ANALYSIS CERTIFICATE

Magister
by compilatio.net

Account : **Bernard Merialdo**

ID : **g3xu5sj1**

Title : **Thesis ahmad assaf-rapporteurs.pdf**

Folder : **Ahmad Assaf**

Comments : *Not available*

uploaded on the :10/07/2015 4:16 PM

Similarity document :

🟢 2%

Similarities section 5 :

🟢 3%

DETAILED INFORMATION

Title : Thesis Ahmad Assaf-Rapporteurs.pdf

Description : Ahmad Assaf

Analysed on : 10/07/2015 4:38 PM

Login ID : d2a4lrj3

uploaded on the : 10/07/2015 4:16 PM

Upload type : manual submission

File name : Thesis Ahmad Assaf-Rapporteurs.pdf

File type : pdf

Word count : 15438

Character count : 72510

TOP PROBABLE SOURCES - *AMONG 28 PROBABLE SOURCES*

1. **link.springer.com**/.../10.1007/978-3-319-10587-1_1 🚩 1%

2. **jens-lehmann.org**/.../files/cv.pdf 🚩 <1%

SIMILARITIES FOUND IN THIS DOCUMENT/SECTION

Matching similarities : **2 %** ℹ️

Assumed similarities : **<1 %** ℹ️

Accidental similarities : **<1 %** ℹ️

Highly probable sources - 28

Less probable sources - 45

Accidental sources- 26 Sources

Ignored sources - 4

HIGHLY PROBABLE SOURCES

	28 Sources	Similarity
1.	**link.springer.com**/.../10.1007/978-3-319-10587-1_1	1%
2.	**citeseerx.ist.psu.edu**/.../viewdoc/summary	1%
3.	**jens-lehmann.org**/.../files/cv.pdf	<1%
4.	**dblp.uni-trier.de**/.../b/Bizer:Christian	<1%
5.	Source Compilatio.net aclt47	<1%
6.	**www.w3.org**/.../TR	<1%
7.	**daselab.cs.wright.edu**/.../pub/domainIdentLOD13.pdf	<1%
8.	Source Compilatio.net alwxz4	<1%
9.	Source Compilatio.net jmqsu2	<1%
10.	**www.semantic-web-journal.net**/.../files/swj499.pdf	<1%
11.	Source Compilatio.net hqrz2	<1%
12.	Source Compilatio.net 72kmq38x	<1%
13.	**151.1.219.218**/.../b43d3f37-bd5d-4144...9-b27a0ca3d1d5.pdf	<1%
14.	Source Compilatio.net dkpqz9	<1%
15.	Source Compilatio.net eikmrv	<1%
16.	Source Compilatio.net ru179	<1%
17.	Source Compilatio.net lz928	<1%
18.	Source Compilatio.net 7e89z	<1%
19.	Source Compilatio.net iptvx	<1%
20.	**ebiquity.umbc.edu**/.../publication/478.pdf	<1%
21.	Source Compilatio.net bchqt1	<1%
22.	Source Compilatio.net 695ct	<1%
23.	Source Compilatio.net fpqst2	<1%
24.	**www.researchgate.net**/.../Martin_Bruemmer/citations	<1%
25.	Source Compilatio.net cdsy1	<1%
26.	Source Compilatio.net ptx14	<1%
27.	Source Compilatio.net iklr8	<1%
28.	Source Compilatio.net jpruv	<1%

LESS PROBABLE SOURCES

	45 Sources	Similarity
1.	**link.springer.com**/.../10.1007/978-3-642-39784-4_1	<1%
2.	**link.springer.com**/.../10.1007/978-3-642-29923-0_19	<1%
3.	Source Compilatio.net dglqw7	<1%
4.	**www.w3.org**/.../TR/vocab-dcat	<1%
5.	Source Compilatio.net btwyz8	<1%
6.	**www.w3.org**/.../TR/void	<1%

ACCIDENTAL SOURCES

26 Sources   *Similarity*

1. Document: bhlmp5 - belongs to another user   <1%

2. Source Compilatio.net cdqz13   <1%

3. Source Compilatio.net fiu47   <1%

4. **research.google.com**/.../pubs/author8038.html   <1%

5. Source Compilatio.net irw68   <1%

6. Source Compilatio.net afy57   <1%

7. **dblp.uni-trier.de**/.../n/Nejdl:Wolfgang   <1%

8. Source Compilatio.net inuwx   <1%

9. Source Compilatio.net klnwy1   <1%

10. **www.w3.org**/.../LinkingOpenData/DataSets   <1%

11. Source Compilatio.net glmu36   <1%

12. Source Compilatio.net gsvz6   <1%

13. Source Compilatio.net isz78   <1%

14. Source Compilatio.net vxt9y   <1%

15. **research.google.com**/.../pubs/PrabhakarRaghavan.html   <1%

16. Source Compilatio.net p9j6w   <1%

17. **urq.deri.ie**/.../   <1%

18. Source Compilatio.net ahisz   <1%

19. Source Compilatio.net eikqu   <1%

20. Source Compilatio.net i8lh19dt   <1%

21. Source Compilatio.net dnqu1   <1%

22. **www.researchgate.net**/.../links/0912f508e4bcc8c30e000000.pdf   <1%

23. Source Compilatio.net efkt17   <1%

24. **www.wiley.com**/.../WileyTitle/productCd-0471255475.html   <1%

25. Source Compilatio.net hilq39   <1%

26. Source Compilatio.net cmpqv5   <1%

IGNORED SOURCES

4 Sources   *Similarity*

1. **github.com**/.../ahmadassaf/opendata-checker   2%

2. **github.com**/.../ahmadassaf/kbe   2%

3. **github.com**/.../ahmadassaf/KBE   2%

4. **www.eurecom.fr**/.../~troncy/Publications   <1%

SIMILARITIES FOUND IN THIS DOCUMENT/SECTION

TEXT EXTRACTED FROM THE DOCUMENT

[11] Ahmad Assaf, Rapha¨l Troncy, and Aline Senart. Roomba: An Extensible e Framework to Validate and Build Dataset Proles. In 12th European Semantic Web Conference (ESWC), Portoroz, Slovenia, 2015. [12] Ahmad Assaf, Rapha¨l Troncy, and Aline Senart. Whats up LOD Cloud? e Observing The State of Linked Open Data Cloud Metadata. In 12th European Semantic Web Conference (ESWC), Portoroz, Slovenia, 2015. [13] S¨ren Auer, Jan Demter, Michael Martin, and Jens Lehmann.

