Semantic Web 1 (2015) 1–5
IOS Press

A Comparative Survey of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO

Michael Färber *,**, Basil Ell, Carsten Menne, and Achim Rettinger
Karlsruhe Institute of Technology (KIT), Institute AIFB,
76131 Karlsruhe, Germany

Abstract. In recent years, several noteworthy large, crossdomain and openly available knowledge graphs (KGs) have been created. These include DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. Although extensively in use, these KGs have not been subject to an in-depth comparison so far. In this survey, we first define aspects according to which KGs can be analyzed. Next, we analyze and compare the above mentioned KGs along those aspects and finally propose a method for finding the most suitable KG for a given setting.

Keywords: Knowledge Graph, Comparison, DBpedia, Freebase, OpenCyc, Wikidata, YAGO

1. Introduction

The idea of the Semantic Web is that of publishing and querying knowledge on the Web in a semantically structured way. According to Guns [27], the term "Semantic Web" already was being used in fields such as educational psychology, before it became prominent in computer science. Freedman and Reynolds [25], for instance, describe "semantic webbing" as organizing information and relationships in a visual display. Berners-Lee presented his idea of using typed links as vehicle of semantics for the first time at the World Wide Web Fall 1994 Conference under the heading "Semantics," and under the heading "Semantic Web" in 1995 [27].

The idea of a Semantic Web was introduced to a wider audience by Berners-Lee in 2001 [11]. According to his vision, the traditional Web as a Web of Documents should be extended to a Web of Data where not only documents and links between documents, but any entity (e.g., a person or organization) and any relation between entities (e.g., *isSpouseOf*) can be represented on the Web.

When it comes to realizing the idea of the Semantic Web, knowledge graphs (KGs) are currently seen as one of the most essential components. We define a knowledge graph as a knowledge base (KB) (defined as the combination of an ontology and instances of the classes in the ontology [59, p. 739]) consisting to a large amount of facts about entities. Besides domain-specific KGs, often general, i.e. encyclopedic/crossdomain knowledge is covered in openly available KGs as DBpedia exemplifies. This makes KGs widely applicable: not only a small set of users – as in the case of expert systems – benefit from using the stored structured knowledge (e.g., via using specific search interfaces of expert systems), but any person on the street having access to the Web can benefit, e.g., by using Web search functionalities where semantic queries against a KG extend traditional information retrieval queries on documents.

In this survey, we focus on those KGs (i) that are freely accessible and freely usable, (ii) that incorporate the Semantic Web standards to some extent such

*Corresponding author. E-mail: michael.faerber@kit.edu.
**This work was carried out with the support of the German Federal Ministry of Education and Research (BMBF) within the Software Campus project *SUITE* (Grant 01IS12051).

modeling with RDF[1] and querying with SPARQL,[2] and (iii) that do not cover special domains such as the biomedical domain, but covers instead general knowledge (often also called crossdomain or encyclopedic knowledge).

Thus, out of scope are KGs which are not openly available such as the Google Knowledge Graph,[3] the Google Knowledge Vault [21], and the Facebook Graph[4] as well as KGs which are not based on Semantic Web standards at all or are only accessible via an API (see WolframAlpha[5]). Also excluded are unstructured or weakly structured knowledge collections.

For selecting the KGs for analysis, we regarded all datasets which were registered at the online dataset catalog `http://datahub.io`[6] and which were tagged as "crossdomain". Besides that, we took further datasets into consideration which fulfilled the above mentioned requirements (e.g., Wikidata). In total, we nominated DBpedia, Freebase, Cyc, Wikidata, and YAGO as KGs for our comparison.

In this paper, we give a systematic overview of these KGs in their current versions, and discuss how the facts of these KGs are modeled, stored, and queried. Note that the focus of this survey is not the life cycle of KGs on the Web or in enterprises. We can refer in this respect to [8].

Besides juxtaposing the characteristics of the KGs we provide a recipe for users who are interested in using one of the mentioned KGs in a research or industrial setting, but who are inexperienced in which KG to choose for their concrete settings.

The main contributions of this survey are:

1. We define 35 aspects (characteristics) according to which KGs can be analyzed.
2. We analyze DBpedia, Freebase, Cyc, Wikidata, and YAGO along these aspects.
3. We propose a checklist which enables users to find the most suitable KG for their needs.

[1] See `http://www.w3.org/RDF/` (accessed June 16, 2015).
[2] See `http://www.w3.org/TR/rdf-sparql-query/` (accessed June 16, 2015).
[3] See `http://www.google.com/insidesearch/features/search/knowledge.html`
[4] See `https://developers.facebook.com/docs/graph-api`
[5] See `http://products.wolframalpha.com/api/`
[6] This catalog is also used for registering Linked Open Data datasets.

The organization of this survey is as follows:

- In Section 2 we describe the genesis of semantic data models and provide a definition for both semantic data models and graph models, since KGs are realizations of both models.
- In Section 3 we describe aspects by which knowledge graphs can be analyzed.
- In Section 4 we describe the knowledge graphs we analyze.
- In Section 5 we analyze the knowledge graphs along the aspects listed in Section 3.
- In Section 6 we present a guideline to assess the knowledge graphs according to the user's setting.
- In Section 7 we outline current limitations of KGs
- In Section 8 we glance over the possible future of the KGs and of the Semantic Web
- In Section 9 we conclude the survey.

2. Semantic Data Models and Graph Models

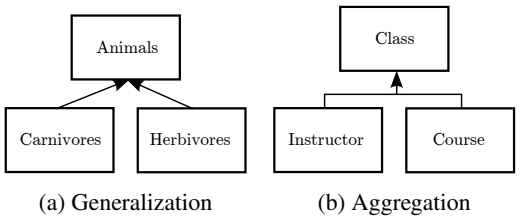
Two data model types are especially relevant with respect to KBs and, hence, to KGs: *Semantic data models* and *graph data models*. In this section, we first describe the genesis of semantic data models and show how both semantic data models and graph data models have been defined. For an in-depth introduction into semantic data models and graph data models, the interested reader is referred to [47] and [5], respectively.

2.1. Genesis of Semantic Data Models

The evolution from database (DB) design toward KB design is coupled with increasing abstraction layers. In the early stages of DB design, models for representing data were modeled conceptually close to the physical layer of data storage. After the basic physical models, hierarchical models [63] became prominent. They were superseded conceptually by network models [62] and later by relational models [17]. Contrary to the hierarchical and network models, the relational model was not located on the primitive record level anymore, but between the physical and logical level, although still affiliated to the record level.

First essential *landmarks* of semantic data modeling arose in the mid-seventies when databases increasingly supported the user's view on the data. Then, new paradigms enriched semantic data modeling over time. The following concepts are notable in this respect:

1. The idea of *data independence* (see [16]) states that data is not modeled in the way as required by the storage architecture, but according to the user's application (i.e., having entities and relationships among them). This application-oriented paradigm was accompanied by the developments and the emergence of new programming languages such as C (developed between 1969 and 1973) and Smalltalk, which were more abstract than traditional languages and not data storage-oriented such as COBOL.
2. The idea of *semantic "injection"/enrichment* (see for instance [54] and [56]) states that: Although semantics was encoded into data models on a low level (i.e. to single data items) up to the mid-seventies, new approaches considered semantic relations between data items and how to model interrelational dependencies. By implementing rules which are based on these interrelational dependencies, consistency checks were made possible. We can mention the following notable modeling approaches as steps in the evolution of semantic data modeling:
   (a) Schmid [54] introduced the idea to model basic semantic properties that entities of a certain class (e.g., person) may have as well as relationships that entities of certain classes may have (e.g., the has-spouse-relationship).
   (b) Smith [56] introduced generalization and aggregation as new forms of abstraction: a) *Generalization* is used to express similarities (see Figure 1a) and is modeled between classes: One class (e.g., carnivores) is subclass of another class (e.g., animals) and shares properties with the superclass. b) An *aggregation* is the composition of an object from a set of objects. The aggregation class (see *Class* in Figure 1b) stands as a whole unit in place of its components (in Figure 1b *Instructor* and *Course*).
   (c) Brodie [14] introduced *classification* and *association* relationships as further modeling approaches: a) Classification means to assign a class to an entity (e.g., Markus is a person). b) An *association* is a relationship between classes and describes the connections between classes in terms of the shared semantics and structure. Any aggregation or composition are associations.



(a) Generalization (b) Aggregation

Fig. 1. Examples for generalization and aggregation.

3. Later, in the 1980s, object-oriented models [36] appeared: Data was considered as collections of objects of specific classes. In parallel to the uprise of object-oriented models, graph models appeared. With this model users were able to represent the inherent graph structure of data.
4. Afterwards, other models such as semi-structured models [15] and the XML model [13] were proposed.
5. The Resource Description Framework (RDF) was originally published by the W3C as recommendation in 1999 [40] and in a new version in 2004 [37]. In 2014, RDF 1.1 [20] was published. RDF builds the basis of the semantic graph model as we consider it in this survey.

2.2. Definition of the Semantic Data Model

According to Hammer [29], semantic data models are characterized as data models adhering to the following principles:

1. A database is a collection of *entities* that correspond to actual objects in the application environment.
2. Entities in the database are organized into *classes*.
3. Classes may be interconnected.
4. Entities and classes are characterized by relations (called *attributes* by Hammer) and relations may interconnect entities.
5. Relations can be derived from other relations via entailment.

