Undefined 1 (2016) 1–5
IOS Press

Dataset Profiling - a Guide to Features, Methods, Applications and Vocabularies

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] *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] *Faculty of Electronics, Telecommunications and Informatics, 80-233 Gdańsk-Wrzeszcz, Poland*
E-mail: julian.szymanski@eti.pg.gda.pl

Abstract. The Web of data, in particular Linked Data, has seen tremendous growth over the past years. However, reuse and take-up is limited and focused on a few well-known and established knowledge bases. This can be attributed in parts to the lack of reliable and up-to-date information about the characteristics of available datasets. While datasets vary heavily with respect to features related to quality, coverage, dynamics and currency, reliable information about such features is essential for enabling data and dataset discovery in tasks such as entity retrieval or distributed search. Even though there exists a wealth of works contributing to this central problem of dataset profiling, these are spread across a range of communities and disciplines. Here, we provide a first comprehensive survey of dataset profiling features, methods, tools and vocabularies and also provide an RDF vocabulary for unambiguously identifying dataset features.

Keywords: Linked data, Dataset profiling, Dataset vocabularies

1. Introduction

The Web of Data, and in particular Linked Data [12], has seen tremendous growth over the past years, leading up to the availability of a large amount of structured datasets on the Web, where a recent crawl[1] of Linked Data sets retrieved over 1000 datasets alone, including over 8 million explicit resources and an estimated 100 billion triples[78]. Datasets and their inherent subgraphs vary heavily with respect to their size, topic and domain coverage, the resource types and schemas or the dynamics and currency.

To this extent, the discovery of suitable datasets which satisfy specific criteria has become a challenging problem for tasks such as *entity and dataset linking*, *entity retrieval*, *distributed search* or *query federation*. This prevalent problem is underlined by the strong bias towards using established and well-known reference graphs such as DBpedia [5], YAGO [82] or Freebase[2], for such tasks, while there exists a long tail of potentially suitable yet under-recognized datasets.

Descriptive metadata, i.e. *profiles*, about available datasets are seen as a substantial building block for facilitating dataset discovery, and hence, help tackling the aforementioned tasks. By a *dataset profile* we understand the formal representation of a set of features that describe a dataset and allow the comparison of different datasets with regard to their represented charac-

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

[2] `http://freebase.com`

teristics. Usually, the relevant feature set is dependent on a given application scenario and task.

A number of popular dataset registries have emerged, which tackle this problem through the curation of light-weight dataset descriptions, often also exposing structured metadata according to state-of-the-art vocabularies such as DCAT[3] or VoID[4]. Popular examples include The DataHub[5] or DataCite[6], while the LinkedUp Catalog[7] 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. As the Web of Data is constantly evolving, manual assessment 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 disciplines 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, where a clear distinction, classification and comparison is still required. Only recently, first efforts [28] have been made to bring such disparate yet closely related fields together.

This work aims at providing a comprehensive overview on *dataset profiling* and closely related approaches, respectively, the involved *features*, *methods*, *applications* and *vocabularies*. Being the first comprehensive study in this area, we provide a thorough analysis and definition of related terms and typical dataset profiling features, leading up to a structured RDF vocabulary of profiling features, the *Vocabulary of Dataset Profiles*, which helps in unambiguously identifying features in a dataset profile description. Furthermore, we provide a systematic study of 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 our discussed works are particularly dedicated to profiling of graph-based datasets, such as RDF-based Linked Data, works of relevance from other related fields are also discussed.

We outline the main contributions of this work, following the structure of this paper. We start by providing a comprehensive set of commonly investigated dataset features (Section 2), based on the existing literature in the field of dataset profiling (whether or not referred to explicitly by using this term). These atomic features are organized into a taxonomy, formally represented as an RDF vocabulary of dataset profiling features and made available to the public. Further, we provide an overview of the existing approaches and tools for the automatic extraction of such features (Section 3). We identify explicitly subsets of atomic features that are considered relevant in particular prominent application scenarios, discussed and analysed in detail (Section 4). Finally, we close the circle by providing information about the already existing RDF vocabularies for representation of certain dataset profiles and features (Section 5). Where feasible, we also provide suggestions on vocabulary use and offer vocabulary recommendations suitable for representing particular features.

2. Dataset Characteristics

The present section makes an inventory of dataset features that can be considered in the dataset profiling process. We take into account features that have been studied in the literature, although often referred to by using different names or defined differently. We propose a common terminology and definitions of (and distinction between) the key terms used in the field. We organize the features in a (non-strict) hierarchy and we introduce the notion of an *atomic feature*, understood as one that has no descendants in this hierarchy. Based on the proposed hierarchy, we introduce an RDF vocabulary of dataset profiles (VoDP)[8].

[3] http://www.w3.org/TR/vocab-dcat/
[4] http://vocab.deri.ie/void
[5] http://www.datahub.io
[6] https://www.datacite.org/
[7] http://data.linkededucation.org/linkedup/catalog/

[8] VoDP (Vocabulary of Dataset Profiles) : http://data.data-observatory.org/vocabs/profiles/ns

2.1. Semantic Characteristics

We present a set of features that carry semantic information about a dataset.

