Linked 'Big' Data: Towards a Manifold Increase in Big Data Value and Veracity

Jeremy Debattista
Enterprise Information Systems – University of Bonn
Organized Knowledge – Fraunhofer IAIS
Bonn, Germany
debattis@iai.uni-bonn.de

Simon Scerri
Enterprise Information Systems – University of Bonn
Organized Knowledge – Fraunhofer IAIS
Bonn, Germany
scerri@iai.uni-bonn.de

Christoph Lange
Enterprise Information Systems – University of Bonn
Organized Knowledge – Fraunhofer IAIS
Bonn, Germany
langec@iai.uni-bonn.de

Sören Auer
Enterprise Information Systems – University of Bonn
Organized Knowledge – Fraunhofer IAIS
Bonn, Germany
auer@iai.uni-bonn.de

***Abstract*—The Web of Data is an increasingly rich source of information, which makes it useful for Big Data analysis. However, there is no guarantee that this Web of Data will provide the consumer with truthful and valuable information. Most research has focused on Big Data's *Volume*, *Velocity*, and *Variety* dimensions. Unfortunately, *Veracity* and *Value*, often regarded as the fourth and fifth dimensions, have been largely overlooked. In this paper we discuss the potential of Linked Data methods to tackle all five V's, and particularly propose methods for addressing the last two dimensions. We draw parallels between Linked and Big Data methods, and propose the application of existing methods to improve and maintain quality and address Big Data's veracity challenge.**

***Keywords*-Linked Data, Web of Data, Veracity, Value, Big Data Dimensions**

I. Introduction

Advancements in Web applications (e.g. social networks) and hardware compliant with the Internet of Things concept, brought a change to data processing and analytics methods. This is due to the vast amount of data being generated continuously, called *Big Data*, where data "is characterised not only by the enormous volume or the velocity of its generation but also by the heterogeneity, diversity and complexity of the data"[18]. Originally Big Data has been described as a three-dimensional model, citing *Volume*, *Variety* and *Velocity* as the three major characteristics – and challenges – for effectively generating value. Later, *Veracity* was introduced as a fourth, frequently overlooked but equally pressing dimension. The Veracity dimension deals with the uncertainty of data due to various factors such as data inconsistencies, incompleteness, and deliberate deception.

Value is often regarded as a fifth dimension, addressing the need to enrich raw and unprocessed data by extracting higher-level knowledge for use across different scenarios. However, the term remains confusing since the generation of actual value from Big Data is achievable only after all five described V's have been successfully addressed. Therefore, we distinguish between *Value* as big data's fifth dimension which requires specific attention, and *big data value* as the ultimate goal after carefully addressing all five dimensions. In this paper we discuss the potential of Linked Data methods to tackle all five V's, and propose methods for addressing the *value* and *veracity* dimensions which have remained largely unaddressed by previous research.

With around 5 billion documents[1], the Web is the largest crowdsourcing effort known to mankind. Taking Wikipedia for example, the open encyclopedia has almost 5 million articles written in English, and a total of 290 official languages. Such an effort requires human intervention to add, modify, and create translations to a big knowledge base. Having said that, it is a known fact that wiki articles may have inconsistencies over the same article in different languages, un-cited claims (often marked on Wikipedia), and completely false information as pages can be easily modified by untrusted sources.

Moving towards a more structured web of data (Linked Data), the Schema.org[2] effort was initiated in order to create a common set of schemas to encourage web publishers to use structured data, using formats such as Microdata, RDFa or JSON-LD. With structured data, documents become semantically enriched resources and can be easily related to other resources. The return on investment of such practices is high, in the sense that the web content is shifted from **(raw) data** to **information resources**, thus drastically increasing the value. These resources can then easily be transformed into machine-comprehensible knowledge with the right tools. In 2013, Facebook introduced the Facebook Graph Search, a semantic search engine based on natural language meaning rather than just keywords. Similarly, Google

[1] Source: http://worldwidewebsize.com – Thursday 25th June 2015
[2] https://schema.org

978-0-7695-5696-3/15 $31.00 © 2015 IEEE
DOI 10.1109/BDC.2015.34

Authorized licensed use limited to: Pontificia Universidade Catolica de Campinas (PUC Campinas). Downloaded on April 22,2023 at 03:51:00 UTC from IEEE Xplore. Restrictions apply.

have their own knowledge graph that enhances a search result with other related information. These two mentioned examples, amongst others, are applications of what it is known as Web 3.0, or even better the Semantic Web.