LODStats - an o 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] Mike Bergman. Deconstructing the Google Knowledge Graph. http://www.mkbergman.com/1009/deconstructing-the-googleknowledge-graph. [15] Tim Berners-Lee. Uniform Resource Identier (URI): Generic Syntax. RFC 3986, 2005. http://tools.ietf.org/html/rfc3986. [16] Tim Berners-Lee. Linked Data - Design Issues. W3C Personal Notes, 2006. http://www.w3.org/DesignIssues/LinkedData. [17] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. Scientic American, 284(5):34–43, 2001. [18] Dimitris Berrueta, Sergio Fern"andez, and Iv"an Frade.

Cooking HTTP content negotiation with Vapour.

In 4th Workshop on Scripting for the Semantic Web

(SFSW'08), 2008. [19] Gasser Les Stvilia Besiki, , Michael B. Twidale, and Linda C. Smith. A framework for information quality assessment.

Journal of the American Society for Information Science and

Technology, 2007. [20] Christian Bizer.

Evolving the Web into a Global Data Space. In 28th British National Conference on Advances in Databases, 2011. [21] Christian Bizer and Tom Heathand Tim Berners-Lee. Linked Data - The Story So Far.

International Journal on Semantic Web and Information Systems (IJSWIS), 2009. [22] Christian Bizer and Richard Cyganiak. Quality-driven information ltering using the wiqa policy framework. Jorunal of Web Semantics, 7(1), 2009.

Bibliography

131

[23] Christian Bizer, Jens Lehmann, Georgi Kobilarov, S¨ren Auer, Christian o Becker, Richard Cyganiak, and Sebastian Hellmann.

DBpedia - A Crystallization Point for the Web of

Data. Journal of Web Semantics, 7(3), 2009. [24] Christoph B¨hm, Gjergji Kasneci, and Felix Naumann. Latent Topics in Grapho structured Data. In 21st ACM International Conference on Information and Knowledge Management (CIKM), pages 2663–2666, Maui, Hawaii, USA, 2012. [25] Christoph B¨hm, Johannes Lorey, and Felix Naumann. Creating voiD Descripo tions for Web-scale Data. Journal of Web Semantics, 9(3):339–345, 2011. [26] Christoph Bohm, Felix Naumann, ZZiawasch Abedjan, Dandy Fenz, Toni Grutze, Daniel Hefenbrock, Matthias Pohl, and David Sonnabend. Proling linked open data with ProLOD. In 26th International Conference on Data Engineering Workshops (ICDEW), 2010. [27] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor.

Freebase: A Collaboratively Created Graph Database for Structuring Human

Knowledge. In ACM International Conference on Management of Data (SIGMOD), 2008. [28] Danah Boyd and Kate Crawford. Six Provocations for Big Data. Social Science Research Network Working Paper Series, 2011. [29] Dan Brickley and R.V. Guha. RDF Schema 1.1. W3C Recommendation, 2014. http://www.w3.org/TR/rdf-schema. [30] Sergey Brin and Lawrence Page.

The anatomy of a large-scale hypertextual Web search engine.

In 7th International Conference on World Wide Web (WWW'98), 1998. [31] Martin Br¨mmer, Ciro Baron, Ivan Ermilov, Markus Freudenberg, Dimitris u Kontokostas, and Sebastian Hellmann.

DataID: Towards Semantically Rich Metadata for Complex Datasets.

In 10th International Conference on Semantic Systems, 2014. [32] Carlos Buil-Aranda and Aidan Hogan.

SPARQL Web-Querying Infrastructure: Ready for Action?

In 12th International Semantic Web Conference (ISWC), 2013. [33] Soumen Chakrabarti, Byron E.

Dom, Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins,

David Gibson, and Jon Kleinberg. Mining the web's link structure. Computer, 1999.

132

Bibliography

[34] Didier Cherix, Ricardo Usbeck, Andreas Both, and Jens Lehmann. CROCUS: Cluster-based ontology data cleansing. In 2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice, 2014. [35] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita.

A Framework for Benchmarking Entity-annotation Systems.

In 22nd International World Wide Web Conference (WWW'13), 2013. [36] Richard Cyganiak, Holger Stenzhorn, Renaud Delbru, Stefan Decker, and Giovanni Tummarello. Semantic Sitemaps: Ecient and Flexible Access to Datasets on the Semantic Web. In 5th European Semantic Web Conference (ESWC), pages 690–704, Tenerife, Spain, 2008. [37] Richard Cyganiak, Jun Zhao, Michael Hausenblas, and Keith Alexander.

Describing Linked Datasets with the VoID Vocabulary.

W3C Note, 2011. http://www.w3.org/TR/void/. [38] Altigran Soares da Silva, Denilson Barbosa, Jo˜o M. B. Cavalcanti, and Marco a A. S. Sevalho. Labeling Data Extracted from the Web. In On The Move Confederated International Conferences, pages 1099–1116, 2007. [39] Mathieu d'Aquin and Enrico Motta.

Watson, More Than a Semantic Web Search Engine.

Semantic Web Journal, 2011. [40] Tim Davies, Raed Sharif, and Jose Alonso. Open Data Barometer - Global Report. Technical report, World Wide Web Foundation, 2015. http: //barometer.opendataresearch.org/. [41] Davi de Castro Reis, Paulo B. Golgher, Altigran S. da Silva, and Alberto 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] Jeremy Debattista, Makx Dekkers, and Deirdre Lee. Data Quality Vocabulary (DQV). W3C First Public Working Draft, 2015. http://www.w3.org/TR/vocab-dqv/. [43] Jeremy Debattista, Christoph Lange, and S¨ren Auer. daQ, an Ontology for o th International Workshop on Linked Data Dataset Quality Information. In 7 on the Web (LDOW), 2014. [44] Jeremy Debattista, Santiago Londo˜o, Christoph Lange, and S¨ren Auer. n o LUZZU - A framework for linked data quality assessment. CoRR, abs/1412.3750, 2014.

Bibliography

133

[45] Renaud Delbru, Nickolai Toupikov, and Michele Catasta.