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 [59] (e.g., life sciences or media), understood as more granular, structured metadata descriptions of a dataset, as the one found in [32]. The cross-domain or multi-topical nature of a dataset is separate feature that indicates of its potential connectivity properties and its specificity.
2. **Context –** We identify two members of this group:
   (a) ***connectivity properties***, meaning concretely the set of RDF triples 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 [90], [89].
3. **Index elements –** Index models have been introduced in order to retrieve information from the LOD. An index is defined as a set of key elements (e.g., types), which are used to lookup and retrieve data items. These elements can be defined on schema level or on entity level. 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 [55]. In that sense, a set of index elements is viewed as a descriptive semantic dataset characteristic.
4. **Representative Schema/Instances –** This group of features is found on schema and on instance level and is understood as a set of types (schema concepts) or a set of key properties/values, or a representative sample of instances [30], [4], [71].

2.2. Qualitative Characteristics

The study of data quality has a strong and on-going tradition. According to [92], data quality is generally conceived as *fitness for use*, i.e., the capability of data to respond to the demands of specific user given a specific use case. Data quality has multiple dimensions, many of which cannot be evaluated in a task-independent manner. Here, we provide a list of features related to several of these dimensions. Many of these features apply to data quality in general and are directly issued from [92]. However, some of them have been defined particularly in the context of linked data [97].

1. **Trust –** Trust is a major concern when dealing with LOD 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 [10].
   (b) ***believability:*** the "degree to which the information is accepted to be correct, true, real and credible" [76]. This can be verified by the presence of the provider/contributor in a list of trusted providers [97].
   (c) ***reputation:*** a judgement made by a user to determine the integrity of a source [97]. Two aspects are to take into consideration:
      i. ***reputation of the data publisher:*** an indice coming from a survey in a community that determines the reputation of a source,
      ii. ***reputation of the dataset:*** scoring the dataset on the basis of the references to it on the web.
   (d) ***licensing policy:*** 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 (see below).
   (e) ***provenance:*** "the contextual metadata that focuses on how to represent, manage and use information about the origin of the source" [97].
2. **Accessibility –** This family of characteristics regards various aspects of the process of accessing data.
   (a) ***availability:*** the extent to which information is available and easily accessible or retrievable [10].
   (b) ***security:*** refers to the degree to which information is passed securely from users to the information source and back [97].

(c) ***performance:*** the response time in query execution [97].
(d) ***versatility of access:*** a measure of the provision of alternative access methods to a dataset [97].

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 [97]. In the Linked Data context, [10] defines the following sub-features:

      i. ***schema completeness (ontology completeness)*** – the degree to which the classes and properties of an ontology are represented,
      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 datasets, and
      iv. ***interlinking completeness*** – refers to the degree to which links are not missing in a dataset.

   (b) ***understandability:*** refers to expression, or, as defined by [76], the extent to which data is easily comprehended.
   (c) ***accuracy / correctness:*** the equivalence between a value in a dataset instances and the actual real world value of it.
   (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) [76], or the importance of data to the user query [10].
   (b) ***sufficiency:*** the availability of enough data for a particular task. [10] uses the term "amount-of-data".
   (c) ***timeliness:*** the availability of timely information in a dataset with regard to a given application.

5. **Degree of connectivity –** Connectivity here is understood as simply 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 (note the difference with contextual connectivity in the class of semantic features and interlinking completeness in the representation class of features).

2.3. Statistical Characteristics

This group of characteristics comprises a set of statistical features, such as size and coverage or average number of triples, property co-occurrence, etc. [6], [35].

1. **Schema-level –** According to schema, we can compute statistical features such as *class / properties usage count*, *class / properties usage per subject and per object* or *class / properties hierarchy depth*.
2. **Instance-level –** Features on this level are computed according to the data 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*

2.4. Temporal Characteristics

This class of features concerns the dynamicity of a dataset (as identified in a catalogue like Datahub) [53], [52]. Every dataset feature is 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 characteristics.
(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 origine 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 entities (instances).
(b) ***stability of URIs:*** the level of stability of URIs i.e., an URI can be moved, modified or a removed.
(c) ***stability of links:*** the level of broken links between resources, i.e., a links is considered as broken link if the a target URIs changes [77]

3. **Semantics-specific** [41] [29] –

(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 having been indexed.

3. Dataset Profiling and Feature Extraction Methods

We review the approaches for dataset profiling, as well as the systems and tools for dataset features extraction, following the categorization introduced in the previous section. An overview of the dataset features and the corresponding extraction systems is shown in Fig.1 and described in detail below.

3.1. Semantic Characteristics Extraction

FluidOps Data Portal[9] [90] 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 interest. 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.

Linked Data Observatory[10] [32] provides an explorative way to browse and search through existing datasets in the LOD Cloud according to the topics which are covered. 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), i.e., those vocabularies will be reviewed in section 5.

voiDge[11] is a tool that automatically generates VoID descriptions for large datasets. This tool allows users to compute the various VoID informations and statistics on dumps of LOD as illustrated in [16]. Additionally, the tool identifies (sub)datasets and annotates the derived subsets according to the voiD specification.

