 2? (are they RDF

predicates? And why the paragraph say "we review" while it does just drop the various aspects of the keys discovery approach?)

Response: Thank you for the comment. To address the comment we provided explicit links between the descriptions of the extraction tools described in Section 4 and the features presented in Section 3.

R2: - p5: footnote 9 is not finished.
Response: Thank you for the comment. We fixed the typo.

R2: - p12: [83] reads more like state-of-the-art for section 2.2 than a vocabulary for 5.2. Same comment applies to the bullet list in the second column of this page.

Response: Thank you for the comment. We agree that "features" directly correlate with vocabulary "terms" and hence, several if not most related works which propose vocabularies also describe features. This also applies to [83] which introduces both a set of features and a corresponding ontology, as stated in the first sentence in Section 5.2. The mentioned bullet list describes the concepts (that is types) introduced by [33] for modeling dataset quality. These concepts directly relate to features, yet in the context of Section 5.2 we are specifically interested in the vocabulary terms rather than the features. (Section 5.2 corresponds to Section 5.3 in the current version of the survey).

R2: - p12: why not mention EDOAL as a reference for alignment vocabulary, next to (or instead of) VoL? Why not mention that VoID also has a part for linksets? Why not mention daQ for data quality? SPIN on the other hand is not for expressing data quality features. Representing rules is quite different from representing the results of applying rules.

Response: Thank you for the comment. We have added EDOAL and included a dedicated subsection (Section 5.2) on vocabularies for representing links and linksets. We have added a discussion of daQ and DQV to Section 5.3. SPIN indeed represents data quality rules, which is explicitly stated in the survey.

R2:- p15: Dublin Core is also used by DCAT and many others for licensing, so it should be in 5.6 (and maybe also other sub-sections as it's a very general vocabulary)

Response: Thank you for the comment. We have added DCAT to the general dataset vocabulary section (i.e. Section 5.1).

Reviewer R3:

R3: The main comment on the current version of this paper is that it does not make a good and well-motivated review: it does list many items that are relevant but the way in which this is all put into a bigger picture is not well-motivated. That limits the value of the review. Authors should in a new version pay much more attention to the justification of passing on all these links.
Moreover, if the aim is to make this paper to be the goto-paper for this subject, kicking off a new piece of research, then the paper should contain more of a definition of the subject and a tangible contribution for other papers to cite and build upon. Currently, besides the paper's subject there is not a concrete thing that other authors would start massively linking to and citing. If such a problem definition would be added the paper has much more chance of being cited and this much more value.

All in all, the paper carries lots of interesting links and items but lacks in a clear justification and aggregation, to be a pivotal review paper.
(1) Suitability as introductory text, targeted at researchers, PhD students, or practitioners, to get started on the covered topic. 50 (but see comments for how to make this a better review paper).
(2) How comprehensive and how balanced is the presentation and coverage. 30 (but see the comments for how to add justification and objectives etc.)
(3) Readability and clarity of the presentation. 50
(4) Importance of the covered material to the broader Semantic Web community. 40

Response: Thank you for the comments and suggestions. In line with your comments, in the revised version of the survey, we added a systematic overview of the survey methodology, provided definitions of the concepts being discussed to the introduction of the survey and organized the overall discussion by defining a taxonomy of dataset profile features, as well as by linking features, extraction methods, vocabularies and applications discussed in the survey to this taxonomy. Finally, we added a discussion of the observations we made in these categories.

Detailed comments: The abstract mention the paper consider 'works' but does not address that in the context of this subject it would be interesting to learn from both scientific works as well as industrial works, or framed differently people working towards the profiling as researchers as well as people working where uptake takes place or could take place.

Response: Thank you for your comment. In this survey, we focus on the features, extraction methods, vocabularies and applications that have been discussed in relevant scientific publications. To address this comment and to clarify the scope of the survey, in the revised version of the survey, we added a systematic overview of the survey methodology to Section 2.

In the introduction it is suggested that the solution is coming from the side of researchers that unite and join a common path. That is a fair ambition, but requires two things: 1) that this choice is explicitly made and thus it is clear what the target audience is of the paper, 2) that this does not lead to yet another researcher proposal that does not get uptake. Notwithstanding this ambition, it is nice to see that for outsiders and beginners an overview is created with lots of useful pointers. One thing that could be stressed more in the opening parts of the paper is why this is special and specific for Linked Data (as opposed to any dataset type).

