What are the Important Properties of an Entity?
Comparing Users and Knowledge Graph Point of View

Ahmad Assaf[1], Ghislain A. Atemezing[1], Raphaël Troncy[1] and Elena Cabrio[1,2]

[1] EURECOM, Sophia Antipolis, France. <firstName.lastName@eurecom.fr>
[2] INRIA, Sophia Antipolis, France. <elena.cabrio@inria.fr>

Abstract. Entities play a key role in knowledge bases in general and in the Web of Data in particular. Entities are generally described with a lot of properties, this is the case for DBpedia. It is, however, difficult to assess which ones are more "important" than others for particular tasks such as visualizing the key facts of an entity or filtering out the ones which will yield better instance matching. In this paper, we perform a reverse engineering of the Google Knowledge graph panel to find out what are the most "important" properties for an entity according to Google. We compare these results with a survey we conducted on 152 users. We finally show how we can represent and explicit this knowledge using the Fresnel vocabulary.

Keywords: Entities, Google Knowledge Graph, knowledge extraction

1 Introduction

In many knowledge bases, entities are described with numerous properties. However, not all properties have the same importance. Some properties are considered as keys for performing instance matching tasks while other properties are generally chosen for quickly providing a summary of the key facts attached to an entity. Our motivation is to provide a method enabling to select what properties should be used when depicting the summary of an entity, for example in a multimedia question answering system such as QakisMedia[3] or in a second screen application providing more information about a particular TV program[4]. Our approach consists in: (i) reverse engineering the Google Knowledge Panel by extracting the properties that Google considers as sufficiently important to show (Section 2), and (ii) analyzing users' preferences by conducting a user survey and comparing the results (Section 3). We finally show how we can explicitly represent this knowledge of preferred properties to attach to an entity using the Fresnel vocabulary before concluding (Section 4).

2 Reverse Engineering the Google KG Panel

Web scraping is a technique for extracting data from Web pages. We aim at capturing the properties depicted in the Google Knowledge Panel (GKP) that are

[3] http://qakis.org/
[4] http://www.linkedtv.eu/demos/linkednews/

injected in search result pages [14]. We have developed a Node.js application that queries all DBpedia concepts that have at least one instance which is `owl:sameAs` with a Freebase resource in order to increase the probability that the search engine result page (SERP) for this resource will contain a GKP. We assume in our experiments that the properties displayed for an entity are type and context dependent (country, time, query) which can affect the results. Moreover, we filter out generic concepts by excluding those who are direct subclasses of `owl:Thing` since they will trigger ambiguous queries. We obtained a list of 352 concepts[5]. For

Algorithm 1 Google Knowledge Panel reverse engineering algorithm

```
INITIALIZE equivalentClasses(DBpedia, Freebase) AS vectorClasses
Upload vectorClasses for querying processing
Set n AS number-of-instances-to-query
for each conceptType ∈ vectorClasses do
    SELECT n instances
    listInstances ← SELECT-SPARQL(conceptType, n)
    for each instance ∈ listInstances do
        CALL http://www.google.com/search?q=instance
        if knowledgePanel exists then
            SCRAP GOOGLE KNOWLEDGE PANEL
        else
            CALL http://www.google.com/search?q=instance + conceptType
            SCRAP GOOGLE KNOWLEDGE PANEL
        end if
        gkpProperties ← GetData(DOM, EXIST(GKP))
    end for
    COMPUTE occurrences for each prop ∈ gkpProperties
end for
return gkpProperties
```

each of these concepts, we retrieve n instances (in our experiment, n was equal to 100 random instances). For each of these instances, we issue a search query to Google containing the instance label. Google does not serve the GKP for all user agents and we had to mimic a browser behavior by setting the $User-Agent$ to a particular browser. We use CSS selectors to check the existence of and to extract data from a GKP. An example of a query selector is *._om* (all elements with class name *_om*) which returns the property DOM element(s) for the concept described in the GKP. From our experiments, we found out that we do not always get a GKP in a SERP. If this happens, we try to disambiguate the instance by issuing a new query with the concept type attached. However, if no GKP was found again, we capture that for manual inspection later on. Listing 1 gives the high level algorithm for extracting the GKP. The full implementation can be found at `https://github.com/ahmadassaf/KBE`. We finally observe that this experiment is only valid for the English Google.com search results since GKP varies according to top level names.