The keys discovery approaches aim at selecting the smallest set of relevant predicates representing the dataset in the instance comparison task. We review two keys discovery approaches: ***(i)*** The *pseudo-Key* [4], a relaxed version of a key that tolerates a few in-

[9]The FluidOps Data Portal is currently tested by a pilot customer and is available on data.fluidops.net and.

[10]The Linked Data Observatory demo is publicly available according to LOD principles at http://data-observatory.org/lod-profiles/index.htm

[11]The source code and the documentation of the *voiDge* tool can be downloaded on http://hpi.de/naumann/projects/btc/btc-2010.html

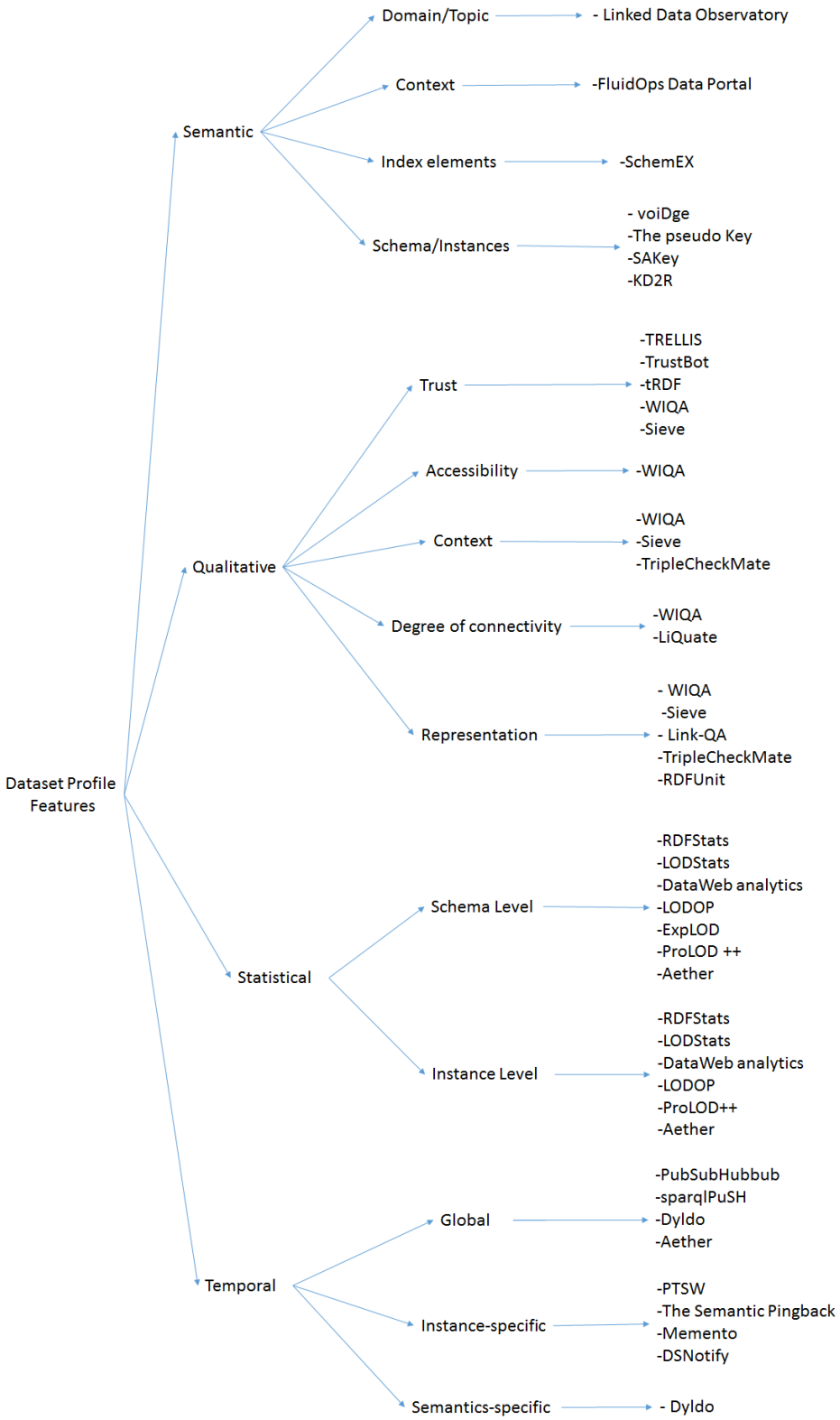

Fig. 1. Overview of dataset profile features and extraction systems.

stances having the same values for the properties, and ***(ii)*** *SAKey*[84] – an approach to discover *almost keys* in datasets where erroneous data or duplicates exist. *SAKey* is an extension of *KD2R*[85] which aims to derive exact composite keys from a set of non keys discovered on RDF data sources. The pseudo-Key and the almost keys approaches mainly differ on the level of the semantic discovery of identity links.

SchemEX[55] is a stream-based indexing and schema extraction approach over LOD. 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 data sources containing instances with corresponding properties.