Response: Thank you for the suggestions. In this survey, we specifically focus on RDF datasets, and discuss features, methods and applications specific for this type of datasets, while also including relevant information on dataset profiling from other domains. We see the target audience of the survey as researchers, dataset providers and application developers who work with RDF datasets. We aim to provide this audience with a comprehensive overview on RDF dataset profiling to encourage experimentation and facilitate broader use of RDF datasets. To address the comment, we adjusted the title, abstract and introduction of the survey and provided an RDF dataset definition in the introduction to make the focus on the RDF datasets more clear; furthermore we added the information on the target audience to the introduction to further clarify the scope of the survey.

In section 2, it would be nicer if the opening would explain both the objective of identifying the characteristics (including the separation into semantic etc.) and the justification of the way that this was done. After all, for a review like this it is important to justify the review approach, to indicate completeness etc. The other main comment regarding the presentation of these features is whether

this is supposed to tell what the literature says (knowing that not all scientific papers have large impact) or whether this is aiming to make a 'chosen' summary of that (the authors making a weighted account of what the literature has reported).

Response: Thank you for the comment. As suggested by the reviewers, Section 2, which is newly added, explains the adopted reviewing process and describes the nature of this survey explicitly by giving the reader a bird's-eye view on the RDF datasets profiling problem while providing some examples of worm's-eye view especially in terms of feature extraction methods, application-driven profiles and vocabularies for profiles representation. We also stress on the fact that this survey is dedicated to Linked Data but more specifically for RDF datasets profiling of which we have added concrete definitions to the introduction.

R3: When representing features, it would be good to perhaps include examples and concrete values to illustrate what they really are. Otherwise the true meaning of mentioning a feature is left too vague. The current text of the section looks like it could be equally well have been given as a simple table or list, but of course the section should do more as juts present a list; it should also make the reader understand how the list was composed. So, assuming that the content of the list is fair, it is recommended that the justification gets more attention.

Response: Thank you for the suggestion. We revised the corresponding section (Section 3) and added an explanation on how the taxonomy of dataset profiling features in the survey was constructed along with more comprehensive description of the features. The concrete value of the features can be extracted using the methods described in Section 4 with respect to a given scenario.

R3: In section 3 the authors use the word 'review'. In addition to mentioning all the tools and approaches it would be interesting to see what the review entails. After all the review more or less by definition implies that the tools and approaches are considered from a chosen perspective and it would be good to clarify the perspective and its motivation.

Response: Thank you for the comment. In addition to summarization of different features, methods, tools, vocabularies and applications presented in the initial version of the survey, in the revised version we have provided a systematic classification of these items with respect to a taxonomy containing semantic, qualitative, statistical and temporal categories of dataset profile features linking the items discussed in the survey to these categories, such that the reader can determine relevant features of a dataset profile as well as identify existing extraction tools and vocabularies for generating and representing relevant profile features.

R3: Similarly, at the end of the section, or at the end of the subsections it would be nice to see some concluding and summarizing remarks. After having learned from the different items discussed what they do individually, it would be good to see what the authors see as bigger picture for that aspect.

Response: Thank you for the useful suggestion. To address this comment and provide more discussion and conclusions to the survey, we added Section 4.5 discussing profiling methods, Section 5.8 containing our observations with respect to the vocabularies, and Section 6.6 discussing application-driven profiles. We also expanded the overall conclusion of the survey in Section 7 to summarize our overall observations.

R3: For the next sections I could repeat the same line of comments. A number of aspects are given, leading to subsections, but it is unclear whether these aspects pop up because this is what the

literature offers or whether these were the aspects the authors carefully chose to use for their analysis of the state of the art. The value of the lists of items discussed depends strongly on the way the aspects are chosen and applied.

Response: Thank you for the comment. In order to address this comment and to clarify the goals and the methodology of the survey we adjusted the introduction and added Section 2 describing survey methodology. Furthermore, we defined a taxonomy of dataset profile features and organized the overall discussion in all sections by systematically linking presented aspects to this taxonomy.

R3: Section 5 appears to follow a different presentation style: unclear why. It is also not easy to understand how the different elements in this section go together. In line with the unclear ambition of this section, it is odd to see that the authors start making recommendations in this reviewing section. I would strongly recommend to separate the observations from the aggregate conclusions and from any subsequent recommendations.

Response: Thank you for the useful suggestion. To address this comment, we aligned the overall presentation in the survey to the dataset profiling taxonomy presented in Section 3, separated our observations and discussion from the literature review, and summarised our observations regarding different aspects in dedicated subsections (Section 4.5, 5.8, and 6.6).