Hierarchical link analysis for ranking web data.

In 7th European Semantic Web Conference (ESWC), 2010. [46] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, and R.Scott Cost. Swoogle: A semantic web search and metadata engine. In 13st ACM International Conference on Information and Knowledge Management (CIKM), 2004. [47] Diaz-Aviles Ernesto, Drumond Lucas, Schmidt-Thieme Lars, and Nejdl Wolfgang.

Real-time top-n recommendation in social streams.

In 6th ACM conference on Recommender systems - RecSys, 2012. [48] Bakshy Eytan, Rosenn Itamar, Marlow Cameron, and Adamic Lada. The role of social networks in information diusion. In 21th International Conference on World Wide Web (WWW'12), 2012. [49] Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. A Scalable Approach for Efciently Generating Structured Dataset Topic Proles. In 11th European Semantic Web Conference (ESWC), 2014. [50] Tim Finin, Zareen Syed, James Mayeld, Paul Mcnamee, and Christine Piatko.

Using Wikitology for Cross-Document Entity Coreference Resolution.

In AAAI Spring Symposium on Learning, 2009. [51] Annika Flemming.

Quality Characteristics of Linked Data Publishing Datasources.

Master's thesis, Humboldt-Universitt zu Berlin, 2010. [52] Giorgos Flouris, Yannis Roussakis, and M Poveda-Villal´n. Using provenance o for quality assessment and repair in linked open data. In 2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn'12), 2012. [53] Benedikt Forchhammer, Anja Jentzsch, and Felix Naumann. LODOP - MultiQuery Optimization for Linked Data Proling Queries. In International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES), Heraklion, Greece, 2014. [54] Philipp Frischmuth, S¨ren Auer, Sebastian Tramp, J¨rg Unbehauen, Kai o o Holzweißig, and Carl-Martin 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.

Bibliography

[55] Philipp Frischmuth, Jakub Kl´mek, S¨ren Auer, Sebastian Tramp, J¨rg Unbeo o hauen, Kai Holzweißig, and Carl-Martin Marquardt. Linked Data in Enterprise Information Integration. Semantic Web Journal, 2012. [56] Matias Frosterus, Eero Hyv¨nen, and Joonas Laitio. Creating and Publishing o Semantic Metadata about Linked and Open Datasets. In Linking Government Data. Springer, 2011. [57] Matias Frosterus, Eero Hyv¨nen, and Joonas Laitio. DataFinland - A Semantic o Portal for Open and Linked Datasets. In 8th Extended Semantic Web Conference (ESWC), pages 243–254, 2011. [58] Christian F¨rber and Martin Hepp. SWIQA - A Semantic Web information u quality assessment framework. In 19th European Conference on Information Systems (ECIS'11), 2011. [59] W3C OWL Working Group. OWL 2 Web Ontology Language. W3C Recommendation, 2012. http://www.w3.org/TR/owl2-overview. [60] Thomas R. Gruber.

A Translation Approach to Portable Ontology Specications.

Knowledge Acquisition, 5(2), 1993. [61] Christophe Gu´ret, Paul Groth, Claus Stadler, and Jens Lehmann.

Assessing e Linked Data Mappings Using Network Measures.

In 9th European Semantic Web Conference (ESWC), 2012. [62] Richard Hammell, Carl Bates, Harvey Lewis, Costi Perricos, Louise Brett, and David 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-datadriving-growth-ingenuity-and-innovation.pdf. [63] Patricia Harpring. Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works. Getty Research Institute, 2010. [64] Andreas Harth, Katja Hose, Marcel Karnstedt, Axel Polleres, Kai-Uwe Sattler, and J¨rgen Umbrich. Data Summaries for On-demand Queries over Linked u Data. In 19th World Wide Web Conference (WWW'10), 2010. [65] Andreas Harth, Sheila Kinsella, and Stefan Decker.

Using naming authority to rank data and ontologies for

web search. In 8th International Semantic Web Conference (ISWC), 2009. [66] Olaf Hartig and Jun Zhao.

Using web data provenance for quality assessment.

In 8th International Semantic Web Conference (ISWC), 2009.

Bibliography

135

[67] Bernhard Haslhofer and Niko Popitsch. DSNotify: Detecting and Fixing Broken Links in Linked Data Sets. In 8th International Workshop on Web Semantics, 2009. [68] Oktie Hassanzadeh, Duan, Achille Fokoue, Anastasios Kementsietsidis, Kavitha Srinivas, and Michael J. Ward. Helix: Online Enterprise Data Analytics. In 20th World Wide Web Conference (WWW'11), pages 225–228, 2011. [69] Michael Hausenblas, Wolfgang Halb, Yves Raimond, Lee Feigenbaum, and Danny Ayer.

SCOVO: Using Statistics on the Web of Data.

In 6th European Semantic Web Conference on The Semantic Web (ESWC), 2009. [70] Aidan Hogan, Andreas Harth, and Stefan Decker. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In 2nd Workshop on Scalable Semantic Web Knowledge Base Systems, 2006. [71] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. Weaving the pedantic web. In 3rd International Workshop on Linked Data on the Web (LDOW), 2010. [72] Aidan Hogan, J¨Rgen Umbrich, Andreas Harth, Richard Cyganiak, Axel u Polleres, and Stefan Decker.

An empirical survey of Linked Data conformance.

Journal of Web Semantics, 2012. [73] Renato Iannella and James McKinney.

vCard Ontology - for describing People and Organizations.

W3C Interest Group Note, 2014. http://www.w3.org/ TR/vcard-rdf. [74] Halo Business Inteligence. Data Quality - Why you should care about the cleanliness of your data. Technical report, Halo, 2013. [75] Antoine Isaac and Ed Summers.

SKOS Simple Knowledge Organization System Primer.

W3C Working Group Note, 2009. [76] Robert Isele, J¨rgen Umbrich, Christian Bizer, and Andreas Harth. LDspiu der: An Open-source Crawling Framework for the Web of Linked Data. In 9th International Semantic Web Conference (ISWC), Posters & Demos Track, 2010. [77] Cantador Iv´n and Bellog´ Alejandro. Semantic contextualisation of social a n tag-based proles and item recommendations. In 12th Internationl Conference on E-Commerce and Web Technolgoies, 2011. [78] Prateek Jain, Pascal Hitzler, Krzysztof Janowicz, and Chitra Venkatramani. There's No Money in Linked Data, 2013. http://knoesis.wright.edu/ faculty/pascal/pub/nomoneylod.pdf.