Well-known examples of the semantic data model are the entity-relationship model [16] and RDF.[7] Further examples are the IFO model [1] and SDM [30].

[7] See `http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/` (accessed July 22, 2015).

2.3. Definition of the Graph Data Model

When deciding for a data model, the choice typically depends on (i) the modeling domain, (ii) the end users, and (iii) the hardware and software constraints. Codd [18] distinguishes three components each data model possesses: 1) The set of data structure types, 2) the set of operators or inference rules, and 3) the set of integrity rules. Angles and Gutierrez [5] define the *graph data model* along this categorization:

1. *Data structure types*: In the graph data model, instance data and/or schema information is represented as/by *graphs*, or by data structures which generalize the notion of graph.
2. *Operators and inference rules*: Data is manipulated via *graph transformation operators* (see [28]).
3. *Integrity rules*: Rules can be constructed which ensure data consistency.

In graph data models, the semantics comes into play as follows: (i) Graph nodes are interpreted as entities or values; (ii) typed relations between nodes are interpreted as facts about the involved entities; and (iii) a schema is introduced by assigning types to instances and introducing relations among classes. The information focus supported by the graph data model therefore encompasses the schema, the instances and the relations.

Established graph models are *GOOD* [28], *GMOD* [4], *G-Log* [46], *Gram* [3], and RDF. They all use a labeled directed graph both for modeling the schema as well as the instance level.

KGs, as we consider it in this survey, are realizations of both semantic data models and graph models, since the KGs are characterized by having a set of entities, a set of classes, a set of relations between entities, and a set of relations between entities and classes.

3. Criteria for Comparison

Several works compared semantic data models:

- Brodie [14] categorized semantic data models into: (i) classical models, (ii) mathematical models, (iii) irreducible data models, (iv) static semantic hierarchy models, and (v) dynamic semantic hierarchy models.
- Tsichritzis and Lochovsky [64] categorized semantic data models into: (i) traditional models, (ii) entity-relationship models, (iii) binary models, (iv) semantic network models, and (v) infological data models.
- Hull and King [31] compared semantic data models according to the features they provide, such as aggregation, grouping, printable, object-valued, and multi-valued.
- Kerschberg et al. [35] analyzed data models according to mathematical foundations, terminology, and semantic levels of abstraction, and distinguished between graph theoretic and set theoretic models.

These approaches do neither consider current KGs with their data models nor ontologies. Also, the used criteria for comparing semantic data models are very abstract, since a wide range of data models are compared against each other. Besides these works, there are approaches which analyze (and sometimes assess) explicitly ontologies, but not KGs. Some of them are mentioned in the following (see also [12] for an overview of ontology evaluation):

- Tartir et al. [61] introduced the approach OntoQA by which ontology schemas and their populations (i.e., KBs) can be analyzed through a set of metrics so that key characteristics of an ontology schema can be highlighted. The analysis focused on numerically expressible characteristics of ontologies and results regarding the ontologies SWETO, TAP, and GlycO were presented.
- Lozano-Tello et al. [42] proposed ONTOMETRIC, which allows the users to measure the suitability of existing ontologies, regarding the requirements of their systems. OntoMETRIC presents a generic methodology and does not analyze specific ontologies.
- Vrandecic et al. [66] differentiated between structural and ontological metrics and provided principle means for the definition of metrics that take the semantic of the ontology appropriately into account.
- Poveda-Villalon et al. [48] present a tool called OOPS! by which an RDF document describing an ontology can be analyzed. Potential pitfalls that could lead to modeling errors are then presented to the user.

Since these approaches focus only on ontologies, we cannot compare the used datasets. Also the criteria for comparison are different to ours, since we do not only focus on the schema. To the best of our knowledge, a

systematic comparison of openly-available knowledge graphs has not been carried out so far. Therefore, we systematically analyze and compare knowledge graphs according to aspects of the following categories:

- *General information*: What general properties does the KG have?
- *Format and representation*: How are facts represented, stored, and queried?
- *Genesis and usage*: How was the KG created and how is it used?
- *Entities*: How are entities represented and described in the KG?
- *Relations*: How are relations represented and described in the KG?
- *Schema*: What are the features of the schema of the KG?
- *Particularities*: What particularities (special features) does the underlying data model of the KG have?

In the following, we list the criteria we use for comparing the different KGs, grouped by the categories mentioned.

3.1. General Information

We use the following aspects to collect general information about the KGs:

1. **Homepage**: The URL where the KG can be accessed.
2. **Current version**: The version of the knowledge base we consider in this survey.
3. **Languages**: What languages (e.g., English) are used in the KG on schema and instance level?
4. **Covered domains**: Which domains are covered by the KG? Are there any fields where the KB is filled only rudimentary?
5. **License**: Under which license is the content of the KG provided?

3.2. Format and Representation

For comparing the different approaches for representing, storing and querying knowledge, we use the following aspects:

1. **Fact representation**: Facts can be represented as triples, quadrupels, or similar.
2. **Dataset formats**: The data storage format (e.g., JSON) in which data is provided.
3. **Dynamicity**: Is the KG updated continuously (dynamic KG) or are only fixed versions of the KG offered (static KG)?
4. **HTTP lookup**: Is machine-readable information about resources available via live HTTP lookup (i.e., querying on demand in order to follow the Linked Data principles [9], so that no export functionality or file download is needed)?
5. **RDF export**: Is data available as RDF export, either via files or via SPARQL endpoint?
6. **Software for data storage**: Which software is used for storing and querying the KG?
7. **Query language (online)**: Each KG may provide one or several query languages in which queries against the KG are formulated.
8. **Size of schema and instance graph**: How many classes and relations are in the KG, how many facts, and how many unique instances?

3.3. Genesis and Usage

Where the stored facts in the KGs come from and where they are applied, is addressed by these aspects:

1. **Provenance of facts**: Is the KG content derived from unstructured or semi-structured data by information extraction techniques or is it gathered manually by users and/or bots?
2. **Quality ensurance of facts**: Are there any restrictions or constraints regarding the quality of stored knowledge? If the correctness of facts is ensured, how and with what precision is this performed?
3. **Software projects**: Which software projects make use of the KG?
4. **Influence on other LOD datasets**: Which other KG-building initiatives take the KG as a starting point?

3.4. Entities

The following aspects address the characteristics of the entities in the KGs:

1. **Entity reference**: What kind of IDs are used to refer to entities?
2. **LOD registration**: Is the dataset registered at `http://datahub.io` as part of the Linked Open Data (LOD) cloud?[8]

[8]The Linked Open Data (LOD) cloud is a collection of datasets published on the Web following the Linked Data principles [9].

3. **LOD linkage**: Are entities linked to entities of other KGs in the LOD Cloud?
4. **Entity relevance**: Is the ordering or ranking of entities according to some function such as a relevance function supported?
5. **Description of entities**: Are entities human-readably described within the KG, e.g., via textual descriptions? What format is used for that?

3.5. Relations

The following aspects address the characteristics of the relations in the KGs:

1. **Relation reference**: What kind of IDs are used to refer to relations?
2. **Relation relevance**: Is the ordering or ranking of relations according to relevance supported? In this way, relations can be declared as more important, for instance since they are more relevant to most users than other relations.
3. **Description of relations**: Are relations human-readably described within the KG, e.g., via textual descriptions? What format is used for that?

3.6. Schema

The characteristics of the schema of the different KGs can be addressed by the following aspects:

1. **Schema restrictions**: Is a fixed schema used or can the schema be extended by users?
2. **Schema constraints**: Are there any schema constraints which need to be observed? May the KG contain data that is inconsistent regarding the schema?
   For example, if – according to the (logical) schema constraints – an entity may only occur once as subject of a certain relation such as `has-spouse`, but within the KG occurs several times as subject with different objects (which are explicitly defined as different entities), then the KG contains data which violates the schema.
3. **Hierarchy and network of relations**: Does the KG contain relations among relations, e.g., a taxonomy of relations (sub-relation, super-relation) or other types of relations (e.g., inverse relation)?
4. **External vocabulary**: Is external vocabulary (classes or relations that belong to existing schemas, e.g., schemas from datasets in the LOD cloud) reused in the KG?
5. **Description of concepts**: Are concepts described within the KG, e.g., via textual descriptions? What format is used for that?
6. **Forms of abstraction**: Is classification, generalization, aggregation, or association supported?
7. **Data types**: Which data types are used in the KG?

3.7. Particularities

Here, particularities about the analyzed KGs are identified which are not covered by the other aspects. Aspects anticipated before the analysis were:

1. **Temporal aspects**: Facts that may change over time (e.g., a country's president) may be annotated according to the time when the fact was valid (e.g., time interval). Furthermore, other time-related information about a fact may be stored, such as the point in time when the fact was added to the KG or when the fact was updated.
2. **Source of facts**: Is it stored where the knowledge in the KG was retrieved from (e.g., the document it was extracted from)?
3. **Reification**: Is it possible to represent statements about statements? Reification here means to have a means for referring to a statement via an identifier thus enabling to formulate statements about statements.