Linked Data [5] is described by Tim Berners-Lee as "Semantic Web done right"[3]. With Linked Data, unstructured content of web documents can be enriched semantically using schemas such as Schema.org and connected to other data according to the four principles defined in [4]. Metadata, such as who created a resource, who edited a resource, privacy, licensing and provenance information, and so on, are nowadays easier to attach to web pages using suitable Linked Data vocabularies. Linked Data can be serialised in different formats, including text, JSON and XML. This flexibility enables applications to consume data without expensive pre-processing, and thus facilitates interoperability between machines, which is one of objectives of the Semantic Web. One popular example of Linked Data on the Web is DBpedia [17], a semantic multi-lingual knowledge base extracted from Wikipedia.

In [14], Hitzler and Janowicz state that Linked Data indeed is part of the Big Data overall perspective. They go on to claim that Linked Data can be used to help solve certain challenges in Big Data. During these last years, research topics in Linked Data were closely tied to challenges related to Volume, Variety and Velocity. Some examples of such challenges include:

Distributed Storage and Querying – One problem related to the *Volume* of the data is the storage. Various research works such as [11, 33] show how RDF data can efficiently be stored and queried in a distributed fashion.

Federated Querying – This is the perfect example related to the *Variety* challenges, where a query is distributed over different sources to answer a complex question. Optimisation of such querying (e.g [26]) is a focus of this area.

Stream Querying and Reasoning – Real-time structured data with a high *Velocity* cannot be queried or reasoned upon using traditional tools such as standard SPARQL query processors. Extensions such as C-SPARQL [2] enable querying real-time RDF data.

Nevertheless, the value in Linked Data is also compromised by concerns about low quality and *Veracity*. In this paper, we draw parallels between Linked and Big Data methods, and propose the application of existing Linked Data methods addressing quality to also target Big Data's veracity. In addition, we point out that various existing Linked Data enrichment methods are highly adequate to addressing Big Data's *Value* dimension, and that transforming Big Data into a Linked Data-compliant format will further enable its enrichment. Thus, we present Linked Data as a holistic approach to increase Big Data value by addressing all five dimensions.

The main contributions of this article are two-fold:

- providing an insight into Linked Data tools and proposing a Semantic Pipeline that enable value creation out of raw data, therefore improving the Big Data *Value* dimension (cf. Section III);
- enabling the assessment of various quality factors in Linked Data to help identify and improve the Big Data *Veracity* dimension (cf. Section IV).

In Section II we discuss the importance of the Big Data value and veracity dimensions. Final remarks are in Section V.

II. Big Data Value and Veracity: The Journey So Far

In this section we discuss work related to initiatives and roadmaps towards improving the Big Data value and veracity dimensions.

In [19], Lukoianova and Rubin argues in favour of the importance of the veracity dimension in Big Data due to the lack of attention paid to the dimension so far. The authors state that irrelevantly of the processes used to collect data, the input could still suffer from biases, ambiguities and lack of accuracy. Lukoianova and Rubin propose a roadmap towards defining veracity in three dimensions: (1) Objectivity/Subjectivity; (2) Truthfulness/Deception; (3) Credibility/Implausibility. In light of these definitions, the authors also propose (1) a number of quality dimensions, some of which can be mapped to their Linked Data counterparts surveyed in [32], (2) a set of NLP tools that can be used to assess these dimensions, and (3) an index measure for big data veracity.

The authors of [22] discuss the possibilities and challenges in creating trust in Big Data. This work is similar to [19], though putting a stronger focus on trust issues. The authors raise a number of trust-related research questions in four different domains, namely trust in data quality, measuring trust, trust in nodes (i.e. systems in a network), and trust in providers, providing known research efforts. For the first two domains, the authors put forward questions on how input data can be assessed for quality measures, how trust can be measured, and how can trust be increased.

The importance of high quality data was also an issue in the Business & Information Systems Engineering editorial [8]. Buhl et al. state that data availability is one of the most important features for Big Data, together with data consistency regarding time and content, completeness, comprehensiveness and reliability. The authors also claim that in order to have high quality big data, the data should have meaning, i.e. value, so as to enable methods to derive new knowledge. They conclude that the "challenge of managing data quality" should be part of the Big Data nucleus, whilst the creation of data value should be pivotal in business models.