136

Bibliography

[79] Anja Jentzsch. Proling the Web of Data. In 13th International Semantic Web Conference (ISWC), Doctoral Consortium, Trentino, Italy, 2014. [80] Tobias K¨fer, Ahmed Abdelrahman, J¨rgen Umbrich, Patrick O'Byrne, and a u Aidan Hogan. Observing Linked Data Dynamics. In 10th European Semantic Web Conference (ESWC), 2013. [81] Beverly K. Kahn,

Diane M. Strong, and Richard Y. Wang. Information quality benchmarks: product and service performance. Communications of the ACM, 2002. [82] C. Maria Keet, Mar´ del Carmen Su´rez-Figueroa, and Mar´ Poveda-Villal´n. a a a o The Current Landscape of Pitfalls in Ontologies. In International Conference on Knowledge Engineering and Ontology Development (KEOD), 2013. [83] Shahan Khatchadourian and Mariano 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. [84] Ralph Kimball, Laura Reeves, Warren Thornthwaite, Margy Ross, and Warren 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. [85] Jon M. Kleinberg.

Authoritative sources in a hyperlinked environment.

ACM Journal, 1999. [86] Mathias Konrath, Thomas Gottron, Steen Staab, and Ansgar Scherp. SchemEX - Ecient Construction of a Data Catalogue by Stream-based Indexing of Linked Data. Journal of Web Semantics, 16, 2012. [87] Dimitris Kontokostas, Patrick Westphal, S¨ren Auer, Sebastian Hellmann, Jens o Lehmann, Roland Cornelissen, and Amrapali Zaveri.

of Linked Data. 4th Conference on Knowledge Engineering and Semantic Web, 2013. [89] Kov´cs-L´ng. Global Terrestrial Observing System. Technical report, GTOS a a Central and Eastern European Terrestrial Data Management and Accessibility Workshop, 2000.

Bibliography

137

[90] Charles J. Kowalski. On the Eects of Non-Normality on the Distribution of the Sample Product-Moment Correlation Coecient. Journal of the Royal Statistical Society, 1972. [91] Sarasi Lalithsena, Prateek Hitzler, Amit Sheth, and P. Jain.

Automatic Domain Identication for Linked Open Data.

In IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), pages 205–212, 2013. [92] Andreas Langegger and Wolfram Woss.

RDFStats - An Extensible RDF Statistics Generator and Library.

In 20th International Workshop on Database and Expert Systems Application (DEXA), pages 79–83, 2009. [93] Ora Lassila and Ralph R. Swick. Resource Description Framework (RDF) Model and Syntax Specication. W3C Recommendation, 1999. http://www. w3.org/TR/1999/REC-rdf-syntax-19990222. [94] Steve LaValle, Eric Lesser, Rebecca Shockley, Michael S. Hopkins, and Nina Kruschwitz. Big Data, Analytics and the Path From Insights to Value. MIT Sloan Management Review, 2011. [95] Timothy Lebo, Satya Sahoo, and Deborah McGuinness. PROV-O: The PROV Ontology. W3C Recommendation, 2013. http://www.w3.org/TR/prov-o. [96] Jens Lehmann and Soeren Sonnenburg.

DL-Learner: Learning Concepts in Description Logics.

Journal of Machine Learning Research, 2009. [97] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pages 233–246, 2002. [98] Maurizio Lenzerini. Data Integration: A Theoretical Perspective. In 21st ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, 2002. [99] Jure Leskovec and Christos Faloutsos. Sampling from Large Graphs. In 12th th ACM International Conference on Knowledge Discovery and Data Mining (KDD'12), 2006. [100] Huiying Li. Data Proling for Semantic Web Data. In International Conference on Web Information Systems and Mining (WISM), pages 472–479, 2012. [101] Girija Limaye, Sunita Sarawagi, and Soumen Chakrabarti.

Annotating and Searching Web Tables Using Entities, Types and

Relationships. VLDB Endowment, pages 1338–1347, 2010.

138

Bibliography

[102] Bernadette Farias L´scio, Eric G. Stephan, and Sumit Purohit. Dataset Usage o Vocabulary (DUV). W3C First Public Working Draft, 2015. http://www. w3.org/TR/vocab-duv/. [103] Joseph Juran. M. and A. Blanton Godfrey. Juran's quality handbook. McGraw Hill, 1999. [104] Fadi Maali and John Erickson. Data Catalog Vocabulary (DCAT). W3C Recommendation, 2014. http://www.w3.org/TR/vocab-dcat/. [105] Christian Mader, Bernhard Haslhofer, and Antoine Isaac.

Finding quality issues in SKOS vocabularies.

Theory and Practice of Digital Libraries, 2012. [106] Eetu M¨kel¨. Aether - Generating and Viewing Extended VoID Statistical a a Descriptions of RDF Datasets. In 11th European Semantic Web Conference (ESWC), Demo Track, Heraklion, Greece, 2014. [107] James Manyika and Almasi Doshi Elizabeth. Open data: Unlocking innovation and performance with liquid information. Technical report, McKinsey Business Technology Oce, 2013. [108] Nicolas Marie, Fabien Gandon, Myriam Ribi`re, and Florentin Rodio. Discove ery Hub: On-the-y Linked Data Exploratory Search. In The 9th International Conference on Semantic Systems, 2013. [109] Pablo Mendes, Hannes M¨hleisen, and Christian Bizer. Sieve: linked data u quality assessment and fusion. In 2012 Joint EDBT/ICDT Workshops, 2012. [110] Pablo N. Mendes, Max Jakob, Andr´s Garc´ e a-Silva, and Christian Bizer.

DBpedia Spotlight: Shedding Light on the Web of Documents.

In 7th International Conference on Semantic Systems, 2011. [111] Nandana Mihindukulasooriya, Raul Garcia-Castro, and Miguel Esteban Guti´rrez. Linked Data Platform as a novel approach for Enterprise Applie cation Integration. In 4th International Workshop on Consuming Linked Data (COLD'13), 2013. [112] Peter Mika. Social Networks and the Semantic Web, volume 5 of Semantic Web and Beyond. Springer, 2007. [113] Alistair Miles and Sean Bechhofer.