3 Evaluation

We conducted a user survey in order to compare what users think should be the important properties to display for a particular entity and what the GKP shows.

[5] See also the SPARQL query at `http://goo.gl/EYuGm1`

User survey.

We set up a survey[6] on February 25th, 2014 and for three weeks in order to collect the preferences of users in term of the properties they would like to be shown for a particular entity. We select only one representative entity for nine classes: `TennisPlayer`, `Museum`, `Politician`, `Company`, `Country`, `City`, `Film`, `SoccerClub` and `Book`. 152 participants have provided answers, 72% from academia, 20% coming from the industry and 8% having not declared their affiliation. 94% of the respondents have heard about the Semantic Web while 35% were not familiar with specific visualization tools. The detailed results[7] show the ranking of the top properties for each entity. We only keep the properties having received at least 10% votes for comparing with the properties depicted in a KGP. We observe that users do not seem to be interested in the `INSEE code` identifying a French city while they expect to see the `population` or the `points of interest` of this city.

Comparison with the Knowledge Graphs. The results of the Google Knowledge Panel (GKP) extraction[8] clearly show a long tail distribution of the properties depicted by Google, with a top N properties (N being 4, 5 or 6 depending on the entity) counting for 98% of the properties shown for this type. We compare those properties with the ones revealed by the user study. Table 1 shows the agreement between the users and the choices made by Google in the GKP for the 9 classes. The highest agreement concerns the type `Museum` (66.97%) while the lowest one is for the `TennisPlayer` (20%) concept. We think properties for museums or books are more stable than for types such as person/agent which vary significantly. We acknowledge the fact that more than one instance should be tested in order to draw meaningful conclusion regarding what are the important properties for a type. With this set of 9 concepts, we are covering

Classes	TennisPlayer	Museum	Politician	Company	Country	City	Film	SoccerClub	Book
Agr.	20%	66.97%	50%	40%	60%	60%	60%	50%	60%

Table 1. Agreement on properties between users and the Knowledge Graph Panel

301,189 DBpedia entities that have an existence in Freebase, and for each of them, we can now empirically define the most important properties when there is an agreement between one of the biggest knowledge base (Google) and users preferences.

Modeling the preferred properties with Fresnel. Fresnel[9] is a presentation vocabulary for displaying RDF data. It specifies *what* information contained in an RDF graph should be presented with the core concept `fresnel:Lens` [128].

[6] The survey is at `http://eSurv.org?u=entityviz`

[7] `https://github.com/ahmadassaf/KBE/blob/master/results/agreement-gkp-users.xls`

[8] `https://github.com/ahmadassaf/KBE/blob/master/results/survey.json`

[9] `http://www.w3.org/2005/04/fresnel-info/`

We use the Fresnel and PROV-O ontologies[10] to explicitly represent what properties should be depicted when displaying an entity. This dataset can now be re-used as a configuration for any consuming application.

```
:tennisPlayerGKPDefaultLens rdf:type fresnel:Lens ;
  fresnel:purpose fresnel:defaultLens ;
  fresnel:classLensDomain dbpedia-owl:TennisPlayer ;
  fresnel:group :tennisPlayerGroup ;
  fresnel:showProperties (dbpedia-owl:abstract dbpedia-owl:birthDate
    dbpedia-owl:birthPlace dbpprop:height dbpprop:weight
    dbpprop:turnedpro dbpprop:siblings) ;
  prov:wasDerivedFrom
    <http://www.google.com/insidesearch/features/search/knowledge.html> .
```

Listing 1.1. Excerpt of a Fresnel lens in Turtle

4 Conclusion and Future Work

We have shown that it is possible to reveal what are the "important" properties of entities by reverse engineering the choices made by Google when creating knowledge graph panels and by comparing users preferences obtained from a user survey. Our motivation is to represent this choice explicitly, using the Fresnel vocabulary, so that any application could read this configuration file for deciding which properties of an entity is worth to visualize. This is fundamentally different from the work in [142] where the authors created a generalizable approach to open up closed knowledge bases like Google's by means of crowd-sourcing the knowledge extraction task. We are aware that this knowledge is highly dynamic, the Google Knowledge Graph panel varies across geolocation and time. We have provided the code that enables to perform new calculation at run time and we aim to study the temporal evolution of what are important properties on a longer period. This knowledge which has been captured will be made available shortly in a SPARQL endpoint. We are also investigating the use of Mechanical Turk to perform a larger survey for the complete set of DBpedia classes.