4. Selection of KGs

We consider the following knowledge graphs for our comparative evaluation:

– **DBpedia**: DBpedia[9] is the most popular and prominent KG in the LOD cloud [7]. The project was initiated by researchers from the Free University of Berlin and the University of Leipzig, in collaboration with OpenLink Software. Since the first public release in 2007, DBpedia is updated roughly once a year.[10] DBpedia is cre-

The LOD cloud project originated from the W3C Linking Open Data project (see `http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData`) and currently consists of almost 1000 datasets; see `http://lod-cloud.net/`.

[9] See `http://dbpedia.org`

[10] There is also DBpedia live which started in 2009 and which is updated when Wikipedia is updated. See `http://live.dbpedia.org/`.

ated from automatically-extracted structured information contained in the Wikipedia, such as from infobox tables, categorization information, geo-coordinates, and external links. Due to its role as the hub of LOD, DBpedia contains many links to other datasets in the LOD cloud such as Freebase, OpenCyc, UMBEL,[11] GeoNames,[12] Musicbrainz,[13] CIA World Factbook,[14] DBLP,[15] Project Gutenberg,[16] DBtune Jamendo,[17] Eurostat,[18] Uniprot,[19] and Bio2RDF.[20] DBpedia is used extensively in the Semantic Web research community, but is also relevant in commercial settings: companies use it to organize their content, such as the BBC [38] and the New York Times [52].

– **Freebase**: Freebase[21] is a KG announced by Metaweb Technologies, Inc. in 2007 and was acquired by Google Inc. on July 16, 2010. In contrast to DBpedia, Freebase had provided an interface that allowed end-users to contribute to the KG by editing structured data. Besides user-contributed data, Freebase integrated data from Wikipedia, NNDB,[22] FMD,[23] and MusicBrainz.[24] Freebase uses a proprietary graph model for storing also complex statements. On December 16, 2014, the Freebase team announced that Freebase will shutdown its services on June 30, 2015. Wikimedia Deutschland and Google plan to integrate Freebase data into Wikidata in the near future – a tool for that will be developed until August 2015 – and to close the Freebase website earliest three months later.[25]
– **OpenCyc**: The Cyc[26] project started in 1984 as part of Microelectronics and Computer Technology Corporation. The aim of Cyc is to store (in a machine-processable way) millions of common sense facts such as "Every tree is a plant." While the focus of Cyc in the first decades was on inferencing and reasoning, more recent work puts a focus on human-interaction such as building question answering systems based on Cyc. Since Cyc is proprietary, a smaller version of the KG called OpenCyc[27] was released under the open source Apache license. In July 2006, ResearchCyc[28] was published for the research community, containing more facts than OpenCyc.
– **Wikidata**: Wikidata[29] is a project of Wikimedia Deutschland which started on October 30, 2012. The aim of the project is to provide data which can be used by any Wikipedia project, including Wikipedia.
  Wikidata does not only store facts, but also the corresponding sources, so that the validity of facts can be checked. Labels, aliases, and descriptions of entities in Wikidata are provided in more than 350 languages. Wikidata is a community effort, i.e., users collaboratively add and edit information. Also, the schema is maintained and extended based on community agreements. In the near future, Wikidata will grow due to the integration of Freebase data.
– **YAGO**: YAGO – Yet Another Great Ontology – has been developed at the Max Planck Institute for Computer Science in Saarbrücken since 2007. YAGO comprises information extracted from the Wikipedia (e.g., categories, redirects, infoboxes), WordNet[23] (e.g., synsets, hyponymy), and GeoNames.[30] As of March 24, 2015, YAGO3 is available.[31]

5. Comparison

In the Tables 1 – 7 we summarize our comparison of the knowledge graphs listed in Section 4 using the

[11] See http://umbel.org/
[12] See http://www.geonames.org/
[13] See http://musicbrainz.org/
[14] See https://www.cia.gov/library/publications/the-world-factbook/
[15] See http://www.dblp.org
[16] See https://www.gutenberg.org/
[17] See http://dbtune.org/jamendo/
[18] See http://eurostat.linked-statistics.org/
[19] See http://www.uniprot.org/
[20] See http://bio2rdf.org/
[21] See http://freebase.com/
[22] See http://www.nndb.com
[23] See http://www.fashionmodeldirectory.com/
[24] See http://musicbrainz.org/
[25] See https://plus.google.com/u/0/109936836907132434202/posts/bu3z2wVqcQc and https://www.wikidata.org/wiki/Wikidata:WikiProject_Freebase.
[26] See http://www.cyc.com/
[27] See http://www.opencyc.org/
[28] See http://research.cyc.com/
[29] See http://wikidata.org/.
[30] See www.geonames.org/
[31] See http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/

aspects given in Section 3. An online version is also available at http://kg-wiki.org. In the following subsections, we provide a detailed analysis in terms of the different aspects of the introduced evaluation categories.

5.1. Comparison of General KB Information

The following findings are notable with regard to general information of the considered KGs (cf. Table 1; information about the homepage and the considered KG version is not discussed any further):

- *Language support:* Most KGs either only support the English language (such as OpenCyc), or other languages than English are added on top. For Freebase and YAGO, entity and property labels for additional languages are provided.
  Remarkable in this context is Wikidata in terms of number of languages supported and in terms of its language-agnostic KG model (cf. the identifiers for entities and relations which intentionally consist of a character and a number).
- *Covered domains:* Since we restricted our KG choice to crossdomain KGs, all considered KGs contain general knowledge, for instance general information about instances of persons such as Barack Obama. Besides specific domains such as the biomedical domain, also common sense knowledge (class-relationships such as "A human has two legs" or "A child is a human") and linguistic knowledge (relationships between linguistic concepts such as "to compose is a synonym of to write") were excluded from this survey.
  Although the scope of the considered KGs is broad and unrestricted in nature, we can make statements about the "relative filling degrees" (in terms of number of entities or number of statements) with respect to located parts of the KGs and, hence, about the maturity of the considered KGs:
  Firstly, Wikidata still is in a start-up phase in the sense that not all subdomains (indicated by the classes) are covered in depth. Wikidata is especially well populated in fields such as "Person" and biological entities, but provides only rudimentary information about entities in fields such as society. All other KGs can be classified as mature, since they do not only exist for a rather long time, but are well positioned in all general domains.
  Secondly, OpenCyc can be seen as mature, but consists of much schema information and is – in terms of the entities, and, hence in the sense of a KG – rather a collection of entities belonging to different classes. Hence, OpenCyc is predestinated for reasoning, but not so much for entity retrieval purposes.
- *License:* Data of all considered KGs except Wikidata is licensed under the Creative Commons Attribution 3.0 license[32] which means that it is allowed to use the data for private and commercial settings and to modify the data by the user.
  In case of Wikidata, all structured data of the main name space and the property name space of Wikidata is licensed under Creative Commons CC0,[33] while text of all other namespaces of Wikidata is available under the Creative Commons Attribution/Share-Alike License.[34] The Creative Commons CC0 licence enables to waive as many rights as legally possible and is especially used for databases.
  In summary we can state that all considered KG can be used without expenses, but in return appropriate credit has to be given and the same license has to be used for further usage.
  Interesting in the context of KG licenses is the study of Jain et al. [32] who studied the applicability of well-known Linked Data datasets for commercial applications. The conclusion the authors drew is that not the technical issues of deployment and use of Linked Data datasets is the crucial point, but legal aspects. Often, the license under which a Linked Data dataset can be reused is not specified by the data providers.

5.2. Comparison of Format and Representation

- *Fact representation:* The KGs DBpedia and OpenCyc store facts as single triples and do not regard additional meta-information about facts such as the confidence of the triples being correct or temporal information related to the facts (e.g., the validity time).

[32] See https://creativecommons.org/licenses/by/3.0/.
[33] See https://creativecommons.org/publicdomain/zero/1.0/.
[34] See https://creativecommons.org/licenses/by-sa/3.0/ and https://www.wikidata.org/wiki/Wikidata:Database_download/en#License.

Table 1: Comparison of the KGs regarding their general information.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Homepage	http://dbpedia.org	http://freebase.com	http://opencyc.org	http://wikidata.org	http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/
Current version	DBpedia 2015-04	continuously updated until Mar 31, 2015[a]	OpenCyc 4.0	Cont. updated since Oct 2012	YAGO3
Languages	"Main" DBpedia is English (properties etc.), but linked localized versions are available in 125 languages (localized are textual descriptions such as rdfs:label, rdfs:comment, dbpedia-owl:abstract. There are also links to local versions of Wikipedia)	human readable IDs are in English, but every entity and property has an i18n in many languages	English	Almost every language (by community), even dialects	All entity names are from English Wikipedia, some rdfs:label values have different languages
Covered domains	General knowledge	General knowledge, very broad, sometimes deep	Common sense	General knowledge	General knowledge
License (content)	Creative Commons Attribution-ShareAlike 3.0, GNU Free Documentation License	Creative Commons Attribution Only	Creative Commons Attribution 3.0	Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication	Creative Commons Attribution 3.0

[a] Google announced to close Freebase on June 30, 2015. However, currently (July 30, 2015) it is still available.