[3] http://www.w3.org/2008/Talks/0617-lod-tbl/#(3)

Authorized licensed use limited to: Pontificia Universidade Catolica de Campinas (PUC Campinas). Downloaded on April 22,2023 at 03:51:00 UTC from IEEE Xplore. Restrictions apply.

A report by IBM [24] discusses the real-world use of big data and how can value be extracted in uncertain data. Schroeck et al. describe that advancements in techniques (such as analytic techniques) enabled entities to make use of the huge amount of data to extract new insights, fast and accurate. On the other hand, similarly to Buhl et al. [8], the authors state that these entities should consider veracity as an important dimension, by adding meaning to the data. According to the authors this can be achieved through data fusion tools or fuzzy logic approaches. Participants in the IBM survey [24] also stated that the integration component on big data platforms is a core component to enable good big data analytic efforts, which in return creates value.

III. Linked Data – Enhancing the Value of Raw Data

In May 2007 the first version of the Linked Open Data cloud (LOD Cloud) [15] was initially published as a set of 12 datasets, following the Linked Data principles[4]. During the last eight years the LOD Cloud has grown at a fast pace, where around 1014 datasets from 570 providers (as of August 2014) have been added. All of these data providers use Linked Data as common standard representation for interoperability.

The term "Linked Data" refers to a set of best practices for publishing and connecting structured data on the Web through URIs, HTTP and RDF [5]. Big Data modelled according to the Linked Data principles can be considered as **structured** Big Data: the same five V's apply; in particular, the overall LOD Cloud amounts to a voluminous dataset (*Volume*), and the rise of streaming RDF data and sensor data has resulted in the continuous addition of semantic data-in-motion (*Velocity*).

The *Variety* dimension is a fundamental characteristic of Linked Data, where the flexible, extensible and self-documenting RDF data model enables different domains to be described in a standardised **interoperable** format. Thus, the issue of *Variety in formats* is no longer pressing, since, by definition, Linked Data is a uniform representation. Thanks to its built-in mechanisms for linking not only data but also aligning data *schemas* (here also called vocabularies), Linked Data also makes *Variety in domains* less of a concern.

Traditionally, Big Data is mainly unstructured, therefore it might inhibit consumers from reusing and exploiting it to create *Value*, for example to lead out analytics or to create meaningful visualisations. If Big Data is transformed into Linked Data, the existing methods for enriching Linked Data can readily be applied, thus automatically lowering the barrier towards increasing its value.

In Figure 1 we propose a *Semantic Pipeline* as a basis for addressing the value dimension. Starting from some data found on the Web, this data passes through four steps, namely *Semantic Lifting*, *Metadata Enrichment*, *Data Integration* and *Reasoning*. These are discussed further in the remainder of this section.

A. Semantic Lifting

Semantic Lifting is the process of transforming unstructured, semi-structured or even structured data into a semantically enriched information using RDF schemas as vocabularies. Nowadays a lot of open source conversion tools offer Linked Data as one possible output. A key advantage of the RDF data model is that the structure of the schema and the data is the same, i.e. both are represented as subgraphs of one graph. Therefore, for example, querying the data together with the schema can easily be done in the same query language (SPARQL).

Schemas can either be reused (which is encouraged) or created. The Linked Open Vocabularies initiative[5] has listed around 500 vocabularies for reuse. This is the first step towards making data meaningful and valuable. In the Linked Data context, "valuable" means enabling machines to extract meaning, and reusing resources by linking them.

Consider the following two snippets in CSV[6]:

```
Time,Event,Type,Presenter,Location
...
27 Aug 2014 09:00,Wikidata,Keynote,Markus Krötzsch,
27 Aug 2014 10:15,Working with Wikidata: A Hands-on
     Guide for Researchers and Developers,Tutorial,
     Markus Krötzsch,
```

```
Name,Affiliation,Town,Country
...
Markus Krötzsch,TU Dresden,Dresden,Germany
```

The problem with this kind of representation is that the data, albeit semi-structured, has no meaning that is explicit to a machine. For example, a machine would not be able to know that the same speaker (i.e. Markus Krötzsch) will be presenting twice. Furthermore, the machine would not understand that this same person is affiliated with the Technical University of Dresden in Germany. In tutorial from which this example is taken, we explored the limits of enriching the value of these CSV data, in particular how to make more sense out of it [16] – only to realise that at a certain point the same meaning can be expressed more intuitively using Linked Data and RDF. Using Linked Data, Listings III-A and III-A can be transformed as follows:

[4] http://www.w3.org/DesignIssues/LinkedData.html

[5] http://lov.okfn.org/dataset/lov/

[6] Obtained from [16]

Authorized licensed use limited to: Pontificia Universidade Catolica de Campinas (PUC Campinas). Downloaded on April 22,2023 at 03:51:00 UTC from IEEE Xplore. Restrictions apply.