Acknowledgments. This work has been partially supported by Datalift (ANR-10-CORD-009), UCN (ANR-11-LABX-0031-01) and LinkedTV (GA 287911).

References

1. Z. Abedjan, T. Grüetze, A. Jentzsch, and F. Naumann. Profiling and mining RDF data with ProLOD++. In *30th IEEE International Conference on Data Engineering (ICDE)*, pages 1198–1201, 2014.
2. M. Acosta, A. Zaveri, E. Simperl, and D. Kontokostas. Crowdsourcing Linked Data quality assessment. In *12th International Semantic Web Conference (ISWC)*, 2013.
3. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing Linked Datasets. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.

[10] `http://www.w3.org/TR/prov-o/`

4. R. Anisa and A. Zaveri. Methodology for Assessment of Linked Data Quality. In *1st Workshop on Linked Data Quality (LDQ)*, 2014.
5. A. Assaf, E. Louw, A. Senart, C. Follenfant, R. Troncy, and D. Trastour. RUBIX: a framework for improving data integration with linked data. In *International Workshop on Open Data (WOD'12)*, pages 13–21, 2012.
6. A. Assaf and A. Senart. Data Quality Principles in the Semantic Web. In *6th International Conference on Semantic Computing ICSC '12*, 2012.
7. A. Assaf, A. Senart, and R. Troncy. SNARC - An Approach for Aggregating and Recommending Contextualized Social Content. In *The Semantic Web: ESWC 2013 Satellite Events, Revised Selected Papers*, pages 319–326, 2013.
8. A. Assaf, A. Senart, and R. Troncy. Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In *24th World Wide Web Conference (WWW'14), Demos Track*, Florence, Italy, 2015.
9. A. Assaf, R. Troncy, and A. Senart. An Objective Assessment Framework & Tool for Linked Data Quality - Enriching Dataset Profiles with Quality Indicators (Major Revision). *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2015.
10. A. Assaf, R. Troncy, and A. Senart. HDL-Towards a Harmonized Dataset Model for Open Data Portals. In *2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data*, Portoroz, Slovenia, 2015.
11. A. Assaf, R. Troncy, and A. Senart. Roomba: An Extensible Framework to Validate and Build Dataset Profiles. In *12th European Semantic Web Conference (ESWC)*, Portoroz, Slovenia, 2015.
12. A. Assaf, R. Troncy, and A. Senart. What's up LOD Cloud? Observing The State of Linked Open Data Cloud Metadata. In *12th European Semantic Web Conference (ESWC)*, Portoroz, Slovenia, 2015.
13. S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats - an Extensible Framework for High-performance Dataset Analytics. In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 353–362, Galway, Ireland, 2012.
14. M. Bergman. Deconstructing the Google Knowledge Graph. `http://www.mkbergman.com/1009/deconstructing-the-google-knowledge-graph`.
15. T. Berners-Lee. Uniform Resource Identifier (URI): Generic Syntax. RFC 3986, 2005. `http://tools.ietf.org/html/rfc3986`.
16. T. Berners-Lee. Linked Data - Design Issues. W3C Personal Notes, 2006. `http://www.w3.org/DesignIssues/LinkedData`.
17. T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001.
18. D. Berrueta, S. Fernández, and I. Frade. Cooking HTTP content negotiation with Vapour. In *4th Workshop on Scripting for the Semantic Web (SFSW'08)*, 2008.
19. G. L. S. Besiki, , M. B. Twidale, and L. C. Smith. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 2007.
20. C. Bizer. Evolving the Web into a Global Data Space. In *28th British National Conference on Advances in Databases*, 2011.
21. C. Bizer and T. H. T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
22. C. Bizer and R. Cyganiak. Quality-driven information filtering using the wiqa policy framework. *Jorunal of Web Semantics*, 7(1), 2009.

23. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A Crystallization Point for the Web of Data. *Journal of Web Semantics*, 7(3), 2009.
24. C. Böhm, G. Kasneci, and F. Naumann. Latent Topics in Graph-structured Data. In *21st ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2663–2666, Maui, Hawaii, USA, 2012.
25. C. Böhm, J. Lorey, and F. Naumann. Creating voiD Descriptions for Web-scale Data. *Journal of Web Semantics*, 9(3):339–345, 2011.
26. C. Böhm, F. Naumann, Z. Abedjan, D. Fenz, T. Grütze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with ProLOD. In *26th International Conference on Data Engineering Workshops (ICDEW)*, 2010.
27. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In *ACM International Conference on Management of Data (SIGMOD)*, 2008.
28. D. Boyd and K. Crawford. Six Provocations for Big Data. *Social Science Research Network Working Paper Series*, 2011.
29. D. Brickley and R. Guha. RDF Schema 1.1. W3C Recommendation, 2014. `http://www.w3.org/TR/rdf-schema`.
30. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *7th International Conference on World Wide Web (WWW'98)*, 1998.
31. M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. DataID: Towards Semantically Rich Metadata for Complex Datasets. In *10th International Conference on Semantic Systems*, 2014.
32. C. Buil-Aranda and A. Hogan. SPARQL Web-Querying Infrastructure: Ready for Action? In *12th International Semantic Web Conference (ISWC)*, 2013.
33. S. Chakrabarti, B. E. Dom, R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg. Mining the web's link structure. *Computer*, 1999.
34. D. Cherix, R. Usbeck, A. Both, and J. Lehmann. CROCUS: Cluster-based ontology data cleansing. In *2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice*, 2014.
35. M. Cornolti, P. Ferragina, and M. Ciaramita. A Framework for Benchmarking Entity-annotation Systems. In *22nd International World Wide Web Conference (WWW'13)*, 2013.
36. R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In *5th European Semantic Web Conference (ESWC)*, pages 690–704, Tenerife, Spain, 2008.
37. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing Linked Datasets with the VoID Vocabulary. W3C Note, 2011. `http://www.w3.org/TR/void/`.
38. A. S. da Silva, D. Barbosa, J. M. B. Cavalcanti, and M. A. S. Sevalho. Labeling Data Extracted from the Web. In *On The Move Confederated International Conferences*, pages 1099–1116, 2007.
39. M. d'Aquin and E. Motta. Watson, More Than a Semantic Web Search Engine. *Semantic Web Journal*, 2011.
40. T. Davies, R. Sharif, and J. Alonso. Open Data Barometer - Global Report. Technical report, World Wide Web Foundation, 2015. `http://barometer.opendataresearch.org/`.
41. D. de Castro Reis, P. B. Golgher, A. S. da Silva, and A. H. F. Laender. Automatic Web News Extraction Using Tree Edit Distance. In *13th International World Wide Web Conference (WWW'04)*, pages 502–601, 2004.

42. J. Debattista, M. Dekkers, and D. Lee. Data Quality Vocabulary (DQV). W3C First Public Working Draft, 2015. `http://www.w3.org/TR/vocab-dqv/`.
43. J. Debattista, C. Lange, and S. Auer. daQ, an Ontology for Dataset Quality Information. In *7th International Workshop on Linked Data on the Web (LDOW)*, 2014.
44. J. Debattista, S. Londoño, C. Lange, and S. Auer. LUZZU - A framework for linked data quality assessment. *CoRR*, abs/1412.3750, 2014.
45. R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. In *7th European Semantic Web Conference (ESWC)*, 2010.
46. L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. In *13st ACM International Conference on Information and Knowledge Management (CIKM)*, 2004.
47. D.-A. Ernesto, D. Lucas, S.-T. Lars, and N. Wolfgang. Real-time top-n recommendation in social streams. In *6th ACM conference on Recommender systems - RecSys*, 2012.
48. B. Eytan, R. Itamar, M. Cameron, and A. Lada. The role of social networks in information diffusion. In *21th International Conference on World Wide Web (WWW'12)*, 2012.
49. T. Fawcett. An Introduction to ROC Analysis. *Pattern Recogn. Lett.*, 2006.
50. B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl. A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles. In *11th European Semantic Web Conference (ESWC)*, 2014.
51. T. Finin, Z. Syed, J. Mayfield, P. Mcnamee, and C. Piatko. Using Wikitology for Cross-Document Entity Coreference Resolution. In *AAAI Spring Symposium on Learning*, 2009.
52. A. Flemming. Quality Characteristics of Linked Data Publishing Datasources. Master's thesis, Humboldt-Universität zu Berlin, 2010.
53. G. Flouris, Y. Roussakis, and M. Poveda-Villalón. Using provenance for quality assessment and repair in linked open data. In *2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn'12)*, 2012.
54. B. Forchhammer, A. Jentzsch, and F. Naumann. LODOP - Multi-Query Optimization for Linked Data Profiling Queries. In *International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES)*, Heraklion, Greece, 2014.
55. P. Frischmuth, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig, and C. Marquardt. Towards Linked Data based Enterprise Information Integration. In *Workshop on Semantic Web Enterprise Adoption and Best Practice Co-located with 12th International Semantic Web Conference (ISWC'13)*, 2013.
56. P. Frischmuth, J. Klímek, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig, and C.-M. Marquardt. Linked Data in Enterprise Information Integration. *Semantic Web Journal*, 2012.
57. M. Frosterus, E. Hyvönen, and J. Laitio. Creating and Publishing Semantic Metadata about Linked and Open Datasets. In *Linking Government Data*. Springer, 2011.
58. M. Frosterus, E. Hyvönen, and J. Laitio. DataFinland - A Semantic Portal for Open and Linked Datasets. In *8th Extended Semantic Web Conference (ESWC)*, pages 243–254, 2011.
59. C. Fürber and M. Hepp. SWIQA - A Semantic Web information quality assessment framework. In *19th European Conference on Information Systems (ECIS'11)*, 2011.