Table 2: Comparison of the KGs regarding format and representation.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Fact representation	Triple	Triple with confidence values	Triple	Every entity has multiple statements. Each statement has one or more references and one claim. The claim consists of a property and a value, accompanied by qualifiers	"SPOTL tuple" (SPO triple with time and location)
Dataset formats	RDF (nt, nq, ttl)	RDF (nt)	RDF (owl), proprietary file format	JSON, XML, SQL, and RDF[a]	RDF (ttl), TSV
Dynamicity	Static	Continuously updated	Static	Continuously updated	Static
HTTP lookup	lookup on demand	lookup on demand	lookup on demand	lookup on demand (but incomplete)	lookup on demand
RDF export	as files and via SPARQL	as files	as OWL file	as files (by third party) and SPARQL (by third party)	as files and via SPARQL
Software for data storage	Virtuoso Universal Server	Graphd	Cyc Knowledge Server	Wikibase	rudimentary query interface, rudimentary browser and demos
Query language (online)	SPARQL (`http://dbpedia.org/sparql`)	MQL (Metaweb Query Language; `https://freebase.com/query`)	CycL Language	Wikibase-API, SPARQL (third party)	SPARQL (`http://lod2.openlinksw.com/sparql`)
Size of schema and instance graph	4.58 Mio entities, 685 classes, 1,079 object properties, 1,600 data type properties, 116 specialized data type properties	1.9 Bio triples	239k terms, 2 mio triples, 47k links to DBpedia	63.2 Mio statements	>10 Mio entities, >120 Mio facts

[a] https://tools.wmflabs.org/wikidata-exports/rdf/

In contrast to these KGs stand Freebase, YAGO, and Wikidata: For each triple, Freebase also stores a confidence value. The authors of YAGO use the so-called SPOTL(X) tuples for representing spatio-temporally enhanced facts (with the elements subject (S), predicate (P), object (O), time (T), and location (L)). For Wikidata, a model is used where each statement consists of a claim that something is the case and a list of references providing evidence for that claim.

- *Dataset formats:* Regarding all considered KGs, data is available in RDF format: Data from DBpedia and Freebase is available in the form of RDF files,[35] data from OpenCyc is available as OWL files,[36] and YAGO is available as both TSV files and RDF files.[37]
  Wikidata has a special position here: It is developed on the basis of Wikibase,[38] a proprietary data model for Wikidata. This data model is per se not based on the RDF format. However, unofficial RDF export files of Wikidata[39] (and some SPARQL endpoints[40]) are provided.
- *Dynamicity:* Many KGs are static in the sense that they are not continuously updated. One reason for that is that some KGs such as DBpedia are created by computationally-expensive information extraction processes. Therefore, DBpedia is static; however, DBpedia live – a derived version of DBpedia – is continuously updated. For that, Wikipedia provides a OAI-PMH update stream, by means of which 84 articles are analyzed per minute.[41]
  Dynamic KGs are Freebase and Wikidata since data is maintained by a user community. In case of these KGs, even the schema is extended by the users.
- *HTTP Lookup:* Regarding all considered KGs, data is made available via HTTP lookups on demand: Given a resource (of a KG which is part of the LOD cloud) identified by a HTTP URI, data about this resource can be obtained by dereferencing this URI.[42] Typically, the returned information is made available using W3C standards such as RDF. The idea of dereferencing is a crucial point of the Semantic Web vision: In this way, agents can traverse the LOD graph (i.e. following links within and across single LOD datasets) and gather the information which they need and which is available in the LOD cloud.
  HTTP lookups on demand are possible for all KGs considered in this survey – thus allowing for data exports.
- *RDF export:* Besides the HTTP lookup availability, data from the KGs is also made available as files. The idea is here that the KG data can also be downloaded and processed otherwise instead of retrieving data via HTTP lookups on demand; this includes parsing the files directly or importing the data into an appropriate database such as a triple store. In this way, queries can be set up and the hardware load is on the client side.
- *Data storage software:* Data is stored using different systems: While DBpedia uses Virtuoso Universal Server[43] and its available RDF dumps can be loaded into any triple store (such as Virtuoso or 4store[44]), all other considered KGs (Freebase, OpenCyc, Wikidata, and YAGO3) are – due to their different data models used internally – based on proprietary software systems. However, the provided RDF dumps of these KGs can be loaded into any triple store.
- *Query language (online):* Although all considered KGs both support RDF as data format and are available online for HTTP lookups, not all online versions of the KGs are offered with a SPARQL endpoint: Only DBpedia and YAGO are queriable in this way.[45] For Wikidata, several un-

[35] See `http://wiki.dbpedia.org/Downloads2014`, `https://developers.google.com/freebase/data`, `http://tools.wmflabs.org/wikidata-exports/rdf/`.

[36] See `http://sw.opencyc.org/`.

[37] See `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/downloads/`.

[38] `https://www.mediawiki.org/wiki/Wikibase/DataModel`

[39] See `http://tools.wmflabs.org/wikidata-exports/rdf/` and [22].

[40] See `https://www.wikidata.org/wiki/Wikidata:Data_access/en`.

[41] See `http://wiki.dbpedia.org/online-access/DBpediaLive`.

[42] See `http://tools.ietf.org/html/rfc3986#section-1.2.2` for more information about dereferencing URIs.

[43] See `http://virtuoso.openlinksw.com/`

[44] See `http://4store.org/`

[45] See `http://dbpedia.org/sparql` and `http://lod2.openlinksw.com/sparql`.

official SPARQL endpoints are available.[46] If the client wants to query the KGs Freebase and OpenCyc by means of SPARQL – and therefore rely on a W3C Recommendation instead of a proprietary query language –, he needs to load the provided RDF data into a triple store. This procedure is suggested by the authors of many KGs in case of permanent, extensive querying, since in this way the user needs to provide the hardware resources by himself. Furthermore, it should be noted that in case of OpenCyc the language CycL[47] was developed to enable expressive reasoning.

- *Size of schema and instance graph:* In our analysis, we only took KGs into consideration which are already widely used and which are representative for open, semantically-structured datasets on the Web. All considered KG are therefore large and a comparison of size per se is not reasonable, since the KGs often cover different domains to a different extent or emphasize different levels of knowledge. Note that the numbers given in the table with respect to the KG size are not directly comparable. A complex fact may be represented and counted as one statement in one KG but represented and counted as multiple statements in another KG.

  As already outlined above under *coverage*, we can state that all considered KGs fulfill the requirements of being a KG. Outstanding are the KGs Wikidata and OpenCyc. While Wikidata is not mature in all areas, but very focused on instances, the primary focus of OpenCyc is schema information; however, it contains many instances and is therefore numbered among the KGs.

5.3. Comparison of Genesis and Usage

- *Provenance of facts:* For covering knowledge about general domain entities – as done primarily by DBpedia, Freebase, Wikidata, and YAGO –, Wikipedia content is exploited to some degree with the help of information extraction tools. For creating a more formal-logical representation of knowledge, experts need to be consulted as the case of Cyc/OpenCyc demonstrates. In the case of Freebase, provenance data stored for facts are for example the IDs of the users that added the facts. For Wikidata, to each statement (consisting, e.g., of a property and a value, such as (country, Germany)) references can be attached which reveal the source – and therefore indirectly the trustfulness – of the statement.
- *Quality ensurance:* The quality ensurance of facts can be aligned with the two ways of fact provisioning in general (see also the aspect *Provenance of facts*): (i) Knowledge for the KG is extracted automatically from a database such as Wikipedia. In that case, no quality ensurance check is implemented, but a posteriori evaluations confirmed a sufficient high average accuracy across the KG YAGO [43]. (ii) Knowledge is gathered by user contributions. In those cases (see Freebase, Cyc, and Wikidata) no fact consistency checks are applied, but the correctness is based on the trustfulness of the contributors.

  In general it is not possible to prioritize one of these two ways a priori. Using solely approach (i) is only duable if the information is already available in semi-structured formats (as in case of Wikipedia-DBpedia), so that the proportion of incorrect facts in the KG is kept small.
- *Software projects:* All considered KGs are exploited in many ways in research projects of universities and in industry, so that we only present projects which are commonly known in the community. Notable is in particular the project IBM Watson[48] which uses several of the considered KGs (namely, DBpedia and YAGO so far). For a description of applications of Linked Data in general in the industry, we can refer to the use cases listed by the W3C.[49]
- *Influence on other LOD datasets:* Data of the single KGs has been reused in other data sources of the LOD cloud – especially in datasets which focus more on the integration of multiple datasets instead of building a genuine own knowledge base (see UMBEL[50] and BabelNet[51] as examples).

[46] See `https://www.wikidata.org/wiki/Wikidata:Data_access` for an overview and `http://wdqs-beta.wmflabs.org/` for an example.

[47] See `http://www.cyc.com/documentation/ontologists-handbook/cyc-basics/syntax-cycl/`.

[48] See `http://www.aaai.org/Magazine/Watson/watson.php`

[49] See `http://www.w3.org/2001/sw/sweo/public/UseCases`.

[50] See `http://www.umbel.org/`.

[51] See `http://babelnet.org/`.