3.2. Quality Assessment Systems

Zaveri et al. [97] provide an extensive survey of 21 works on linked data quality assessment. In this section, we focus on tools that are implemented and available.

TRELLIS[12] [38] is an interactive environment that examines the degree of trust of datasets based on user annotation. The user can provide Trellis with semantic markup of annotations through interaction with the ACE tool[13] [14]. The tool allows several users to add and store their observations, viewpoints and conclusions. The annotations made by the users with ACE can be used in TRELLIS to detect conflicting information or handle incomplete information.

TrustBot [39] is an Internet Relay Chat bot that make trust recommendations to users based on the trust network it builds. It allows users to transparently interact with the graph by making a simple serie of queries. Users can add their own URIs to the bot at any time, and incorporate the data into the graph. The bot keeps a collection of these URIs that are spidered when the bot is launched or called upon to reload the graph. TrustBot is a semi-automatic tool to gauge the trustworthiness of the data publisher.

tRDF[14] [46] 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. The users can query the dataset and access the trust values associated to the query solutions in a declarative manner.

WIQA[15] [11] 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 provenance information, ratings, and background information about information providers. This framework is composed of two components: a Named Graph Store for representing information together with quality related meta-information, and an engine which enables applications to filter information and to retrieve explanations about filtering decisions. WIQA policies are expressed using the WIQA-PL syntax, which is based on the SPARQL query language.

Sieve[16] [64] is the quality evaluation module in *LDIF* (Linked Data Integration Framework). To assess the quality of a dataset, the user can choose which characteristics of the data indicate higher quality, how this quality is quantified and how should it be stored in the system. This is enabled by a conceptual model composed of assessment metrics, indicators and scoring functions (TimeCloseness, Preference, SetMembership, Threshold and Interval Membership). Sieve aimed mainly to perform data fusion (integration) based on quality assessment.

Link-QA[17] [44] is a framework for detection of the quality of linksets using five network metrics (degree, clustering coefficient, open *sameAs* chains, centrality, description richness through *sameAs*). This framework is completely automatic and takes as input a set of resources, SPARQL endpoints and/or dereferencable resources and a set of triples. The workflow consists of five components: Select of set of, Construct, Extend, Analyse and Compare.

LiQuate[18] [80] 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.

[12] TRELLIS is an open-source tool and available online at http://www.isi.edu/ikcap/trellis/demo.html

[13] Annotation Canonicalization through Expression synthesis

[14] This tools are available online on http://trdf.sourceforge.net/tsparql.shtml

[15] WIQA is an open-source tool and available on http://wifo5-03.informatik.uni-mannheim.de/bizer/wiqa/impl

[16] This tool is open-source and available on http://sieve.wbsg.de/development

[17] Link-QA is Open-source and available on http://bit.ly/Linked-QA.

[18] The demo is published at http://liquate.ldc.usb.ve.

TripleCheckMate[19] [57] is a user-driven quality evaluation tool. The system will provide the user with a list of classes wherein he can choose the ones he is most familiar with. There are three options: ***(i)*** Any: where a completely random resource will be retrieved, ***(ii)*** Class: where one has the option to choose a random resource belonging to that class will be retrieved, ***(iii)*** Manual: where you can manually put in the *DBpedia* URI of a resource of your choice. After selecting a resource, the user will be shown each triple belonging to that resource. The system allow to evaluate each triple whether it contains an error or not. If it contains an error, the user can select an error type from a suggested list.

RDFUnit[20] [56] is a framework for the data quality tests of 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.

3.3. Extraction of Statistical Characteristics

RDFStats[21] [60] is a solid framework for generating statistics from RDF data that can be used for query processing and optimisation over SPARQL endpoints. Thoses statistics include histograms for subjects (URIs, blank nodes) and histograms for properties and associated ranges. RDFStats can be integrated into user interfaces and other Semantic Web applications to provide this information but also to support tools to achieve a better performance when processing large amount of data.

LODStats[22] [6] is a statement-stream-based tool and framework for gathering comprehensive statistics about datasets adhering RDF. The tool calculates 32 different statistical criterions 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 metadata repository, either as a patch or as an external web application using CKAN's API.

Data Web analytics[23] [9] examines the growth of the LOD Cloud since 2007. It provides statistics about the Cloud containing multiple aspects such as the usage of vocabularies as well as provenance and licensing information. The main difference to LODStats is that this information is partially entered manually in the Data Hub and updated infrequently, whereas with LODStats these calculations can be performed automatically.

LODOP[24] [35] is a framework for computing, optimizing, and benchmarking statistics for Linked Datasets. This system provides a total of 56 scripts, which compute 15 different statistical properties across different subsets of the input dataset. This statistical properties are determined via the following types of groupings : by resource, property, class, class and property, datatype, context URL, vocabulary, language, object URI, or no grouping. **ExpLOD** [54] creates usage summaries from RDF graphs including meta-data about the structure of a 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 usage.