60. G. Gouriten and P. Senellart. API Blender: A Uniform Interface to Social Platform APIs. In *21th International Conference on World Wide Web (WWW'12)*, 2012.
61. W. O. W. Group. OWL 2 Web Ontology Language. W3C Recommendation, 2012. `http://www.w3.org/TR/owl2-overview`.
62. T. R. Gruber. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 1993.
63. C. Guéret, P. Groth, C. Stadler, and J. Lehmann. Assessing Linked Data Mappings Using Network Measures. In *9th European Semantic Web Conference (ESWC)*, 2012.
64. R. Hammell, C. Bates, H. Lewis, C. Perricos, L. Brett, and D. Branch. Open Data: Driving growth, ingenuity and innovation. Technical report, Deloitte LLP, 2012. `http://www2.deloitte.com/content/dam/Deloitte/uk/Documents/deloitte-analytics/open-data-driving-growth-ingenuity-and-innovation.pdf`.
65. P. Harpring. *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. Getty Research Institute, 2010.
66. A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data Summaries for On-demand Queries over Linked Data. In *19th World Wide Web Conference (WWW'10)*, 2010.
67. A. Harth, S. Kinsella, and S. Decker. Using naming authority to rank data and ontologies for web search. In *8th International Semantic Web Conference (ISWC)*, 2009.
68. O. Hartig and J. Zhao. Using web data provenance for quality assessment. In *8th International Semantic Web Conference (ISWC)*, 2009.
69. B. Haslhofer and N. Popitsch. DSNotify: Detecting and Fixing Broken Links in Linked Data Sets. In *8th International Workshop on Web Semantics*, 2009.
70. O. Hassanzadeh, Duan, A. Fokoue, A. Kementsietsidis, K. Srinivas, and M. J. Ward. Helix: Online Enterprise Data Analytics. In *20th World Wide Web Conference (WWW'11)*, pages 225–228, 2011.
71. M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayer. SCOVO: Using Statistics on the Web of Data. In *6th European Semantic Web Conference on The Semantic Web (ESWC)*, 2009.
72. A. Hogan, A. Harth, and S. Decker. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.
73. A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In *3rd International Workshop on Linked Data on the Web (LDOW)*, 2010.
74. A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics*, 2012.
75. R. Iannella and J. McKinney. vCard Ontology - for describing People and Organizations. W3C Interest Group Note, 2014. `http://www.w3.org/TR/vcard-rdf`.
76. H. B. Inteligence. Data Quality - Why you should care about the cleanliness of your data. Technical report, Halo, 2013.
77. A. Isaac and E. Summers. SKOS Simple Knowledge Organization System Primer. W3C Working Group Note, 2009.
78. R. Isele, J. Umbrich, C. Bizer, and A. Harth. LDspider: An Open-source Crawling Framework for the Web of Linked Data. In *9th International Semantic Web Conference (ISWC), Posters & Demos Track*, 2010.