SKOS Simple Knowledge Organization System Reference.

W3C Recommendation, 2009. http://www.w3.org/TR/ skos-reference/. [114] Ren´e J. Miller and Periklis Andritsos. Schema Discovery. IEEE Data Engie neering Bulletin, 26:40–45, 2003.

Bibliography

139

[115] Andriy Nikolov, Mathieu d'Aquin, and Enrico Motta. What Should I Link to?

Identifying Relevant Sources and Classes for Data Linking.

In Joint International Semantic Technology Conference (JIST), 2011. [116] Tommaso Di Noia, Roberto Mirizzi, Vito Ostuni Claudio, Davide Romito, and Markus Zanker. Linked Open Data to Support Content-based Recommender Systems. In 8th International Conference on Semantic Systems - ISEMANTICS '12, 2012. [117] Lawrence Page, Sergey Brin, Motwani Rajeev, and Winograd Terry. PageRank Citation Ranking: Bringing Order to the Web, 1998. The

[118] Eric Peukert, Julian Eberius, and Erhard Rahm. AMC - A framework for modelling and comparing matching systems as matching processes. In IEEE 27th International Conference on Data Engineering (ICDE'11), 2011. [119] Eric Peukert, Julian Eberius, and Erhard Rahm. A Self-Conguring Schema Matching System. In IEEE 28th International Conference on Data Engineering (ICDE'12), 2012. [120] Archer Phil and Shukair Gofran. Asset Description Metadata Schema (ADMS). W3C Working Group Note, 2013. http://www.w3.org/TR/vocab-adms. [121] Emmanuel Pietriga, Christian Bizer, David Karger, and Ryan Lee.

Fresnel: A Browser-Independent Presentation Vocabulary for RDF.

In 5th International Semantic Web Conference (ISWC'06), pages 158–171, 2006. [122] Joshua Porter. Designing for the Social Web. New Riders, 2008. [123] Mar´ Poveda-Villal´n, Mari Carmen Su´rez-Figueroa, and Asunci´n G´meza o a o o th International Conference on P´rez. Validating Ontologies with OOPS! In 18 e Knowledge Engineering and Knowledge Management (EKAW), 2012. [124] Daniel Preotiuc-Pietro, Sina Samangooei, Trevor Cohn, Nicholas Gibbins, and Mahesan Niranjan. Trendminer: An architecture for real time analysis of social media text. In 6th International AAAI Conference on Weblogs and Social Media (ICWSM), 2012. [125] NISO Press. Understanding Metadata. Technical report, National Information Standards Organization, 2004. [126] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language for RDF. W3C Recommendation, 2008. http://www.w3.org/TR/rdf-sparqlquery/.

140

Bibliography

[127] Dave Reynolds. The Organization Ontology. W3C Recommendation, 2014. http://www.w3.org/TR/vocab-org. [128] Edna Ruckhaus, Oriana Baldizan, and Maria-Esther Vidal. Analyzing Linked Data Quality with LiQuate. In 11th European Semantic Web Conference (ESWC), 2014. [129] Olivia Parr Rud. Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy. John Wiley & Sons, 2009. [130] Massimiliano Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the Linked Data Best Practices in Dierent Topical Domains. In 13th International Semantic Web Conference (ISWC), 2014. [131] Cambridge Semantics. RDF-101. http://www.cambridgesemantics. com/semantic-university/rdf-101. Accessed: 2013-09-07. [132] Evren Sirin, Michael Smith, and Evan Wallace. Opening, Closing Worlds - On Integrity Constraints. In 5th OWLED Workshop on OWL: Experiences and Directions, 2008. [133] Dagobert Soergel.

Thesauri and ontologies in digital libraries.

ACM/IEEE-CS Joint Conference on Digital Libraries, 2002. In 2nd

[134] Thomas Steiner and Stefan 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. [135] Umberto Straccia and Rapha¨l Troncy. oMAP: Combining Classiers for Aligne ing Automatically OWL Ontologies. In 6th International Conference on Web Information Systems Engineering, 2005. [136] Fabian Suchanek, Gjergji Kasneci, and Gerhard Weikum.

Yago: A Core of Semantic Knowledge.

In 16th International World Wide Web Conference (WWW'07), 2007. [137] Osma Suominen and Eero Hyv¨nen. Improving the quality of SKOS vocabuo th International Conference on Knowledge Engilaries with skosify. In The 18 neering and

Knowledge Management, 2012. [138] Osma Suominen and Christian Mader. Assessing and Improving the Quality of SKOS Vocabularies. Journal on Data Semantics, 2013. [139] Zareen Syed, Tim Finin, Varish Mulwad, and Anupam Joshi.

Exploiting a Web of Semantic Data for Interpreting Tables.

In 2nd Web Science Conference, 2010.

Bibliography

141

[140] Jiao Tao, Li Ding, and Deborah 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. [141] Nickolai Toupikov, J¨rgen Umbrich, and Renaud Delbru. DING! Dataset ranku ing using formal descriptions. In 2nd International Workshop on Linked Data on the Web (LDOW), 2009. [142] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, Szymon Danielczyk, Renaud Delbru, and Stefan Decker. Sig.ma:

Live views on the Web of data.

Journal of Web Semantics, 8(4), 2010. [143] Giovanni Tummarello, Renaud Delbru, and Eyal Oren. Sindice.com: Weaving the open linked data. In 6th International Semantic Web Conference (ISWC), 2007. [144] Ricardo Usbeck, Michael R¨der, Axel-Cyrille Ngonga-Ngomo, Ciro Baron, o Andreas Both, Martin Br¨mmer, Diego Ceccarelli, Marco Cornolti, Didier u Cherix, Bernd Eickmann, Paolo Ferragina, Christiane Lemke,