ProLOD ++[25] [2] is an interactive user interface, which is divided into a cluster tree view and a details view. The cluster view enables users to explore the cluster tree and to select a cluster for further investigation for statistics. *ProLOD ++* is extension of *ProLOD*[17] which generated basic statistics. In addition to the mining and the cleansing tasks of *ProLOD ++*, the tool generates profiling features like related to key analysis, predicate and value distribution, string pattern analysis, link analysis and data type analysis.

Aether[26] [63]is a tool that generates automatically an extended VoID statistical profile from a $sparql1.1$ endpoint This statistical profile can be viewed in a graphical interface with the viewer module. In addition, *Aether* provides a temporal profile by allowing the comparison between datasets versions and a qualitative profile by detecting outliers and errors. The gen-

[19] TriplecheckMate is open source and available on https://github.com/AKSW/TripleCheckMate.

[20] This framework is on http://aksw.org/Projects/RDFUnit.html

[21] http://rdfstats.sourceforge.net/

[22] https://github.com/AKSW/LODStats/wiki/LODStats-clean-install

[23] http://lod-cloud.net/state/

[24] https://github.com/bforchhammer/lodop/

[25] https://www.hpi.uni-potsdam.de/naumann/sites/prolod++/app.html, https://github.com/HPI-Information-Systems/ProLOD

[26] A demo of *Aether* is available on http://demo.seco.tkk.fi/aether/

erated extensions of the VoID description represent statistics in both schema and instance level.

3.4. Temporal Characteristics Extraction Systems

PubSubHubbub[27] [33] 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.

sparqlPuSH[28] [73] is an interface that can be plugged on any SPARQL endpoint and that broadcasts notifications to clients interested in what is happening in the store using the PubSubHubbub protocol 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: *(i)* register the SPARQL queries related to the updates that must be monitored in a RDF store, *(ii)* broadcast changes when data mapped to these queries are updated in the store.

Ping the Semantic Web (PTSW)[18] is a web service archiving the location of recently created/updated RDF documents. If a document is created or updated, its author can notify PTSW that the document has been created or updated by pinging the service with the URL of the document. This protocol is used by crawlers or other types of software agents to know when and where the latest updated RDF documents can be found. PTSW is dedicated to Semantic Web documents and all the sources they may come from: blogs, databases exported in RDF, hand-crafted RDF files, etc.

The Semantic Pingback[87] is a mechanism that allows users and publishers of RDF content, of weblog entries or of an 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[29] service.

Memento[30][26], [27] is a protocol-based time travel that can be used to access archived representations resources. The current representation of a ressource 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*. Mementos are available both in HTML and RDF/XML.

The Dynamic Linked Data Observatory (Dyldo)[53], [52] 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 (snapshots) over time. The dataset provides weekly crawls of LOD data sources starting from the 2nd of November 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).

DSNotify[31][77] 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.

4. Applications and Application-Driven Profiles

In this section, we discuss dataset profiles from an application point of view. We review existing applications from the areas of data linking and curation, schema inference, as well as query and search along with dataset profile features these applications typically require.

[27] https://code.google.com/p/pubsubhubbub/
[28] https://code.google.com/p/sparqlpush/

[29] Remote procedure call.
[30] A demo of Memento for the DBpedia environment: `http://dbpedia.org/data/Tim_Berners-Lee`, the memento result is `http://mementoarchive.lanl.gov/dbpedia/memento/20090701/http://dbpedia.org/data/Tim_Berners-Lee`.
[31] http://www.cibiv.at/ niko/dsnotify

4.1. Data Linking

Data linking applications aim to annotate, disambiguate and interlink entities and events in text using NLP techniques and external sources including Linked Data. In this context, popular services include Wikipedia-enabled DBpedia Spotlight [25] and Illinois Wikifier [79] as well as more recently developed multilingual annotator Babelfy [69].

Features for data linking applications: Data linking applications typically use semantic features discussed in Section 2.1 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 keys discovery approaches.

4.2. Data Curation, Cleansing, 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 [96]. In the context of Linked Data, statistical approaches to error detection and type prediction are shown to be more effective than the standard ontology reasoning techniques due to their independence of the background knowledge and robustness to noise [75]. Therefore, a number of recent works focus on statistical methods for: (1) Outlier detection to detect errors in numerical values [34], [75], [93]; (2) Automatic prediction of missing types of instances [75]; and (3) Identification of wrong links between datasets [74]. A further line of research in Linked Data quality is related to the discovery of errors in the data based on the existing interlinking (e.g., [19], [95]). Thereby some works go beyond error detection and attempt to automatically determine correct data values in case of inconsistencies [19].

Features for error detection in numerical values: In [34] 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 2.3) 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.

Features for conflict resolution in multilingual DBpedia: The features used in conflict resolution in [19] include provenance metadata at the statement, property and author levels. The temporal dataset profile 1 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.

4.3. Schema Inference

Many existing Linked Data sources do not explicitly specify schemas, or only provide incomplete specifications. However, many real-world applications (e.g., answering queries over distributed data [13]) rely on the schema information. Recently, approaches aimed at automatic inference of missing schema information have been developed (e.g., [75], [55]).