79. C. Iván and B. Alejandro. Semantic contextualisation of social tag-based profiles and item recommendations. In *12th Internationl Conference on E-Commerce and Web Technolgoies*, 2011.
80. P. Jain, P. Hitzler, K. Janowicz, and C. Venkatramani. There's No Money in Linked Data, 2013. `http://knoesis.wright.edu/faculty/pascal/pub/nomoneylod.pdf`.
81. A. Jentzsch. Profiling the Web of Data. In *13th International Semantic Web Conference (ISWC), Doctoral Consortium*, Trentino, Italy, 2014.
82. A. Jentzsch, R. Cygania, and C. Bizer. State of the lod cloud. `http://lod-cloud.net/state/`.
83. T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing Linked Data Dynamics. In *10th European Semantic Web Conference (ESWC)*, 2013.
84. B. K. Kahn, D. M. Strong, and R. Y. Wang. Information quality benchmarks: product and service performance. *Communications of the ACM*, 2002.
85. H. Kang and B. Shneiderman. MediaFinder: an interface for dynamic personal media management with semantic regions. In *Conference on Human Factors in Computing Systems (CHI)*, pages 764–765. ACM, 2003.
86. C. M. Keet, M. del Carmen Suárez-Figueroa, and M. Poveda-Villalón. The Current Landscape of Pitfalls in Ontologies. In *International Conference on Knowledge Engineering and Ontology Development (KEOD)*, 2013.
87. S. Khatchadourian and M. P. Consens. ExpLOD: Summary-based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In *7th Extended Semantic Web Conference (ESWC)*, pages 272–287, Heraklion, Greece, 2010.
88. H. Khrouf, G. Atemezing, G. Rizzo, and R. Troncy. Aggregating Social Media for Enhancing Conference Experiences. In *1st Internationl Workshop on Real-Time Analysis and Mining of Social Streams*, 2012.
89. R. Kimball, L. Reeves, W. Thornthwaite, M. Ross, and W. Thornwaite. *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing, Developing and Deploying Data Warehouses with CD Rom.* John Wiley & Sons, Inc., 1st edition, 1998.
90. J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *ACM Journal*, 1999.
91. M. Konrath, T. Gottron, S. Staab, and A. Scherp. SchemEX - Efficient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. *Journal of Web Semantics*, 16, 2012.
92. D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven Evaluation of Linked Data Quality. In *23rd International Conference on World Wide Web (WWW'14)*, 2014.
93. D. Kontokostas, A. Zaveri, S. Auer, and J. Lehmann. TripleCheckMate: A Tool for Crowdsourcing the Quality Assessment of Linked Data. *4th Conference on Knowledge Engineering and Semantic Web*, 2013.
94. Kovács-Láng. Global Terrestrial Observing System. Technical report, GTOS Central and Eastern European Terrestrial Data Management and Accessibility Workshop, 2000.
95. C. J. Kowalski. On the Effects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coefficient. *Journal of the Royal Statistical Society*, 1972.
96. S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain. Automatic Domain Identification for Linked Open Data. In *IEEE/WIC/ACM International Joint Conferences on*