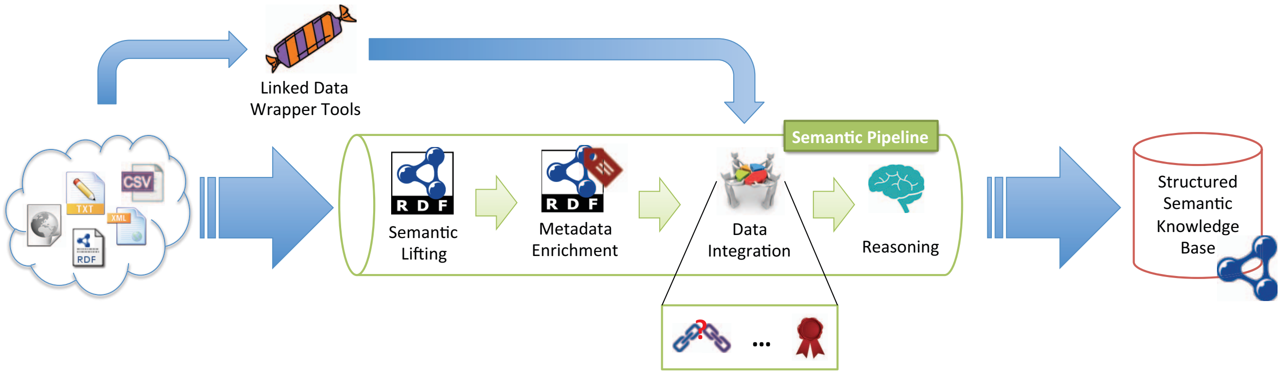

Figure 1. A Semantic Pipeline for enhancing the value in Big Data

```
@prefix ex: <http://purl.org/net/wiss2014/> .
# ... further prefix bindings

<ex:presenters/#MarkusKrötzsch> rdf:type foaf:Person
    ;
  ex:hasAffilation dbpedia:
       Dresden_University_of_Technology ;
  foaf:name "Markus Krötzsch" .

<ex:schedule/#Keynote1> rdf:type ex:Keynote ;
  ex:presenter <ex:presenters/#MarkusKrötzsch> ;
  ex:title "Wikidata" ;
  dcterms:date "2014-08-27T09:00:00"^^xsd:dateTime .

<ex:schedule/#Tutorial1> rdf:type ex:Tutorial ;
  ex:presenter <ex:presenters/#MarkusKrötzsch> ;
  ex:title "Working with Wikidata: A Hands-on Guide
       for Researchers and Developers" ;
  dcterms:date "2014-08-27T10:15:00"^^xsd:dateTime .
```

Listing 1. RDF Representation of Listings III-A and III-A in Turtle

With such a representation, a machine can understand that Markus Krötzsch will be presenting both sessions, and that he is affiliated with the University of Technology in Dresden. `dbpedia:Dresden_University_of_Technology` is a semantic resource in DBpedia, thus a machine can infer more knowledge, such as the country and the year the university was established. Therefore value is not added only through semantic enriching of the data itself, but also through reuse by linking to external and internal resources (such as `<ex:person/MarkusKrötzsch>`). Thanks to specifications such as RDFa, Linked Data can also be embedded into HTML pages to make their content useful for both human readers and machine services. The Schema.org vocabulary is being promoted explicitly with this use case in mind.

B. Promoting Metadata into Valuable Data

Data can be made more valuable (retrievable, comprehensible, reusable, ...) by providing expressive *metadata* using a standardised metadata vocabulary. Sophisticated metadata are often more complex than simple key/value lists. For example, descriptions of provenance, such as what activity, involving what agents, led to the creation of the artefact being described, form knowledge graphs of their own (for example the W3C PROV standard [12]). More than, e.g., relational databases or XML, the RDF data model is capable of handling this blurring distinction between metadata and data, since metadata is described in the same way as the data itself.