Features for type inference: Statistical characteristics of datasets (see Section 2.3) play an important role in the type inference applications. For example, in [75] 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).

4.4. Distributed Query Applications

Linked Data Cloud can be queried either through direct HTTP URI lookups or using distributed SPARQL endpoints [45] that can include full-text search extensions (see e.g., [1]). Also combinations of both query paradigms are possible [48]. 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 [94]; 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 [45]. The majority of existing federated query approaches for LOD (e.g., [48], [45], [91], [40]) are aimed to optimize for efficient query processing and do not (yet) take quality parameters of LOD sources into account. Therefore, existing *Data Summaries* mostly contain frequencies and interlinking statistics of varying granularity.

Features for efficient and quality-aware query applications: The majority of existing query applications rely on semantic and statistical characteristics (see Sections 2.1 and 2.3) 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 [45]. Finally, quality-aware query applications also take into account qualitative characteristics (see Section 2.2) (e.g., completeness and accuracy) at different granularity levels. This includes overall data source statistics [70], as well as property-specific [81] and type-specific statistics [94].

*4.5. Information Retrieval (*IR*)*

In *IR*, Linked Data is mostly used in the context of semantic search, a typical demonstration of which can be found in [31]. The majority of the semantic search applications are domain-oriented, and a large number of practical cases have been shown for repositories related to biomedical sciences. For example, the concept-based search mechanism [58] allows biologists to describe the topics of the search interest more specifically and retrieve the information with higher precision (in comparison to usage of keywords only). It should be stressed here that the concept-based search requires linking to high-quality external resources (such as, e.g., UMLS [15]), which involves features related to trust , especially verifiability and believability.

The datasets providing semantic features allow one to go beyond standard Bag of Words representation [86]. Wide range of methods based on linking to external, domain-oriented resources has been proposed, e.g., [79], [65], [88]. They also employ statistical features extracted from large-scale text corpora [20] and allow one to expand the user queries to increase recall [8]. In addition, geographical and temporal contexts play an increasingly important role in *IR* applications. These contexts enable retrieval of information relevant with respect to the spatial [51] and temporal [21] dimensions of the query.

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 preserve the semantic search, *IR* implies profiles features like topical domains, and context.

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 1
Adoption of VoID across LOD Datasets per Category (Source: http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/).

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 2.

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)[49] 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[32] 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-Terms, FOAF for general metadata and SCOVO - the Statistical Core Vocabulary[33] for statistical information. VoID is already widely used in the Web of data,

[32] http://vocab.deri.ie/void
[33] http://purl.org/NET/scovo

as documented by Table 1, depicting the use of VoID descriptions among the 1014 datasets and per category in the current inventory of the Web of Data[34].

The Data Catalog Vocabulary (DCAT)[35] follows a similar rationale and has been created based on a survey of government data catalogues[62]. 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. a RDF dump or a SPARQL endpoint). For categorisation of datasets, the *dcterms:subject* predicate and controlled SKOS vocabularies are recommended.

A more specific approach is followed by the Vocabulary of Links (VoL)[36], 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.

5.2. Dataset Quality

Early works by Supelar *et al.* in [83] define a set of knowledge quality features applicable for knowledge graphs, respectively ontologies, and a corresponding ontology. Their features are classified into *quantifiable* and *non-quantifiable* characteristics and include characteristics such as usability, availability, accuracy, or complexity. The suggested ontology, however, only includes a higher level taxonomy, but neither a fully fledged vocabulary for annotation nor a specific set of metrics to quantify the quantifiable metrics.

Fürber *et al.*[37] describe the DQM Ontology[37], 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[38] describe some early work to filter content according 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[36], where authors use the SPARQL Inferencing Notation (SPIN) - a vocabulary that allows the representation of SPARQL queries - to represent data quality rules.

Finally, while provenance information often provides indicators about timelines, currency and update cycles of datasets, Section 5.5 introduces additional vocabularies of relevance.

5.3. 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 manuscript, 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[39] for instance lists a number of vocabularies for representing dataset changeset and updates. The *Talis Changeset vocabulary*[40] provides some early, yet discontinued work

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

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

[36] http://data.linkededucation.org/vol/index.htm

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

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

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

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

on representing changeset and specific characteristics, and has a similar approach as the Delta vocabulary[41]. The *Triplify Update vocabulary*[42] 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.*[43] 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. 5.5). 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*[43], which allows the representation of more abstract dynamics-related observations for a specific dataset. It is specifically foreseen to be used in conjunction with VoID, where a *void:Dataset* is annotated with instantiations of *dady:UpdateDynamics*. The latter captures information about the update regularity and frequency.

For capturing specific features and 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.

5.4. 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[44], SDMX[45] or SCOVO[46].

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[47], currently a W3C Editors Draft developed by the Government Linked Data Working Group[48] 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)[49] 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[50] 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[51], also described by Hausenblas *et al.*[50], 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[24], 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[7], 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

[41] http://www.w3.org/2004/delta
[42] http://triplify.org/vocabulary/update
[43] http://vocab.deri.ie/dady
[44] http://www.w3.org/TR/vocab-data-cube/
[45] http://sdmx.org
[46] http://vocab.deri.ie/scovo
[47] https://dvcs.w3.org/hg/gld/raw-file/default/data-cube/index.html
[48] http://www.w3.org/2011/gld/
[49] http://sdmx.org/
[50] http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html
[51] http://vocab.deri.ie/scovo

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.