Table 3: Comparison of the KGs regarding genesis and usage.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Provenance of facts	Automatically extracted from Wikipedia	Initially from Wikipedia, MusicBrainz etc.; new information is gathered by algorithms, the Freebase team, and the community	Filled by experts	Data is maintained by users and bots	Wikipedia, Wordnet, Geonames gazetter, Wikidata
Quality ensurance of facts	quality depends on Wikipedia content and on extraction algorithms/template mappings	trusted "Freebase experts" keep an eye on changes, scripts look for incorrect data	no	Users should only add verifiable information from sources such as books, scientific publications, or newspaper articles, as in the original Wikipedia, data is controlled by community	no ensurance, evaluation of > 95 % correctness for YAGO2
Software projects	DBpedia Spotlight,[a] Wikipedia Miner,[b] IBM Watson[c]	Google Knowledge Vault, Bing	Terrorism Knowledge Base	some first prototypes	YAGO NAGA, IBM Watson, Broccoli[d]
Influence on other LOD datasets	Freebase, YAGO	Wikidata	UMBEL	YAGO3	SUMO,[e] DBpedia, UMBEL, Freebase

[a] See http://spotlight.dbpedia.org/
[b] See http://wikipedia-miner.cms.waikato.ac.nz/
[c] See http://www.ibm.com/smarterplanet/us/en/ibmwatson/
[d] See broccoli.informatik.uni-freiburg.de
[e] See http://www.adampease.org/OP/

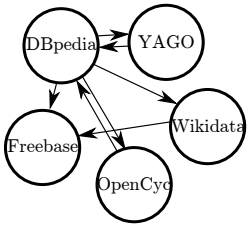


Fig. 2. `owl:sameAs` relations between the considered KGs.

In this context, the relation `owl:sameAs` is worth mentioning: This relation indicates that the two resources which are linked by an `owl:sameAs` relation refer to the same real-world object, even though they might have different URIs in the different datasets. In this way, further information about the resources can be retrieved from linked datasets and no information needs to be copied to other datasets.
The most important sets of `owl:sameAs` relations between the considered KGs are as follows (see also Figure 2): DBpedia and YAGO link to each other; DBpedia links in addition to Freebase and Wikidata; Wikidata links to Freebase; OpenCyc links to DBpedia; DBpedia links to broken Cyc links.[52]

5.4. Comparison of Entities

- *Entity reference:* Since all considered KGs were shaped by the vision of the Semantic Web, entities do not only have unique IDs, but URIs by which they can be referred to. While DBpedia, Freebase, Cyc, and YAGO provide human-readable IDs, Freebase, and Cyc additionally operate with opaque URIs. Wikidata only provides entity IDs which consist of "Q" followed by a number in order to be language-agnostic. The labels for the entities are stored in Wikidata separately.
  As outlined by Berners-Lee in 1998 in "Cool URIs don't change" [10], URIs should be designed with three things in mind: simplicity, stability, and manageability. In the context of KGs where each entity has a URI, well-chosen URIs become even more important. Sauermann and Cyganiak [53] present so-called 303 URIs (which are human readable) and hash URIs (which are not human readable) for the Semantic Web. Both forms have their advantages and disadvantages.
- *LOD registration:* Publishing a dataset according to the Linked Data principles already implies that this dataset is part of the LOD cloud. Besides that, there are Linked Open Data registration portals such as `http://datahub.io` where LOD datasets can be registered and, hence, found quickly.
  All KGs considered in this survey are published in RDF and are part of the LOD cloud. Besides that, until the submission of this survey, all considered KGs except Wikidata were also registered at `http://datahub.io` as part of the LOD cloud.[53]
- *LOD linkage:* Most of the considered KGs link their entities to entities of other datasets in the LOD cloud. Remarkable are hereby DBpedia and Freebase in terms of their high degree of connectivity with other LOD datasets. DBpedia is justifiably called the *hub of the LOD cloud* [41,50].
- *Entity relevance:* In some scenarios it is helpful to rank or order entities based on some importance and/or relevance score (e.g., to find the most well-known football players or politicians). In the past, several approaches were presented which calculate scores for entities. However, currently only Freebase provides relevance scores for entities that were created by using the link counts in Freebase and Wikipedia.[54]
- *Description of entities:* It can be difficult for users to figure out which entity is meant by a given ID or URI – especially if the ID is a mostly numerical value due to the constraint of the knowledge representation being language-independent. In such cases, a textual description of the entities is important. While some KGs offer textual descriptions via special properties (see DBpedia and Freebase) or fields within the data model (as in case of Wikidata), YAGO does not offer any entity description and OpenCyc only for a fraction of the entities.

[52] HTTP requests of URIs with the domain `http://sw.cyc.com` result in a DNS error, but these URIs are dereferencable if the domain is replaced by `http://sw.openyc.org`.

Table 4: Comparison of the KGs regarding entities.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Entity reference	URI (based on Wikipedia page title), e.g., `http://dbpedia.org/resource/Karlsruhe`	MID and sometimes ID (MID is permanent consistent with changes, but not human-readable as IDs) `/m/0qb1z`	unique ID, English ID, e.g., `T6OkHdRS-eUiqO5n8NA1g`, `Karlsruhe`	unique ID, e.g., `Q1040`	URI (based on Wikipedia page title), e.g., `http://yago-knowledge.org/resource/Karlsruhe`
LOD registration	yes	yes	yes	no	yes
LOD linkage	Links to Freebase, OpenCyc, YAGO, UMBEL, GeoNames, etc.	Links to BBC Music, Geospecies	Links to DBpedia	Some entities have links to Freebase and Musicbrainz	Links to DBpedia
Entity relevance	no	Entities have relevance scores (calculated by link counts in Freebase and Wikipedia)	no	no	no
Description of entities	Yes, via property `http://dbpedia.org/ontology/abstract`	yes, via property `/common/topic/description`; often there is an image `/common/topic/image`	Textual description in *comment* field	Description field for every entity, no special property	no

Table 5: Comparison of the KGs regarding relations.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Relation reference	URI (two property prefixes: *ontology* and *property*)	ID and MID (see entity referencing in Table 4)	Phrase (e.g., IsA)	ID	URI
Relation relevance	no	no	not mentioned	yes, e.g., by date	no
Description of relations	All properties have a label, some properties have a link to a description	yes, via `/common/topic/description`; mostly one sentence: e.g. `/film/producer/films_executive_produced`: "Films this person and served as an executive producer on"	no	yes, specific field in every property	properties have a property "hasGloss" with a pattern of usage

5.5. Comparison of Relations

- *Relation reference:* Analogously to entity references, relations in the KGs are represented as IDs or URIs.
- *Relation relevance:* In this respect, Wikidata is in particular noteworthy. Since each statement stored in Wikidata has some meta-information such as a timestamp attached to it, properties of items in Wikidata can be ordered with respect to these meta-information aspects.
- *Description of relations:* Freebase, Wikidata, and YAGO provide a textual description for all relations, so that the meaning of properties can be derived. YAGO stores how the properties are expressed in freetext, while the other KGs provide custom-built textual descriptions. DBpedia provides a label and a link to a definition only for some properties. Cyc and BabelNet do not provide any description at all.

5.6. Comparison of Schema

- *Schema restrictions:* In KGs where the data is derived via automated extraction methods, both the set of classes and the set of relations is fixed. In case of KGs where end-users can contribute (see Wikidata and Freebase) the schema is fixed, but can be extended.
- *Schema constraints:* Constraints regarding the schema become relevant when new facts are added to the KG (see *integrity constraints and data consistency* as conceptual keystone of any graph data model according to [18]). None of the considered KGs use significant constraints when facts are added: In case of DBpedia the type is fixed during the mapping process. No further constraints are given. For YAGO, a type checker and constraint checker is provided. Wikidata has no constraint check tool, but users can report constraint violations. OpenCyc has no constraints since the facts are created manually by experts. No information about constraints was found in case of Freebase.
- *Hierarchy and network of relations:* Only OpenCyc implements a hierarchy of relations [19]. Notable is, however, that DBpedia properties are automatically extracted from Wikipedia, leading to many properties whose meaning is not given,[55] remains unclear[56] or which are semantically overlapping with other properties.[57]
- *External vocabulary:* DBpedia and YAGO use vocabularies from other datasets (DBpedia uses owl, xsd, rdfs, rdf, foaf, dc, skos, umbel[58]; YAGO uses skos, umbel, rdfs und rdf), while Freebase, and Cyc only use their own vocabulary. Wikidata uses also their own vocabulary, but also links sometimes to external vocabulary via "equivalent property" property.
- *Description of classes:* DBpedia, Freebase, Cyc, and Wikidata provide human-readable descriptions of their classes (DBpedia uses `dbpedia-owl:abstract` and `rdfs:comment`; YAGO only uses `rdf:label`, no description; Freebase uses the relation `/common/topic/description`; OpenCyc has a `comment` relation; Wikidata provides a description and is exported in the RDF dumps as `schema:description`).
- *Forms of abstraction:* As outlined in Section 2.1, classification, generalization, aggregation, and association are among the most important methods to model in a more abstract way. All considered KGs support the modeling of generalization, classification, and association. Freebase and Wikidata also support aggregation.
- *Data types:* The KGs either do not support any data types for literal values, but just store strings (as in case of Cyc), or they support simple data types such as a subset of the XML Schema (see DBpedia, Freebase, Wikidata, and YAGO; a typical data type is `xsd:integer`). The highest number of data types is used by DBpedia[59] and Freebase.[60]

[53] See `http://datahub.io/group/lodcloud`.