C. Link Discovery and Interlinking

The LOD Cloud is clustered into a variety of domains. All datasets are represented in the same standard data model; however, different datasets typically employ different identifiers for the same things (e.g. the same person) and might also use different schemas. Since RDF is represented as graphs, it is easy to create edges between similar nodes in different datasets (or graphs). Similar nodes between different datasets can be discovered through semantic matching. For example, a tourist office can link guided walking tour information with geolocation data, by matching the name of the tour places of interests against GPS locations. This discovery would enable further analysis and possibly more innovative use of the data itself, including its exploitation for business purposes, as in finding the most interesting route taking the least time.

Tools helping this link discovery include Silk [30] and LIMES [20]. Silk[7] is a flexible link discovery framework allowing users to define linkage heuristics using a declarative language. Similarly, LIMES[8] is a large-scale link discovery framework based on metric spaces. Another approach for link discovery is DHR [13], where, unlike Silk and LIMES, the links are discovered by closed frequent graphs and similarity matching.

D. Data Integration

The amount of new data being generated every second is resulting in an even bigger information overload, making it hard to find the needle in the haystack. Data integration is the process of incorporating heterogenous data from different sources into a central, canonical data source, thus establishing a single entry point with only the required information.

[7] http://silk-framework.com
[8] http://aksw.org/Projects/limes

Authorized licensed use limited to: Pontificia Universidade Catolica de Campinas (PUC Campinas). Downloaded on April 22,2023 at 03:51:00 UTC from IEEE Xplore. Restrictions apply.

Data integration is a higher level process, which incorporates tasks such as link discovery and quality assessment. Unlike mere link discovery, data integration does not create links between the datasets being integrated.

LDIF[9] is a framework for Linked Data Integration. LDIF translates linked datasets into a clean, local target representation, keeping track of the provenance [25].

To integrate a linked dataset with other sources, the sources can have different formats. Ontology-based data access (OBDA) can also assist in combining data from multiple heterogenous sources through a conceptual layer [9]. Therefore, using OBDA, linked datasets can be integrated together with datasets having different formats. Ontop [1] is an example of an OBDA framework, where relational databases are queried as virtual RDF graphs. Similarly, Ultrawrap [27] encodes a relational database into an RDF graph using SQL views, allowing SPARQL queries to be executed on these RDF graphs without requiring Extract-Transform-Load (ETL). Another such NoETL approach is SparqlMap [28], which is a SPARQL-to-SQL rewriter based on the W3C R2RML specification[10]. A survey of further OBDA approaches can be found in [31].

E. Reasoning

Reasoning enables consumers to derive new facts that were not previously explicitly expressed in a knowledge base, usually using the simple RDF Schema logic, description logic as implemented in the OWL Web Ontology Language, or first order predicate logic. Thanks to the wide tool support, e.g., by the OWL reasoner HermiT[11], reasoning is one of the main selling points of Semantic technologies. In Linked Data, reasoning poses a number of challenges, as mentioned by Polleres et al. in their lecture notes [21]. These challenges include the existence of inconsistent data and the evolution of the data. Solutions for reasoning over big linked datasets and streaming linked data are available [21, 7, 10].

IV. Linked Data – Identifying Veracity in Big Data

The fourth dimension of Big Data, *Veracity*, is more challenging than the volume, variety, velocity and value. It is defined as *"conformity with truth or facts"* by the Merriam-Webster dictionary. This definition makes the veracity dimension difficult to measure, as truth is not always known objectively. In practice, a consumer might doubt data from producer A but trust data from producer B. On the other hand, another consumer might actually doubt the truthfulness of producer B, but blindly trust producer A.

Zaveri et al. present a comprehensive survey of quality metrics for linked open datasets [32]. Most of the quality metrics discussed are *deterministic* and computable within *polynomial time*. On the other hand, once these metrics are computed on large datasets, the algorithms' upper bound grows and, as a result, the computation becomes intractable time-wise. In this section we describe eight quality metrics (extracted from [32]) that we deem relevant for the veracity challenge in relation to Linked Data, and probabilistic approximation techniques that can be used to compute them efficiently while still achieving a high precision.