5.5. 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[52] 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 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)"*.[53]

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 [68]. 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 [67] 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)**[54] 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**[55] 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 [22] 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

[52] http://www.w3.org/TR/2013/REC-prov-dm-20130430/

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

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

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

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**[56] 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)[57] as a lightweight vocabulary, and the OPM Ontology (OPMO)[58] using more advanced OWL constructs.
8. The **PROV Ontology (PROV-O)**[59] 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 [66] created to support "the widespread publication and use of provenance information of Web documents, data, and resources" – including a Data Model (PROV-DM) [68] and an Ontology (PROV-O) [61] – 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.[60]

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 [47], 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 [47] 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

[56] http://trdf.sourceforge.net/provenance/ns.html
[57] http://purl.org/net/opmv/ns#
[58] http://openprovenance.org/model/opmo
[59] http://www.w3.org/TR/prov-o/
[60] http://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

(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** [23] 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 [72].

5.6. 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)**[61] is a framework that allows users to define the rights regarding how others can reuse the content that the users themselves have published. It provides various licenses to define if and how people can reuse content that has been published, if they can modify it, and if it may be used for commercial purposes. Creative Commons also allows licensing information to be expressed in RDF using the ccREL (REL, or rights expression language) vocabulary. Many datasets in the LOD cloud are already licensed under Creative Commons, as we will see later.
– The **Open Data Commons (ODC)** license[62] 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[63] 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[64] uses RDF or JSON, evolving from an earlier XML-based REL version[65].
– **Open Government License (OGL)**[66] 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)**[67] 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)**[68] is an RDF vocabulary that provides a common metadata modelling scheme for describing projects creating software applications, in order to provide a unified way to represent a software project no matter the source. The main class is *Project* which has properties such as its licence, the project's maintainers, the URL for subversion access, etc. Many of the concepts in DOAP could also be reapplied to datasets since they share many of the same properties.
– **Licenses for Linked Open Data (l4lod)**[69] was introduced in [42] 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 asso-

[61] http://creativecommons.org/licenses/by/3.0/
[62] http://opendatacommons.org/licenses/
[63] http://www.w3.org/community/odrl/two/model/
[64] http://w3.org/ns/odrl/2/
[65] http://www.w3.org/TR/odrl/
[66] http://www.nationalarchives.gov.uk/doc/open-government-licence/
[67] http://purl.org/LiMo/0.1
[68] http://usefulinc.com/ns/doap
[69] http://ns.inria.fr/l4lod/

Vocabulary Name	Type	Overall	Datasets
Dublin Core	General, Provenance	21,397,721	154
FOAF	General, Provenance	3,689,178	117
SKOS	General	10,581,530	67
VoID	General	9,754	41
voidp	Provenance	172	21
SIOC	Provenance	148	16
DOAP	Licensing	306	14
Creative Commons	Licensing	16,525	12
Provenance Vocabulary	Provenance	84	12
Data Cube	Statistical	581,381	10
SCOVO	General, Statistical	408	9
PML	Provenance	259	8
OPMO	Provenance	63	8
SDMX	Statistical	285,904	6
OPMV	Provenance	4	2
PROV-O	Provenance	4,537	1
DCAT	General	8	1
Waiver	Licensing	1	1
Delta	Dynamics	0	0
RMO	Dynamics	0	0
Triplify	Dynamics	0	0
ChangeSet	Dynamics, Provenance	0	0
VoL	General	0	0
l4lod	Licensing	0	0
LiMo	Licensing	0	0
ODC	Licensing	0	0
ODRL	Licensing	0	0
OGL	Licensing	0	0
PREMIS	Provenance	0	0
DQM	Quality	0	0
SPIN	Quality	0	0
WIQA	Quality	0	0

Table 2
Overall usage and dataset counts for the aforementioned vocabularies.

ciated with heterogeneous datasets whose information items have been returned together for consumption (e.g. via a single SPARQL query).

5.7. Vocabulary Usage: Analysis and Recommendation

Based on the usage frequency of the aforementioned vocabularies, we can make some recommendations as to their usage for datasets. We use the LOD2 Stats service[70] 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 2, where the type refers to the vocabulary type as per the headings above.[71]

251 datasets use RDF syntax, giving us an overall total. From the data in Table 2, we observe that general metadata about the datasets is readily provided, but that more specific information on provenance and

[70] http://stats.lod2.eu/ as accessed on 2nd February 2015

[71] 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).