[54] See `http://wiki.freebase.com/wiki/Search_Cookbook#Scoring_and_Ranking`

[55] An example for that is `http://dbpedia.org/property/s`.

[56] An example for that is `http://dbpedia.org/property/useEw%25_`

[57] An example for that is `http://dbpedia.org/property/develop`, `http://dbpedia.org/property/developer`, and `http://dbpedia.org/property/develops`.

[58] See `http://lov.okfn.org/dataset/lov/vocabs/dbpedia-owl`

[59] See `http://mappings.dbpedia.org/index.php/DBpedia_Datatypes`.

[60] See `https://wikidata.org/wiki/Special:ListDatatypes`.

Table 6: Comparison of the KGs regarding the schema.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Schema restrictions	yes, DBpedia ontology; in addition non-mapping-based properties with no fixed domain	yes, fixed schema is used, but users are allowed to edit and expand the schema	properties are fixed	properties are fixed, but users are allowed to edit and expand the schema	not mentioned, >560k classes
Schema constraints	inconsistent data may occur	not mentioned	no constraints	no constraint checks, users can report constraint violations	type and constraint checker provided
Network or hierarchy of relations	no, but many fuzzy relations where the meaning is unclear or which is overlapping with other relations	no	yes	not mentioned	no
External vocabulary	yes, FOAF, RDFS, OWL, YAGO, UMBEL	no	no	no, but sometimes a "equivalent property" property links to external vocabulary	yes, RDFS, OWL, WordNet
Description of classes	DBpedia ontology for concepts, in OWL	classes are described in `/common/topic/description` (e.g., `https://www.freebase.com/computer/computer?props=`)	often there is a comment for a collection	yes, just as in entities	no
Forms of abstraction	generalization, classification, association	generalization, classification, aggregation, association	generalization, classification, association	generalization, classification, aggregation, association	generalization, classification, association
Data types	xsd data types, different other data types for currencies etc.[a]	data types such as datetime, int, float, media_type, etc.	not used	only simple data types such as time, timezone, coordinations	simple data types such as xsd:integer

[a] See `http://mappings.dbpedia.org/index.php/DBpedia_Datatypes`

5.7. Comparison of Particularities

Besides general KG information and information about the storage of instances, relations, and the schema of the KGs, there are several aspects of data modeling which can be found only in the models of distinct KGs. These aspects are:

- *Temporal aspects:* There are three types of temporal aspects which can be attached to facts in a KG (see [57]): The valid time, i.e., the point in time or time span the fact is valid; the insertion time, i.e., the time the fact is or was inserted into the KG; and a relevance time aspect, i.e., the point in time or time interval which is relevant for the user's application. Only two of the KGs support the storage of temporal aspects besides pure facts. The data model of Wikidata allows users to store the time interval in which the statement holds true. In this way, facts which are valid only for specific time spans such as election periods can be stored. YAGO also supports temporal information to be stored attached to the fact such as the occurrence date.
- *Source of facts:* For reasons of traceability the source (reference) of facts is stored together with the facts. The knowledge where the facts are derived from might be important for the user to assess the validity and trustability of the fact. Wikidata and YAGO are the only KGs where the storage of facts is both supported by its data model and used by the users.[61]
  The other KGs do not store the source. In case of DBpedia the source of facts is obvious, namely Wikipedia, and does not need to be stored. Cyc is created completely by experts. It can be assumed that the source of facts is not stored here, since all facts are reliable.
- *Reification:* RDF reification was intended as a mechanism for making provenance statements and other statements about RDF triples [67]. Regarding our KGs, only Cyc and YAGO use reification to some extent: Cyc allows the reification of literals. In case of YAGO, time and location is attached to facts by reification.

[61] It can be noted that many facts of Wikidata are derived from Wikipedia, so that in many cases the Wikipedia URL is the only source. For YAGO, the source of facts is provided in a separate file.

6. Assessment of KGs

Based on the Tables 1 – 7, we created a matrix (see Table 8) where the most important aspects in which the KGs differ (extracted from the Tables 1 – 7) are formulated as yes-no-questions. These questions serve the purpose of guiding users that are interested in choosing among the KGs those that best fit their purposes.

For assessing the KGs, a score is calculated for each KG. For each KG, this "fulfillment score" can be calculated as the number of times the answer for the desired KG matches the answer of the KG in question. Also more sophisticated scoring functions are possible where the matching regarding specific questions is weighted higher. In the end, the KG which has achieved the highest score is the KG which is favored by this framework.

7. Limitations of KGs

Peckham and Maryanski [47] argued that semantic data models will be used widely when they perform sufficiently well for real-world settings (especially in enterprises). We can argue that there are already some KG applications and many Linked Open Data datasets available. Examples where Linked Open Data is used are the BBC,[62] Best Buy,[63] and the German National Library.[64]

However, there are several limitations of the KGs and, hence, of the Semantic Web in its current uptake which became apparent during the analysis of the considered KGs and which we therefore would like to emphasize:

1. *Domain specificity limitation*: During the process of selecting the KGs for comparison, it became apparent that either many KBs in the LOD cloud are highly aligned to the general domain Wikipedia covers – since Wikipedia is used as knowledge source in these cases – or (i) the KBs focus on specific domains (cf. the lexical databases WordNet and BabelNet) and/or (ii) cover more schema information than instance information, so that they cannot be called KGs any more (cf. the common sense KBs ConceptNet

[62] See http://www.bbc.co.uk/ontologies

[63] See http://www.bestbuy.com/

[64] See http://www.dnb.de/DE/Service/DigitaleDienste/LinkedData/linkeddata_node.html

Table 7: Comparison of the KGs regarding their particularities.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
Temporal aspects	no	no	no	yes, the valid time of facts (e.g., the population for different points in time)	yes, e.g., the time of occurrence
Source of facts	no (all from Wikipedia)	not mentioned	no	yes, mostly	yes (in file `yagoSources`)
Reification	Currently not exploited	not mentioned	reification of literals	no, decision against it	yes, time and location attached via reification

Table 8: Decision matrix for KG selection.

	DBpedia	Freebase	OpenCyc	Wikidata	YAGO3
1. Is the available KG continuously updated (no fixed versions)?	(✓)[a]	✓	-	✓	-
2. Are other languages than English supported?	✓	✓	-	✓	✓
3. Is it rather an instance KG than a schema KG?	✓	✓	-	✓	✓
4. Is data available via HTTP lookup?	✓	✓	✓	(✓)[b]	✓
5. Is an official SPARQL endpoint provided?	✓	-	-	-	✓
6. Is some data quality level ensured (manually or via manual approvement)?	-	✓	✓	✓	(✓)[c]
7. Is the KG part of the LOD cloud according to `http://datahub.io`?	✓	✓	✓	-	✓
8. Do the entities have any ordering or ranking?	-	✓	-	-	-
9. Are entity descriptions available?	✓	✓	✓	✓	-
10. Do the relations have any ordering or ranking?	-	-	-	✓	-
11. Are relation descriptions available?	✓	✓	-	✓	✓
12. Is the occurrence time of facts stored?	-	-	-	✓	✓
13. Is the source of facts stored?	-	-	-	✓	-
14. Is reification supported and done in practice?	-	-	✓	-	✓
15. Are concepts descriptions available?	-	✓	✓	✓	-
16. Are any standard data types used for literals?	✓	✓	-	✓	✓

[a] Continuous updates are available for DBpedia live, but not for DBpedia
[b] Via HTTP lookups the user can only retrieve the labels and the Wikipedia categories of Wikidata items.
[c] An evaluation of the quality was only performed for YAGO2, but not for YAGO3.

and UMBEL). Although there is a considerable amount of KGs which are freely available and focus on specific domains, most domains with potential use cases are not covered by KGs. Regarding the prediction by Peckham and Maryanski [47] according to which semantic data models are used mainly for the management of scientific, engineering, and manufacturing data, we can state: This data exists, but most of the data is not semantically enriched and/or in the format RDF – and if, there are not many attributes. Future projects on the so called "Industry 4.0" (including concepts of cyber-physical systems, the Internet of Things, and the Internet of Services) aim at changing this.