Andrea Moro, Roberto Navigli, Francesco Piccinno, Giuseppe Rizzo, Harald

Sack, Ren´ e Speck, Rapha¨l Troncy, J¨rg Waitelonis, and Lars Wesemann. GERBIL - Gene o eral Entity Annotation Benchmark Framework. In 24th International World Wide Web Conference (WWW'15), 2015. [145] Zanardi Valentina and Capra L. Social ranking: uncovering relevant content using tag-based recommender systems. In 2nd ACM conference on Recommender systems - RecSys, 2008. [146] Mateja Verlic. LODGrene - LOD-enabled Google Rene in Action. In 8th International Conference on Semantic Systems - I-SEMANTICS '12, 2012. [147] Graham Vickery. Review of Recent Studies on PSI-use and Related Market Developments. Technical report, EC DG Information Society, 2011. [148] Holger Wache, Thomas V¨gele, Ubbo Visser, Heiner Stuckenschmidt, Gerard o Schuster, Heiko Neumann, and Sebastian H¨bner. Ontology-Based Integrau tion of Information - A Survey of Existing Approaches. In IJCAI Workshop: Ontologies and Information, pages 108–117, 2001. [149] Jiying Wang and Frederick Lochovsky. Data Extraction and Label Assignment for Web Databases. In 12th International World Wide Web Conference (WWW'03), pages 187–196, 2003. [150] Richard Y. Wang and Diane M. Strong.

Beyond Accuracy: What data quality means to data consumers. Journal of Management Information Systems, 1996.

142

Bibliography

[151] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and S¨ren Auer. Quality Assessment Methodologies for Linked Open o Data. Semantic Web Journal, 2012.

Appendix A

Installation and Cutomization Instructions

A.1

Installation and cutomization instructions for Roomba

You can either download Roomba from the Github repository1 as a zip le or clone directly through git. Pay attention when cloning as there is a submodule dened and has to be cloned recursively as well. This can be done via: git clone --recursive http://github.com/ahmadassaf/opendata-checker If you have cloned without --recursive, you may nd out that some folders are empty. To x this: git submodule update --init After successfully having cloned Roomba to your local machine. • Install dependencies by running the command "npm install" • Start Roomba by running "node DC.js"

A.2

Customizing Roomba

There are a set of options that you can customize. They can be edited from options.json le (see Listing A.1)

{ locale : cacheFolderName : licensesFolder : mappingFileName : proxy : } en , /cache/ , /util/licenses/ , licenseMappings ,

Listing A.1: Roomba's customization via the options le

1

https://github.com/ahmadassaf/opendata-checker

144

Appendix A. Installation and Cutomization Instructions

• locale: the language of the messages and the prompts. Default: en • cacheFolderName: the name of cache folder separated by /. Default: /cache/ • licensesFolder: the location of the Open Licenses repository2 . It is dened as a submodule and by default it is located in /util/licenses/. • mappingFileName: the name of manual license mappings. censeMappings • proxy: proxy server e.g., proxy:8080. Default: "" Default: li-

A.2.1

Localization

If you wish to translate the messages and prompts into other languages than English, you have to create a new language entry in the util/messages.js with the new locale code e.g., fr. Afterwards, you should keep the object keys intact by translating the values into the desired language. See Listing A.2 for an example.

var messages = { en : { error : { unKnownError : An unknown Error occurred .. , . . fr : { error : { unkownError : Erreur inconnue .. } }

Listing A.2: Roomba's localization le example

A.3

Installation and cutomization instructions for Knowledge Graph Scraper

You can either download the knowledge graph scrapper from the Github repository3 as a zip le or clone directly through git. git clone http://github.com/ahmadassaf/kbe After successfully having cloned the repository to your local machine.

2 3

https://github.com/okfn/licenses https://github.com/ahmadassaf/kbe

A.3. Installation and cutomization instructions for Knowledge Graph Scraper 145

• Install dependencies by running the command "npm install" • Start Roomba by running "node KBE.js" The script will automatically create all the required Cache folders: • Main cache folder called cache in the root folder of the application. • GKB inside the cache folder which will hold the aggregated Google Knowledge boxes extracted for a DBpedia concept. • instances GKB inside the cache folder which will hold the Google Knowledge box for a single instance. • instances inside the cache folder which will hold the DBpedia instances for each concept. • instance properties inside the cache folder which will hold the distinct list of properties for all the instances of a certain concept. The application is run in the console and the output will be available in /cache/result.json.

A.3.1

Customization

There are a set of options that you can customize. They can be edited from options.json le (see Listing A.3)

{ cache_dbpedia_concepts : limit_dbpedia_concepts : limit_dbpedia_instances : limit_dbpedia_concepts_value : limit_dbpedia_instances_value: proxy : null } true, true, true, 10, 10,

Listing A.3: Roomba's localization le example

• cache dbpedia concepts: caches the concepts retrieved from DBpedia. • limit dbpedia concepts: limits the number of concepts retrieved by DBpedia, false will retrieve all the concepts. • limit dbpedia instances: limits the number of instances retrieved for each concept, false will retrieve all the instances.

146

Appendix A. Installation and Cutomization Instructions

• limit dbpedia concepts value: species the number of concepts that you wish to retrieve. • limit dbpedia instances value: species the number of instancesthat you wish to retrieve for each concept. • proxy: species the proxy address string containing ports e.g., proxy:8080. Default: "". Since Google changes the CSS selectors dynamically at random times, the user can always check the corresponding CSS class name selectors for the Google Knowledge Panel and edit them if needed in the same options.json le as shown in Listing A.4.

{ knowledgeBox knowledgeBox_disambiguate property property_value label description type images special_property special_property_value special_property_value_link } : : : : : : : : : : : : #kno-result , .kp-blk , ._Nl , .kno-fv , .kno-ecr-pt , .kno-rdesc , ._kx , .bicc , .kno-sh , ._Zh , a._dt

Listing A.4: Roomba's localization le example

Appendix B

Roomba Results

B.1

Roomba proling report for the LOD Cloud

Metadata Report [259] group information is missing. Check organization information as they can be mixed sometimes [133] maintainer field exists but there is no value defined [143] maintainer email field exists but there is no value defined [6] author field exists but there is no value defined [39] author email field exists but there is no value defined [156] version field exists but there is no value defined [44] The url defined for this dataset is not reachable [28] organization image url field exists but there is no value defined [34] notes field exists but there is no value defined [1] Tags information [Tags, Vocabularies] is missing [224] resources information (API endpoints, downloadable dumpds, etc.) is missing [11] author email is not a valid email address [6] maintainer email is not a valid email address [3] organization description field exists but there is no value defined [3] url field exists but there is no value defined [2] The organization image url defined for this dataset is not reachable Dataset Statistics There is a total of: 259 [missing] group fields 100.00% There is one [missing] tag field 0.39% There is a total of: 224 [missing] resources fields 86.49% There is a total of: 133 [undefined] maintainer fields 51.35% There is a total of: 143 [undefined] maintainer email fields 55.21% There is a total of: 6 [undefined] author fields 2.32% There is a total of: 39 [undefined] author email fields 15.06% There is a total of: 156 [undefined] version fields 60.23% There is a total of: 28 [undefined] organization image url fields 10.81% There is a total of: 34 [undefined] notes fields 13.13% There is a total of: 3 [undefined] organization description fields 1.16% There is a total of: 3 [undefined] url fields 1.16% Dataset Connectivity Issues