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 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 2 that many datasets are explicitly licensed via some machine-readable form, with just 5% (12) containing Creative Commons metadata. However, according to work by [42], 95% of the datasets in the LOD cloud[72] 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. Conclusions

In this paper, we have provided a comprehensive survey of existing research aimed at supporting the dataset profiling task, a central challenge when facilitating dataset discovery in tasks such as entity retrieval, distributed search or entity linking. Given the complexity of the topic, we have focused on first providing an exhaustive taxonomy of dataset features, also available as structured RDF vocabulary, and then surveyed and assessed methods for assessing and extracting such features from arbitrary datasets, vocabularies for representing these features, preferrably as Linked Data, and applications which make use of dataset profiles. Whereever feasible, we also provided insights into the adoption and impact of the discussed works and offer suggestions, for instance, for chosing vocabularies when representing dataset profiling features.

While this has been the first study of its kind, we are currently looking into expanding some of its results in order to simplify adoption of some of the presented approaches. For instance, while we already provide a structured RDF vocabulary describing dataset profiling features which can be used to identify features as part of a dataset description, for instance, with VoID, no explicit mappings with available vocabularies (Section 5) have been provided yet. In order to simplify dataset representation, we are currently working on providing explicit mappings between our dataset profiling features vocabulary and existing vocabularies meant for describing individual features.

Given the continuous evolution and expansion of the Web of data, we assume that the problem of dataset profiling will become an increasingly important 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.

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

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] Manuel Atencia, Jérôme David, and François Scharffe. Keys and pseudo-keys detection for web datasets cleansing and interlinking. In *Knowledge Engineering and Knowledge Management - 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012. Proceedings*, pages 144–153, 2012.

[5] 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.

[6] 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.

[7] 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.

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

[9] Chris Bizer, Anja Jentzsch, and Richard Cyganiak. State of the lod cloud. *Version 0.3 (September 2011), http://lod-cloud.net/state/*, 1803, 2011.

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

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

[12] 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.

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

[14] 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.

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

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

[17] 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.

[18] Uldis Bojars, Alexandre Passant, Frederick Giasson, and John G. Breslin. An architecture to discover and query decentralized RDF data. In *Proceedings of the ESWC'07 Workshop on Scripting for the Semantic Web, SFSW 2007, Innsbruck, Austria, May 30, 2007*, 2007.

[19] 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.

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

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Herbert Van de Sompel, Michael L. Nelson, Robert Sanderson, Lyudmila Balakireva, Scott Ainsworth, and Harihar Shankar. Memento: Time travel for the web. *CoRR*, abs/0911.1112, 2009.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

2014., 2014.

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

[37] 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.

[38] 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.

[39] Jennifer Golbeck, Bijan Parsia, and James A. Hendler. Trust networks on the semantic web. In *Cooperative Information Agents VII, 7th International Workshop, CIA 2003, Helsinki, Finland, August 27-29, 2003, Proceedings*, pages 238–249, 2003.

[40] 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.

[41] 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.

[42] 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.

[43] Markus Graube, Stephan Hensel, and Leon Urbas. R43ples: Revisions for triples - an approach for version control in the semantic web.

[44] Christophe Guéret, Paul T. Groth, Claus Stadler, and Jens Lehmann. Assessing linked data mappings using network measures. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, pages 87–102, 2012.

[45] 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.

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

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

[54] 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.

[55] 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.

[56] 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.

[57] Dimitris Kontokostas, Amrapali Zaveri, Sören Auer, and Jens Lehmann. Triplecheckmate: A tool for crowdsourcing the quality assessment of linked data. In *Knowledge Engineering and the Semantic Web - 4th International Conference, KESW 2013, St. Petersburg, Russia, October 7-9, 2013. Proceedings*, pages 265–272, 2013.

[58] 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.

[59] Sarasi Lalithsena, Pascal Hitzler, Amit P. Sheth, and Prateek Jain. Automatic domain identification for linked open data. In *2013 IEEE/WIC/ACM International Conferences on Web Intelligence, WI 2013, Atlanta, GA, USA, November 17-20, 2013*, pages 205–212, 2013.

[60] Andreas Langegger and Wolfram Wöß. Rdfstats - an extensible RDF statistics generator and library. In *Database and Expert Systems Applications, DEXA, International Workshops, Linz, Austria, August 31-September 4, 2009, Proceedings*, pages 79–83, 2009.

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

[62] 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.

[63] Eetu Mäkelä. Aether–generating and viewing extended void statistical descriptions of rdf datasets. In *The Semantic Web: ESWC 2014 Satellite Events*, pages 429–433. Springer, 2014.

[64] Pablo N. Mendes, Hannes Mühleisen, and Christian Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, pages 116–123, 2012.

[65] 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.

[66] 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.

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

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

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

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

[71] Felix Naumann. Data profiling revisited. *SIGMOD Record*, 42(4):40–49, 2013.

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

[73] 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.

[74] 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.

[75] 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.

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

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

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

[87] 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.

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

[89] Andreas Wagner, Peter Haase, Achim Rettinger, and Holger Lamm. Discovering related data sources in data-portals. In *Proceedings of the First International Workshop on Semantic Statistics, co-located with the the International Semantic Web Conference*, 2013.

[90] 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.

[91] 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.

[92] 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.

[93] 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.

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

[95] 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.

[96] 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.

[97] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment methodologies for linked open data. In *Under Review*, 2014.