2. *Limitations of modeling time aspects*: Concepts for modeling dynamic and/or temporal aspects within semantic (graph) data models have been developed since the uprise of semantic data models (see [58,57] for an overview of temporal database theory). Snowgrass and Ahn [57] distinguish between transaction time (time when data is stored), valid time (time the data is useful or valid), and user-defined time (additional time information to be stored) as the dimensions of representing temporal data in databases. Despite research on temporal data models, dynamic and temporal data models have not become prevalent so far. Still today, most semantic graph data models neither encompass the temporal characteristics of knowledge facts nor the spatial-temporal grounded representation of events. Also, Rula et al. [51] showed that the amount of temporal information available in the Linked Open Data cloud is still very small. One reason might be that adding temporal aspects multiplies the number of statements and also may complicate the situation for users and software developers who write queries since queries become more complex. Keeping things simple – and neglecting temporal aspects – is the often selected mantra for building up scalable environments such as KBs.
3. *KG Population*: The Semantic Web suffers from the difficulties of transforming text and other, mainly unstructured data into RDF. KGs of today have already some potential and can be applied to many settings; however, the KGs are dependent on the supply of structured data from external sources. Knowledge extraction tools and ontology learning tools are the key for building KGs.
4. *Limitations regarding the Linked Open Data cloud* (partly based on [34]):
   (a) *Lack of Conceptual Description of Datasets*: In order to identify the domain a specific LOD dataset covers, a human expert is needed. There is currently no standard mechanism or dataset description interface which states that, e.g., MusicBrainz is about music related information while Geonames is about geographical information. This leads to a missing overview of what datasets are there and which can be used in a certain setting. There are some attempts [49,2] to describe LOD datasets,[65] but they do not focus on the conceptual or semantic level, but instead on statistical information or a prosaic description. Since the LOD cloud consists of datasets which were published under the Linked Data principles, nobody knows the complete picture of the LOD cloud. Even the well-known LOD cloud diagram[66] is only a *particular perspective on the Web of Data*, and many other valid perspectives are possible.[67] Other approaches automatically assess, annotate and index linked datasets, e.g., by extracting topic annotations for arbitrary Linked Data datasets [24]; but these tools have not yet become widely used.
   (b) *Lack of LOD Schema Alignment*: Links between LOD datasets are almost exclusively on the level of instances. There are only a few approaches or good practices for mapping concepts at the schema level of the LOD cloud. Although ontology matching has been widely studied in the Semantic Web area and its tools usually produced strict mappings between concepts such as equivalence and subsumption, the situation in case of Linked Data is difficult: Even though concepts may have a strong semantic similarity, the concepts are not necessarily equivalent. One example for an inconsistency is the fact that `dbpedia:Actor` denotes professional ac-

[65] See also the LOD data catalogs `http://datahub.io` and `http://linkeddatacatalog.dws.informatik.uni-mannheim.de`.

[66] See `http://lod-cloud.net/`.

[67] See `http://lod-cloud.net/`.

tors, while the concept `movie:actor` of LinkedMDB[68] means a person who plays a role in a movie, but who is not a stage actor.
The UMBEL ontology was developed to connect schemas used by LOD datasets. However, UMBEL does not take the individual usage patterns of the concepts into account [45]. Remarkable approaches for finding schema-level links between LOD datasets are provided by Nokolov et al. [45] and Jain et al. [33].

(c) *Lack of Expressivity*: Publishing Linked Data in the LOD cloud is done for rapid data releases and for relying on the Web of Data as washing machine (cleaning data over time) [6]. However, in Linked Data the rich features of OWL are rarely used. Although entities can be interlinked between datasets with `owl:sameAs` relations, there is no automatic constraint check or reasoning whether the entities in different datasets contain incoherent information. The city of Berlin, for instance, can have a different population size in DBpedia and in Geonames, and this is not eradicated. This task remains as burden for the data consumers.

8. Outlook

Future work on KGs and Semantic Web technologies might focus on the following areas:

- There are new approaches of how to model knowledge in a semantically-structured form – against the background of having learned of 15 years of ontology engineering. One example of such a new approach is the design and use of so called *ontology design patterns* [26]: An ontology design pattern is a reusable solution to a recurrent modeling problem. The focus is on reusing existing components, since ontologies and ontology components have been reused only to a very limited extend so far.
- There might be new forms of KGs and KBs which do not focus on the storage of entities and their relations, but instead on other things such as events [39,55,65]. Papers published the last years indicate that event-centric KGs will become more important and also widely applicable.
- Recently, there is noticeable progress towards constructing KGs automatically. This is necessary, since constructing and/or populating KGs neither with the help of experts nor with the help of open communities does not scale to an extend that is needed for most applications. For instance, in Freebase the place of birth relation was missing for 71% of all people instances, although this relation was mandatory according to the schema [68]. Also, Buh et al. [60] showed that the growth of Wikipedia has been slowing down. Consequently, automatic knowledge base construction (AKBC) methods have been attracted more attention [44]. Noteworthy in this context is the approach of statistical inference in KGs. Predictive models are trained on known facts from the KG. From that, unknown facts are derived and compared to "noisy" facts extracted from external, often unstructured sources such as the Web. New facts are added to the KG if they are supported by both models with a certain confidence. This methodology is for instance used in Google's Knowledge Vault project [21].

[68] See `http://www.linkedmdb.org/`.

9. Conclusion

Freely available knowledge graphs (KGs) have not been in the focus of any extensive comparative study so far. In this survey, we defined aspects according to which KGs can be analyzed. We analyzed and compared DBpedia, Freebase, Cyc, Wikidata, and YAGO along these aspects and proposed a checklist to enable readers to find the most suitable KG for their settings. We discussed the essential issues current KGs are conflicted with and glanced over the possible future of the Semantic Web.

References

[1] S. Abiteboul and R. Hull. IFO: A Formal Semantic Database Model. *ACM Trans. Database Syst.*, 12(4):525–565, Nov. 1987.

[2] K. Alexander. Describing Linked Datasets – On the Design and Usage of voiD, the 'Vocabulary Of Interlinked Datasets'. *WWW 2009 Workshop: Linked Data on the Web*, 2009.

[3] B. Amann and M. Scholl. Gram: A Graph Data Model and Query Languages. In *Proceedings of the ACM Conference on Hypertext*, ECHT '92, pages 201–211, New York, NY, USA, 1992. ACM.

[4] M. Andries, M. Gemis, J. Paredaens, I. Thyssens, and J. V. den Bussche. Concepts for Graph-Oriented Object Manipulation. In *Proceedings of the 3rd International Conference on Extending Database Technology: Advances in Database Technology*, EDBT '92, pages 21–38, London, UK, UK, 1992. Springer-Verlag.

[5] R. Angles and C. Gutierrez. Survey of Graph Database Models. *ACM Computing Surveys*, 40(1):1:1–1:39, 2 2008.

[6] S. Auer. Creating Knowledge out of Interlinked Data: Making the Web a Data Washing Machine. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, WIMS '11, pages 4:1–4:8, New York, NY, USA, 2011. ACM.

[7] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. DBpedia: A Nucleus for a Web of Open Data. In *Proceedings of the 6th International Semantic Web Conference and 2nd Asian Semantic Web Conference*, ISWC 2007/ASWC 2007, pages 722–735. Springer, 2007.

[8] S. Auer, J. Lehmann, A.-C. Ngonga Ngomo, and A. Zaveri. Introduction to Linked Data and Its Lifecycle on the Web. In *Reasoning Web. Semantic Technologies for Intelligent Data Access*, volume 8067 of *Lecture Notes in Computer Science*, pages 1–90. Springer Berlin Heidelberg, 2013.

[9] T. Berners-Lee. Linked Data – Design issues. `http://www.w3.org/Designissues/LinkedData.html`. accessed May 15,2015.

[10] T. Berners-Lee. Cool URIs don't change. Technical report, World Wide Web Consortium, 1998.

[11] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):29–37, 5 2001.

[12] J. Brank, M. Grobelnik, and D. Mladenic. A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170, 2005.

[13] T. Bray, J. Paoli, and C. M. Sperberg-McQueen. Extensible Markup Language (XML) 1.0, W3C Recommendation 10. `http://www.w3.org/TR/1998/REC-xml-19980210`. accessed July 31, 2015.

[14] M. L. Brodie. On the Development of Data Models. In M. L. Brodie, J. Mylopoulos, and J. W. Schmidt, editors, *On Conceptual Modelling*, Topics in Information Systems, pages 19–47. Springer New York, 1984.

[15] P. Buneman. Semistructured Data. In *Proceedings of the Sixteenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '97, pages 117–121, New York, NY, USA, 1997. ACM.

[16] P. P.-S. Chen. The Entity-relationship Model: Toward a Unified View of Data. *ACM Trans. Database Syst.*, 1(1):9–36, Mar. 1976.

[17] E. F. Codd. A Relational Model of Data for Large Shared Data Banks. *Commun. ACM*, 13(6):377–387, June 1970.

[18] E. F. Codd. Data Models in Database Management. *SIGPLAN Not.*, 16(1):112–114, June 1980.

[19] J. Curtis, J. Cabral, and D. Baxter. On the Application of the Cyc Ontology to Word Sense Disambiguation. In *FLAIRS Conference*, pages 652–657, 2006.

[20] R. Cyganiak, D. Wood, and M. Lanthaler. RDF 1.1 Concepts and Abstract Syntax. 2014. accessed July 30, 2015.

[21] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang. Knowledge Vault: A Web-scale Approach to Probabilistic Knowledge Fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 601–610, New York, NY, USA, 2014. ACM.

[22] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić. Introducing Wikidata to the Linked Data Web. In *Proceedings of the 13th International Semantic Web Conference (ISWC'14)*, LNCS. Springer, 2014.

[23] C. Fellbaum. *WordNet – An Electronic Lexical Database*. MIT Press, 1998.

[24] B. Fetahu, S. Dietze, B. P. Nunes, D. Taibi, and M. A. Casanova. Profiling of Linked Datasets using Structured Descriptions. In *The 12th International Semantic Web Conference (ISWC2013)*, 2013.

[25] G. Freedman and E. Reynolds. Enriching basal reader lessons with semantic webbing. *Reading Teacher*, 33(6):677–684, 1980.