148

Appendix B. Roomba Results

There are 44 connectivity issues with the following URLs: http://id.loc.gov/authorities/ http://libris.kb.se http://www.londongazette.co.uk/mashup/gazettesdata.htm http://www4.wiwiss.fuberlin.de/eures/ http://www.linkedopenservices.org/services/geo/Spatial Resources/point/ICAO/ Tag Report [3220] vocabulary id field exists but there is no value defined Tag Statistics There is a total of: 3220 [undefined] vocabulary id fields 100.00% Total elements count: 3220 License Report [142] license url field is missing [136] We could not normalize the license information as no valid mapping was found [123] License information has been normalized [43] license title field exists but there is no value defined [43] license id field exists but there is no value defined License Statistics There is a total of: 141 [missing] license url fields 54.44% There is a total of: 43 [undefined] license title fields 16.60% There is a total of: 43 [undefined] license id fields 16.60% Total elements count: 259 Resource Report [1035] cache last updated field exists but there is no value defined [1024] webstore last updated field exists but there is no value defined [871] size field exists but there is no value defined [1020] hash field exists but there is no value defined [69] format field exists but there is no value defined [1024] mimetype inner field exists but there is no value defined [1035] url type field is missing [832] mimetype field exists but there is no value defined [1035] cache url field exists but there is no value defined [817] name field exists but there is no value defined

B.1. Roomba proling report for the LOD Cloud

149

[928] created field is missing [975] webstore url field exists but there is no value defined [853] last modified field exists but there is no value defined [488] resource type field exists but there is no value defined [335] The url for this resource is not reachable [98] description field exists but there is no value defined [52] The size for resource is not defined correctly [53] mimetype value defined where the resource is not reachable [44] size value defined where the resource is not reachable [20] The mimeType for resource is not defined correctly Resource Statistics There is a total of: 1035 [undefined] cache last updated fields 96.91% There is a total of: 1024 [undefined] webstore last updated fields 95.88% There is a total of: 871 [undefined] size fields 81.55% There is a total of: 1020 [undefined] hash fields 95.51% There is a total of: 69 [undefined] format fields 6.46% There is a total of: 1024 [undefined] mimetype inner fields 95.88% There is a total of: 832 [undefined] mimetype fields 77.90% There is a total of: 1035 [undefined] cache url fields 96.91% There is a total of: 817 [undefined] name fields 76.50% There is a total of: 975 [undefined] webstore url fields 91.29% There is a total of: 853 [undefined] last modified fields 79.87% There is a total of: 488 [undefined] resource type fields 45.69% There is a total of: 98 [undefined] description fields 9.18% There is a total of: 1034 [missing] url type fields 96.82% There is a total of: 927 [missing] created fields 86.80% Total elements count: 1068 Resource Connectivity Issues There are 330 connectivity issues with the following URLs: http://www.data.gov/semantic/data/alpha http://www.data.gov/catalog/raw http://www.data.gov/catalog/geodata http://lab3.libris.kb.se/sparql http://www.londongazette.co.uk/mashup/LondonGazetteData.zip http://www.londongazette.co.uk/issues/59535/notices/1196300 http://www.idref.fr/027182800/id ..... UnReachable URLs Types There There There There There There are: are: are: are: are: are: 212 unreachable URLs of type [null] 8 unreachable URLs of type [api] 112 unreachable URLs of type [file] 1 unreachable URLs of type [metadata] 1 unreachable URLs of type [example] 1 unreachable URLs of type [documentation]

150

Appendix B. Roomba Results

Listing B.1: The result of running Roomba on the LOD Cloud group. Note that some URLs are cut for display purposes

B.2

Roomba quality proling report for the LOD Cloud

D a t a s e t Q u a l i t y Report completeness quality Score : 50.22% a v a i l a b i l i t y quality Score : 23.32% l i c e n s i n g quality Score : 21.14% f r e s h n e s s quality Score : 79.49% c o r r e c t n e s s quality Score : 72.36% comprehensibility quality Score : 31.63% provenance q u a l i t y Score : 74.07% Average t o t a l q u a l i t y S c o r e : 50.32% Q u a l i t y I n d i c a t o r s Average E r r o r % Supports m u l t i p l e s e r i a l i z a t i o n s Has d i f f e r e n t data a c c e s s p o i n t s Uses d a t a s e t s d e s c r i p t i o n v o c a b u l a r i e s E x i s t e n c e o f d e s c r i p t i o n s about i t s s i z e E x i s t e n c e o f d e s c r i p t i o n s about i t s s t r u c t u r e (MIME Type , Format) E x i s t e n c e o f d e s c r i p t i o n s about i t s o r g a n i z a t i o n and c a t e g o r i z a t i o n E x i s t e n c e o f d e r e f e r e n c a b l e l i n k s f o r t h e d a t a s e t and i t s r e s o u r c e s E x i s t e n c e o f an RDF dump t h a t can be downloaded by u s e r s Existence o f queryable endpoints that respond to d i r e c t q u e r i e s E x i s t e n c e o f v a l i d d e r e f e r e n c a b l e URLs (r e s p o n d t o HTTP r e q u e s t) E x i s t e n c e o f human and machine r e a d a b l e l i c e n s e i n f o r m a t i o n Existence of d e r e f e r e n c a b l e l i n k s to the f u l l l i c e n s e information E x i s t e n c e o f timestamps t h a t can keep t r a c k o f i t s m o d i f i c a t i o n s I n c l u d e s t h e c o r r e c t MIME t y p e f o r t h e c o n t e n t I n c l u d e s the c o r r e c t s i z e f o r the content Absence o f S y n t a c t i c e r r o r s o n t h e l i n k s l e v e l E x i s t e n c e o f a t l e a s t o n e exemplary RDF f i l e E x i s t e n c e o f g e n e r a l i n f o r m a t i o n (title , URL, d e s c r i p t i o n) E x i s t e n c e o f m a i l i n g l i s t , message board o r p o i n t o f c o n t a c t E x i s t e n c e o f metadata t h a t d e s c r i b e s i t s a u t h o r i t a t i v e i n f o r m a t i o n Usage o f v e r s i o n i n g : 11.35% 19.31% 88.80% 86.30% 83.67% 8.33% 0.45% 91.51% 96.14% 42.41% 61.58% 96.14% 20.51% 48.92% 32.71% 1.28% 99.61% 5.89% 99.61% 21.75% 30.12%