Web Intelligence (WI) and Intelligent Agent Technologies (IAT), pages 205–212, 2013.
97. A. Langegger and W. Woss. RDFStats - An Extensible RDF Statistics Generator and Library. In *20th International Workshop on Database and Expert Systems Application (DEXA)*, pages 79–83, 2009.
98. O. Lassila and R. R. Swick. Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation, 1999. `http://www.w3.org/TR/1999/REC-rdf-syntax-19990222`.
99. S. LaValle, E. Lesser, R. Shockley, M. S. Hopkins, and N. Kruschwitz. Big Data, Analytics and the Path From Insights to Value. *MIT Sloan Management Review*, 2011.
100. T. Lebo, S. Sahoo, and D. McGuinness. PROV-O: The PROV Ontology. W3C Recommendation, 2013. `http://www.w3.org/TR/prov-o`.
101. J. Lehmann and S. Sonnenburg. DL-Learner: Learning Concepts in Description Logics. *Journal of Machine Learning Research*, 2009.
102. M. Lenzerini. Data Integration: A Theoretical Perspective. In *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pages 233–246, 2002.
103. M. Lenzerini. Data Integration: A Theoretical Perspective. In *21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, 2002.
104. J. Leskovec and C. Faloutsos. Sampling from Large Graphs. In *12thth ACM International Conference on Knowledge Discovery and Data Mining (KDD'12)*, 2006.
105. H. Li. Data Profiling for Semantic Web Data. In *International Conference on Web Information Systems and Mining (WISM)*, pages 472–479, 2012.
106. G. Limaye, S. Sarawagi, and S. Chakrabarti. Annotating and Searching Web Tables Using Entities, Types and Relationships. *VLDB Endowment*, pages 1338–1347, 2010.
107. B. F. Lóscio, E. G. Stephan, and S. Purohit. Dataset Usage Vocabulary (DUV). W3C First Public Working Draft, 2015. `http://www.w3.org/TR/vocab-duv/`.
108. J. J. M. and A. B. Godfrey. *Juran's quality handbook*. McGraw Hill, 1999.
109. F. Maali and J. Erickson. Data Catalog Vocabulary (DCAT). W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-dcat/`.
110. C. Mader, B. Haslhofer, and A. Isaac. Finding quality issues in SKOS vocabularies. *Theory and Practice of Digital Libraries*, 2012.
111. E. Mäkelä. Aether - Generating and Viewing Extended VoID Statistical Descriptions of RDF Datasets. In *11th European Semantic Web Conference (ESWC), Demo Track*, Heraklion, Greece, 2014.
112. J. Manyika and A. D. Elizabeth. Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey Business Technology Office, 2013.
113. N. Marie, F. Gandon, M. Ribière, and F. Rodio. Discovery Hub: On-the-fly Linked Data Exploratory Search. In *The 9th International Conference on Semantic Systems*, 2013.
114. P. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In *2012 Joint EDBT/ICDT Workshops*, 2012.
115. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding Light on the Web of Documents. In *7th International Conference on Semantic Systems*, 2011.

116. N. Mihindukulasooriya, R. Garcia-Castro, and M. E. Gutiérrez. Linked Data Platform as a novel approach for Enterprise Application Integration. In *4th International Workshop on Consuming Linked Data (COLD'13)*, 2013.
117. P. Mika. *Social Networks and the Semantic Web*, volume 5 of *Semantic Web and Beyond*. Springer, 2007.
118. A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference. W3C Recommendation, 2009. `http://www.w3.org/TR/skos-reference/`.
119. R. J. Miller and P. Andritsos. Schema Discovery. *IEEE Data Engineering Bulletin*, 26:40–45, 2003.
120. D. Nebert. Developing Spatial Data Infrastructures: The SDI Cookbook, 2004. `http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf`.
121. A. Nikolov, M. d'Aquin, and E. Motta. What Should I Link to? Identifying Relevant Sources and Classes for Data Linking. In *Joint International Semantic Technology Conference (JIST)*, 2011.
122. T. D. Noia, R. Mirizzi, V. O. Claudio, D. Romito, and M. Zanker. Linked Open Data to Support Content-based Recommender Systems. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.
123. L. Page, S. Brin, M. Rajeev, and W. Terry. The PageRank Citation Ranking: Bringing Order to the Web, 1998.
124. M. Pennacchiotti and S. Gurumurthy. Investigating topic models for social media user recommendation. In *20th International Conference on World Wide Web (WWW'11)*, 2011.
125. E. Peukert, J. Eberius, and E. Rahm. AMC - A framework for modelling and comparing matching systems as matching processes. In *IEEE 27th International Conference on Data Engineering (ICDE'11)*, 2011.
126. E. Peukert, J. Eberius, and E. Rahm. A Self-Configuring Schema Matching System. In *IEEE 28th International Conference on Data Engineering (ICDE'12)*, 2012.
127. A. Phil and S. Gofran. Asset Description Metadata Schema (ADMS). W3C Working Group Note, 2013. `http://www.w3.org/TR/vocab-adms`.
128. E. Pietriga, C. Bizer, D. Karger, and R. Lee. Fresnel: A Browser-Independent Presentation Vocabulary for RDF. In *5th International Semantic Web Conference (ISWC'06)*, pages 158–171, 2006.
129. J. Porter. *Designing for the Social Web*. New Riders, 2008.
130. M. Poveda-Villalón, M. C. Suárez-Figueroa, and A. Gómez-Pérez. Validating Ontologies with OOPS! In *18th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, 2012.
131. D. Preotiuc-Pietro, S. Samangooei, T. Cohn, N. Gibbins, and M. Niranjan. Trendminer: An architecture for real time analysis of social media text. In *6th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2012.
132. N. Press. Understanding Metadata. Technical report, National Information Standards Organization, 2004.
133. E. Prud'hommeaux and A. Seaborne. SPARQL Query Language for RDF. W3C Recommendation, 2008. `http://www.w3.org/TR/rdf-sparql-query/`.
134. D. Reynolds. The Organization Ontology. W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-org`.
135. E. Ruckhaus, O. Baldizan, and M.-E. Vidal. Analyzing Linked Data Quality with LiQuate. In *11th European Semantic Web Conference (ESWC)*, 2014.
136. O. P. Rud. *Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy*. John Wiley & Sons, 2009.