[26] A. Gangemi. Ontology Design Patterns for Semantic Web Content. In Y. Gil, E. Motta, V. Benjamins, and M. Musen, editors, *The Semantic Web – ISWC 2005*, volume 3729 of *Lecture Notes in Computer Science*, pages 262–276. Springer Berlin Heidelberg, 2005.

[27] R. Guns. Tracing the Origins of the Semantic Web. *Journal of the American Society for Information Science and Technology*, 64(10):2173–2181, 2013.

[28] M. Gyssens, J. Paredaens, and D. van Gucht. A Graph-oriented Object Database Model. In *Proceedings of the 9th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, PODS '90, pages 417–424, New York, NY, USA, 1990. ACM.

[29] M. Hammer and D. Mc Leod. Database Description with SDM: A Semantic Database Model. *ACM Trans. Database Syst.*, 6(3):351–386, Sept. 1981.

[30] M. Hammer and D. McLeod. The Semantic Data Model: A Modelling Mechanism for Data Base Applications. In *Proceedings of the 1978 ACM SIGMOD International Conference on Management of Data*, SIGMOD '78, pages 26–36, New York, NY, USA, 1978. ACM.

[31] R. Hull and R. King. Semantic Database Modeling: Survey, Applications, and Research Issues. *ACM Comput. Surv.*, 19(3):201–260, Sept. 1987.

[32] P. Jain, P. Hitzler, K. Janowicz, and C. Venkatramani. There's No Money in Linked Data. `http://corescholar.libraries.wright.edu/cse/240`, 2013. accessed July 20, 2015.

[33] P. Jain, P. Hitzler, A. P. Sheth, K. Verma, and P. Z. Yeh. Ontology Alignment for Linked Open Data. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part I*, ISWC'10, pages 402–417, Berlin, Heidelberg, 2010. Springer-Verlag.

[34] P. Jain, P. Hitzler, P. Z. Yeh, K. Verma, and A. P. Sheth. Linked Data Is Merely More Data. In *AAAI Spring Symposium: linked data meets artificial intelligence*, volume 11, 2010.

[35] L. Kerschberg, A. C. Klug, and D. Tsichritzis. A Taxonomy of Data Models. In *Systems for Large Data Bases*, pages 43–64. North Holland & IFIP, 1976.

[36] W. Kim. Object-Oriented Databases: Definition and Research Directions. *IEEE Transactions on Knowledge and Data Engineering*, 2(3):327–341, 1990.

[37] G. Klyne and J. J. Carroll. Resource Description Framework (RDF): Concepts and Abstract Syntax. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210, 2004. accessed July 20, 2015.

[38] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC 2009 Heraklion, pages 723–737, Berlin, Heidelberg, 2009. Springer-Verlag.

[39] E. Kuzey, J. Vreeken, and G. Weikum. A Fresh Look on Knowledge Bases: Distilling Named Events from News. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pages 1689–1698, New York, NY, USA, 2014. ACM.

[40] O. Lassila and R. R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/, 1999. accessed July 4, 2015.

[41] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2), 2012.

[42] A. Lozano-Tello and A. Gómez-Pérez. Ontometric: A method to choose the appropriate ontology. *Journal of Database Management*, 2(15):1–18, 2004.

[43] F. Mahdisoltani, J. Biega, and F. M. Suchanek. YAGO3: A Knowledge Base from Multilingual Wikipedias. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*, 2015.

[44] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich. A Review of Relational Machine Learning for Knowledge Graphs: From Multi-Relational Link Prediction to Automated Knowledge Graph Construction. *arXiv preprint arXiv:1503.00759*, 2015.

[45] A. Nikolov, V. Uren, E. Motta, and A. de Roeck. Overcoming Schema Heterogeneity between Linked Semantic Repositories to Improve Coreference Resolution. In A. Gómez-Pérez, Y. Yu, and Y. Ding, editors, *The Semantic Web*, volume 5926 of *Lecture Notes in Computer Science*, pages 332–346. Springer Berlin Heidelberg, 2009.

[46] J. Paredaens, P. Peelman, and L. Tanca. G-Log: A Graph-Based Query Language. *IEEE Trans. on Knowl. and Data Eng.*, 7(3):436–453, June 1995.

[47] J. Peckham and F. Maryanski. Semantic Data Models. *ACM Comput. Surv.*, 20(3):153–189, Sept. 1988.

[48] M. Poveda-Villalón, M. Suárez-Figueroa, and A. Gómez-Pérez. Did You Validate Your Ontology? OOPS! In E. Simperl, B. Norton, D. Mladenic, E. Della Valle, I. Fundulaki, A. Passant, and R. Troncy, editors, *The Semantic Web: ESWC 2012 Satellite Events*, volume 7540 of *Lecture Notes in Computer Science*, pages 402–407. Springer Berlin Heidelberg, 2015.

[49] B. Quilitz and U. Leser. Querying Distributed RDF Data Sources with SPARQL. In *Proceedings of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications*, ESWC'08, pages 524–538, Berlin, Heidelberg, 2008. Springer-Verlag.

[50] M. A. Rodriguez. A graph analysis of the Linked Data cloud. *arXiv preprint arXiv:0903.0194*, 2009. accessed July 31, 2015.

[51] A. Rula, M. Palmonari, A. Harth, S. Stadtmüller, and A. Maurino. On the Diversity and Availability of Temporal Information in Linked Open Data. In *The Semantic Web – ISWC 2012*, volume 7649 of *Lecture Notes in Computer Science*, pages 492–507. Springer Berlin Heidelberg, 2012.

[52] E. Sandhaus. Semantic Technology at the New York Times: Lessons Learned and Future Directions. In *Proceedings of the 9th International Semantic Web Conference on The Semantic Web - Volume Part II*, ISWC'10, pages 355–355, Berlin, Heidelberg, 2010. Springer-Verlag.

[53] L. Sauermann and R. Cyganiak. Cool URIs for the Semantic Web. W3C Note, http://www.w3.org/TR/2008/NOTE-cooluris-20081203/, 12 2008. accessed July 10, 2015.

[54] H. A. Schmid and J. R. Swenson. On the Semantics of the Relational Data Model. In *Proceedings of the 1975 ACM SIGMOD International Conference on Management of Data*, SIGMOD '75, pages 211–223, New York, NY, USA, 1975. ACM.

[55] R. Segers, P. Vossen, M. Rospocher, L. Serafini, E. Laparra, and G. Rigau. ESO: a Frame based Ontology for Events and Implied Situations. *Proceedings of Maplex2015*, 2015.

[56] J. M. Smith and D. C. P. Smith. Database Abstractions: Aggregation and Generalization. *ACM Trans. Database Syst.*, 2(2):105–133, June 1977.

[57] R. Snodgrass and I. Ahn. A Taxonomy of Time Databases. In *Proceedings of the 1985 ACM SIGMOD International Conference on Management of Data*, SIGMOD '85, pages 236–246, New York, NY, USA, 1985. ACM.

[58] R. T. Snodgrass. Temporal databases. In *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, volume 639 of *Lecture Notes in Computer Science*, pages 22–64. Springer Berlin Heidelberg, 1992.

[59] S. Staab and R. Studer. *Handbook on Ontologies*. Springer Publishing Company, Incorporated, 2nd edition, 2009.

[60] B. Suh, G. Convertino, E. H. Chi, and P. Pirolli. The Singularity is Not Near: Slowing Growth of Wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 8:1–8:10, New York, NY, USA, 2009. ACM.

[61] S. Tartir, I. B. Arpinar, M. Moore, A. P. Sheth, and B. Aleman-meza. OntoQA: Metric-based ontology quality analysis. In *IEEE Workshop on Knowledge Acquisition from Distributed, Autonomous, Semantically Heterogeneous Data and Knowledge Sources*, 2005.

[62] R. W. Taylor and R. L. Frank. CODASYL Data-Base Management Systems. *ACM Comput. Surv.*, 8(1):67–103, Mar. 1976.

[63] D. C. Tsichritzis and F. H. Lochovsky. Hierarchical Data-Base Management: A Survey. *ACM Comput. Surv.*, 8(1):105–123, Mar. 1976.

[64] D. C. Tsichritzis and F. H. Lochovsky. *Data Models*. Prentice Hall Professional Technical Reference, 1982.

[65] P. Vossen, T. Caselli, and Y. Kontzopoulou. Storylines for structuring massive streams of news. *ACL-IJCNLP 2015*, pages 40–49, 2015.

[66] D. Vrandečić and Y. Sure. How to Design Better Ontology

Metrics. In E. Franconi, M. Kifer, and W. May, editors, *The Semantic Web: Research and Applications*, volume 4519 of *Lecture Notes in Computer Science*, pages 311–325. Springer Berlin Heidelberg, 2007.

[67] E. R. Watkins and D. A. Nicole. Named Graphs as a Mechanism for Reasoning About Provenance. In X. Zhou, J. Li, H. Shen, M. Kitsuregawa, and Y. Zhang, editors, *Frontiers of WWW Research and Development - APWeb 2006*, volume 3841 of *Lecture Notes in Computer Science*, pages 943–948. Springer Berlin Heidelberg, 2006.

[68] R. West, E. Gabrilovich, K. Murphy, S. Sun, R. Gupta, and D. Lin. Knowledge Base Completion via Search-based Question Answering. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 515–526, New York, NY, USA, 2014. ACM.