A. Reservoir sampling

Reservoir sampling is a statistics-based technique that facilitates the sampling of evenly distributed items. The sampling process randomly selects k elements ($\leq n$) from a source list, possibly of an unknown size n, such that each element has a k/n probability of being chosen [29]. Therefore, the size of k affects the computation time and the accurateness of the result. The reservoir sampling technique is part of the *randomised algorithms* family, that offer simple and fast solutions for time-consuming counterparts by implementing a degree of randomness.

1) Dereferenceability Metric: HTTP URIs should be dereferenceable, i.e. HTTP clients should be able to retrieve the resources identified by a URI by download from that URI. A typical web URI resource returns a `200 OK` code indicating that a request was successful and a `4xx` or `5xx` code if the request was unsuccessful. In Linked Data, a successful request should return an RDF document containing triples that describe the requested resource. Unless URIs include a hash (#), they should respond with a `303 Redirect` code [23]. Dereferenceable URIs ensure that the data is complete in a sense that data consumers are able to retrieve machine-comprehensible representations of all resources of interest. This metric computes the ratio of correctly dereferenceable URIs.

2) Entities as Members of Disjoint Classes: A consistent dataset should be free of any logical or formal contradictions, both at an explicit level and an implicit level. Consistency is a central challenge for the veracity dimension as it ensures correctness of data. In the Web Ontology Language (OWL), classes can be defined as disjoint, meaning that they do not have any entities as common members. In practice, however, entities in a dataset may accidentally end up as members of disjoint classes.These would cause conflicting information and data consumers would not be able to correctly interpret the data, thus decreasing veracity and value of the data. This metric checks if the defined types (and their super types) for a resource are explicitly declared disjoint in the schema.

B. Bloom Filters

A Bloom Filter is a fast and space efficient bit vector data structure commonly used to query for elements in a set ("is element A in the set?"). The size of the bit vector plays an important role with regard to the precision of the

[9] http://ldif.wbsg.de/

[10] http://www.w3.org/TR/r2rml/

[11] http://hermit-reasoner.com

Authorized licensed use limited to: Pontificia Universidade Catolica de Campinas (PUC Campinas). Downloaded on April 22,2023 at 03:51:00 UTC from IEEE Xplore. Restrictions apply.

result. A set of hash functions is used to map each item added to be compared to a corresponding set of bits in the array filter. The main drawback of Bloom Filters is that they can produce *false positives*, therefore possibly identifying an item as existing in the filter when it is not, but this happens with a very low probability. The trade-off of having a fast computation yet a very close estimate of the result depends on the size of the bit vector. With some modifications, Bloom Filters are useful for detecting duplicates in data streams, generating a low *false positive* rate during the process [3].

1) Extensional Conciseness: A linked dataset is extensionally concise if there are no redundant instances. Having duplicate records with different IDs can cause doubtfulness, and thus lower the veracity of a dataset. This metric measures the number of unique instances found in the dataset. The uniqueness of instances is determined from their properties and values. An instance is unique if no other instance (in the same dataset) exists with the same set of properties and corresponding values.

C. Traditional Techniques

Some metrics do not require any probabilistic approximation techniques since their computation is straightforward, such as pattern matching some predicate in the dataset's triples. In this section we will describe a number of simple metrics for Linked Data quality with regard to various veracity aspects.

1) Endpoint and Data Dump Availability: One of the most important aspects in both Linked Data and Big Data is availability. Data should be available for consumption from both SPARQL endpoints (i.e. standardised query interfaces) and data dumps. Whilst the latter makes available a snapshot (taken at different intervals such as hourly, daily, weekly etc.) of the data, an endpoint provides data consumers with the most recent (possibly up-to-date) version of a dataset. The URLs of such endpoints should be discoverable from the dataset's metadata. The VoID[12] and DCAT[13] schemas, both W3C recommendations, are used to describe Linked Datasets and data catalogues in order to bridge dataset publishers and consumers. Both vocabularies have properties that allow the attachment of a SPARQL endpoint URI and a data dump URI to a dataset. Having the right SPARQL endpoint described in the dataset's metadata ensures that the consumed data is provided from the intended source.

2) Misused OWL Datatype and Properties: Similarly to entities as members of disjoint classes, this metric is also part of the "consistency" dimension of quality. This metric checks if the value of a resource's property is correctly defined as an object or data literal, as intended by the vocabulary maintainer. An Object Property (`owl:ObjectProperty`) expects a resource as its value, whilst a Datatype Property (`owl:DatatypeProperty`) expects data of a simple type (e.g. string, integer, date/time) as its value. Misusing a property creates inconsistencies in the data and inhibits the consumer from appropriate reuse.