137. M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th International Semantic Web Conference (ISWC)*, 2014.
138. C. Semantics. RDF-101. `http://www.cambridgesemantics.com/semantic-university/rdf-101`. Accessed: 2013-09-07.
139. B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996.
140. E. Sirin, M. Smith, and E. Wallace. Opening, Closing Worlds - On Integrity Constraints. In *5th OWLED Workshop on OWL: Experiences and Directions*, 2008.
141. D. Soergel. Thesauri and ontologies in digital libraries. In *2nd ACM/IEEE-CS Joint Conference on Digital Libraries*, 2002.
142. T. Steiner and S. Mirea. SEKI@home or Crowdsourcing an Open Knowledge Graph. In *1st International Workshop on Knowledge Extraction & Consolidation from Social Media (KECSM'12)*, Boston, USA, 2012.
143. U. Straccia and R. Troncy. oMAP: Combining Classifiers for Aligning Automatically OWL Ontologies. In *6th International Conference on Web Information Systems Engineering*, 2005.
144. F. Suchanek, G. Kasneci, and G. Weikum. Yago: A Core of Semantic Knowledge. In *16th International World Wide Web Conference (WWW'07)*, 2007.
145. O. Suominen and E. Hyvönen. Improving the quality of SKOS vocabularies with skosify. In *The 18th International Conference on Knowledge Engineering and Knowledge Management*, 2012.
146. O. Suominen and C. Mader. Assessing and Improving the Quality of SKOS Vocabularies. *Journal on Data Semantics*, 2013.
147. Z. Syed, T. Finin, V. Mulwad, and A. Joshi. Exploiting a Web of Semantic Data for Interpreting Tables. In *2nd Web Science Conference*, 2010.
148. J. Tao, L. Ding, and D. McGuinness. Instance Data Evaluation for Semantic Web-Based Knowledge Management Systems. In *42nd Hawaii International Conference on System Sciences, HICSS'09*, pages 1–10, 2009.
149. A. Tonon, M. Catasta, G. Demartini, P. Cudré-Mauroux, and K. Aberer. TRank: Ranking Entity Types Using the Web of Data. In *12th International Semantic Web Conference (ISWC)*, 2013.
150. N. Toupikov, J. Umbrich, and R. Delbru. DING! Dataset ranking using formal descriptions. In *2nd International Workshop on Linked Data on the Web (LDOW)*, 2009.
151. G. Tummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker. Sig.ma: Live views on the Web of data. *Journal of Web Semantics*, 8(4), 2010.
152. G. Tummarello, R. Delbru, and E. Oren. Sindice.com: Weaving the open linked data. In *6th International Semantic Web Conference (ISWC)*, 2007.
153. R. Usbeck, M. Röder, A.-C. Ngonga-Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL - General Entity Annotation Benchmark Framework. In *24th International World Wide Web Conference (WWW'15)*, 2015.
154. Z. Valentina and C. L. Social ranking: uncovering relevant content using tag-based recommender systems. In *2nd ACM conference on Recommender systems - RecSys*, 2008.

155. M. Verlic. LODGrefine - LOD-enabled Google Refine in Action. In *8th International Conference on Semantic Systems - I-SEMANTICS '12*, 2012.
156. G. Vickery. Review of Recent Studies on PSI-use and Related Market Developments. Technical report, EC DG Information Society, 2011.
157. H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hübner. Ontology-Based Integration of Information - A Survey of Existing Approaches. In *IJCAI Workshop: Ontologies and Information*, pages 108–117, 2001.
158. J. Wang and F. Lochovsky. Data Extraction and Label Assignment for Web Databases. In *12th International World Wide Web Conference (WWW'03)*, pages 187–196, 2003.
159. R. Y. Wang and D. M. Strong. Beyond Accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 1996.
160. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment Methodologies for Linked Open Data. *Semantic Web Journal*, 2012.