3) Usage of Undefined Classes and Properties: Using classes and properties without any formal definition, i.e. without a definition in any schema, makes resources meaningless, since data consumers cannot understand the semantics of data. This further makes the data incoherent and thus reduces the effectiveness of tools such as reasoners. This metric checks that each resource type (the object attached to the `rdf:type` property) and each property attached to the resource can be dereferenced to a definition in some schema.

4) Usage of Blank Nodes: Blank nodes are local identifiers given during the publishing process, which are not URIs. These identifiers cannot be externally referenced and thus are useless to the data consumer. Blank nodes furthermore make merging to different sources a difficult task [6]. Therefore, high usage of blank nodes reduces the value of a dataset. This metric computes the ratio of blank nodes in a dataset.

V. Final Remarks

As discussed in Section II, *Value* and *Veracity* are paramount to successful Big Data technology. This article focuses precisely on these two dimensions, and how Linked Data methodology can help better addressing them.

In Section III we argue that Linked Data is the most suitable data model for increasing the value of data over conventional Web formats such as CSV, thus contributing towards the value challenge in Big Data. Linked Data converts raw data into information based on RDF schemas, enabling data interoperability between machines. Based on this concept, we also presented our vision for a semantic pipeline that creates a semantically rich knowledge base from heterogenous formats available on the Web using a variety of techniques.

Section IV addressed the veracity aspect of Big Data. Defined as *"conformity with truth or facts"*, we described eight Linked Data quality metrics (a subset of [32]) which can (1) improve a consumer's perspective with regard to the data, and (2) become an enabler of a data cleaning process that in return improves the dataset's value. We also described two techniques, reservoir sampling and Bloom Filters, that can be used on quality metrics whose time complexity becomes intractable when assessing big datasets.

In the future we plan to implement more quality metrics to cover more aspects of Linked Data quality. We will also implement and evaluate the proposed semantic pipeline *as-a-service*, in the hope that it enables consumers to easily transform their Big Data into *Linked Data-ready* datasets.

Acknowledgment

This work is supported by the European Commission under the Seventh Framework Program FP7 grant 601043

[12] http://www.w3.org/TR/void/

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

Authorized licensed use limited to: Pontificia Universidade Catolica de Campinas (PUC Campinas). Downloaded on April 22,2023 at 03:51:00 UTC from IEEE Xplore. Restrictions apply.

(http://diachron-fp7.eu) and the Horizon 2020 grant 644564 (http://www.big-data-europe.eu).

References

1. Bagosi, T. et al. The Ontop Framework for Ontology Based Data Access. In: *The Semantic Web and Web Science – 8th Chinese Conference, Wuhan, China, August 8-12, 2014, Revised Selected Papers*. 2014.
2. Barbieri, D. F. et al. Querying RDF Streams with C-SPARQL. In: SIGMOD Rec. 39(1) (Sept. 2010), pp. 20–26. http://doi.acm.org/10.1145/1860702.1860705.
3. Bera, S. K. et al. Advanced Bloom Filter Based Algorithms for Efficient Approximate Data De-Duplication in Streams. In: (2012).
4. Berners-Lee, T. Linked Data. http://www.w3.org/DesignIssues/LinkedData.html. July 2006.
5. Bizer, C., Heath, T., Berners-Lee, T. Linked data – the story so far. In: Int. J. Semantic Web Inf. Syst. (2009).
6. Bizer, C., Cyganiak, R., Heath, T. How to publish Linked Data on the Web. Web page. Revised 2008. Accessed 22/02/2015. 2007. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.
7. Bonatti, P. A. et al. Robust and scalable Linked Data reasoning incorporating provenance and trust annotations. In: J. Web Sem. 9(2) (2011), pp. 165–201.
8. Buhl, H. U. et al. Big Data. In: Business & Information Systems Engineering 5(2) (2013), pp. 65–69. http://dx.doi.org/10.1007/s12599-013-0249-5.
9. Calvanese, D. et al. Ontologies and databases: The DL-Lite approach. In: *Reasoning Web*. LNCS 5689. 2009.
10. Chevalier, J. A Linked Data Reasoner in the Cloud. In: *ESWC*. Ed. by P. Cimiano et al. Vol. 7882. LNCS. Springer, 2013, pp. 722–726.
11. Farhan Husain, M. et al. Storage and Retrieval of Large RDF Graph Using Hadoop and MapReduce. In: *Proceedings of the 1st International Conference on Cloud Computing*. CloudCom '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 680–686.
12. Gil, Y. et al. PROV Model Primer. W3C Working Group Note. World Wide Web Consortium (W3C), 30th Apr. 2013. http://www.w3.org/TR/prov-primer/.
13. Hau, N., Ichise, R., Le, B. Discovering Missing Links in Large-Scale Linked Data. In: *Intelligent Information and Database Systems*. Ed. by A. Selamat, N. Nguyen, H. Haron. LNCS. Springer Berlin Heidelberg, 2013.
14. Hitzler, P., Janowicz, K. Linked Data, Big Data, and the 4th Paradigm. In: Semantic Web 4(3) (2013).
15. Jentzsch, A., Cyganiak, R., Bizer, C. State of the LOD Cloud. Version 0.3. 19th Sept. 2011. http://lod-cloud.net/state/ (visited on 2014-08-06).
16. Lange, C. Publishing 5-star Open Data. Tutorial at the Web Intelligence Summer School "Web of Data". 25th Aug. 2014. http://clange.github.io/5stardata-tutorial/.
17. Lehmann, J. et al. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. In: Semantic Web Journal (2014).
18. Llewellyn, A. NASA Tournament Lab's Big Data Challenge. https://open.nasa.gov/blog/2012/10/03/nasa-tournament-labs-big-data-challenge/. Oct. 2012.
19. Lukoianova, T., Rubin, V. L. Veracity roadmap: Is big data objective, truthful and credible? In: Advances in Classification Research Online 24(1) (2014), pp. 4–15.
20. Ngomo, A.-C. N., Auer, S. LIMES: A Time-efficient Approach for Large-scale Link Discovery on the Web of Data. In: *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*. 2011.
21. Polleres, A. et al. RDFS and OWL Reasoning for Linked Data. In: *Reasoning Web. Semantic Technologies for Intelligent Data Access*. LNCS. Springer Berlin Heidelberg, 2013.
22. Sänger, J. et al. Trust and Big Data: A Roadmap for Research. In: *Database and Expert Systems Applications (DEXA)*. Sept. 2014, pp. 278–282.
23. Sauermann, L., Cyganiak, R. Cool URIs for the Semantic Web. Interest Group Note. W3C, Dec. 2008. http://www.w3.org/TR/cooluris/.
24. Schroeck, M. et al. Analytics: The real-world use of big data. How innovative enterprises extract value from uncertain data. Tech. rep. Saïd Business School, University of Oxford: IBM Institute for Business Value.
25. Schultz, A. et al. LDIF: Linked Data Integration Framework. In: *ISWC Posters & Demos*. 2011.
26. Schwarte, A. et al. FedX: Optimization Techniques for Federated Query Processing on Linked Data. In: *The Semantic Web – ISWC 2011, Part I*. 2011.
27. Sequeda, J., Miranker, D. P. Ultrawrap: SPARQL Execution on Relational Data. In: Web Semantics: Science, Services and Agents on the World Wide Web (2013).
28. Unbehauen, J., Stadler, C., Auer, S. Accessing Relational Data on the Web with SparqlMap. In: *JIST*. 2012.
29. Vitter, J. S. Random Sampling with a Reservoir. In: ACM Trans. Math. Softw. (3rd Apr. 2006).
30. Volz, J. et al. Discovering and Maintaining Links on the Web of Data. In: *Proceedings of the 8th International Semantic Web Conference*. Berlin, Heidelberg, 2009.
31. Wache, H. et al. Ontology-based integration of information – a survey of existing approaches. In: *IJCAI-01 Workshop: Ontologies and Information*. Ed. by H. Stuckenschmidt. 2001, pp. 108–117.
32. Zaveri, A. et al. Quality Assessment for Linked Data: A Survey. In: Semantic Web Journal (2015).
33. Zeng, K. et al. A distributed graph engine for web scale RDF data. In: *Proceedings of the 39th international conference on Very Large Data Bases*. PVLDB'13. VLDB Endowment, 2013, pp. 265–276.

Authorized licensed use limited to: Pontificia Universidade Catolica de Campinas (PUC Campinas). Downloaded on April 22,2023 at 03:51:00 UTC from IEEE Xplore. Restrictions apply.