Listing B.2: The result of running Roomba quality proler on the LOD Cloud group

Appendix C

DBpedia Ranked Properties in Fresnel Vocabulary

@prefix @prefix @prefix @prefix @prefix @prefix

ns1 : ns2 : rdf : rdfs xml : xsd :

<h t t p : / /www. w3 . o r g /2004/09/ f r e s n e l# . > <h t t p : / /www. w3 . o r g / n s / p r o v# . > <h t t p : / /www. w3 . o r g /1999/02/22 r d f s y n t a x ns# . > : <h t t p : / /www. w3 . o r g /2000/01/ r d f schema# . > <h t t p : / /www. w3 . o r g /XML/1998/ namespace> . <h t t p : / /www. w3 . o r g /2001/XMLSchema# . >

<h t t p : / / d a t a . eurecom . f r / i d / f r e s n e l / academicjournalGKPDefaultLens> a ns1 : Lens ; ns1 : c l a s s L e n s D o m a i n <h t t p : / / d b p e d i a . o r g / o n t o l o g y / AcademicJournal> ; ns1 : group <h t t p : / / d a t a . eurecom . f r / i d / f r e s n e l / group / a c a d e m i c j o u r n a l G r o u p> ; ns1 : p u r p o s e ns1 : d e f a u l t L e n s ; ns1 : s h o w P r o p e r t i e s <h t t p : / / d a t a . eurecom . f r / d e f / e i p#r e l a t e d t o p i c s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#s e e r e s u l t s a b o u t> ; ns2 : wasDerivedFrom <h t t p : / /www. g o o g l e . com/ i n s i d e s e a r c h / f e a t u r e s / s e a r c h / k n o w l e d g e . html> . <h t t p : / / d a t a . eurecom . f r / i d / f r e s n e l / actorGKPDefaultLens> a ns1 : Lens ; ns1 : c l a s s L e n s D o m a i n <h t t p : / / d b p e d i a . o r g / o n t o l o g y / Actor> ; ns1 : group <h t t p : / / d a t a . eurecom . f r / i d / f r e s n e l / group / actorGroup> ; ns1 : p u r p o s e ns1 : d e f a u l t L e n s ; ns1 : s h o w P r o p e r t i e s <h t t p : / / d a t a . eurecom . f r / d e f / e i p#albums> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#awards> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#b o o k s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#born> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#c h i l d r e n> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#died> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#e d u c a t i o n> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#f i l m m u s i c c r e d i t s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#h e i g h t> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#movies> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#m o v i e s a n d t v s h o w s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#n a t i o n a l i t y> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#n o m i n a t i o n s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#p a r e n t s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#p a r t n e r> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#p e o p l e a l s o s e a r c h f o r> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#s e e r e s u l t s a b o u t> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#s i b l i n g s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#s p o u s e> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#t v s h o w s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#upcomingmovie> ; ns2 : wasDerivedFrom <h t t p : / /www. g o o g l e . com/ i n s i d e s e a r c h / f e a t u r e s / s e a r c h / k n o w l e d g e . html> .

152

Appendix C. DBpedia Ranked Properties in Fresnel Vocabulary

<h t t p : / / d a t a . eurecom . f r / i d / f r e s n e l / a d m i n i s t r a t i v e r e g i o n G K P D e f a u l t L e n s> a ns1 : Lens ; ns1 : c l a s s L e n s D o m a i n <h t t p : / / d b p e d i a . o r g / o n t o l o g y / A d m i n i s t r a t i v e R e g i o n> ; ns1 : group <h t t p : / / d a t a . eurecom . f r / i d / f r e s n e l / group / a d m i n i s t r a t i v e r e g i o n G r o u p> ; ns1 : p u r p o s e ns1 : d e f a u l t L e n s ; ns1 : s h o w P r o p e r t i e s <h t t p : / / d a t a . eurecom . f r / d e f / e i p#a r e a> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#c o l l e g e s a n d u n i v e r s i t i e s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#c o u n t y s e a t> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#f o u n d e d> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#i s l a n d g r o u p> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#l o c a l t i m e> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#p e o p l e a l s o s e a r c h f o r> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#p o p u l a t i o n> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#r i v e r s> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#s e e r e s u l t s a b o u t> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#unemploymentrate> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#u n i v e r s i t y> , <h t t p : / / d a t a . eurecom . f r / d e f / e i p#w e a t h e r> ; ns2 : wasDerivedFrom <h t t p : / /www. g o o g l e . com/ i n s i d e s e a r c h / f e a t u r e s / s e a r c h / k n o w l e d g e . html> . <h t t p : / / d a t a . eurecom . f r / i d / f r e s n e l / a d u l t a c t o r G K P D e f a u l t L e n s> a ns1 : Lens ; ns1 : c l a s s L e n s D o m a i n <h t t p : / / d b p e d i a . o r g / o

ntology/AdultActor>;ns1:group <http://data.eurecom.fr/id/fresnel/group/adultactorGr oup>;ns1:purposens1:defaultLens;ns1:showProperties<http://data.eurecom.fr/def/e ip#awards>,<http://data.eurecom.fr/def/eip#books>,<http://data.eurecom.fr/def/eip#born> ,<http://data.eurecom.fr/def/eip#children>,<http://data.eurecom.fr/def/eip#died>,<htt p://data.eurecom.fr/def/eip#education>,<http://data.eurecom.fr/def/eip#height>,<htt p://data.eurecom.fr/def/eip#movies>,<http://data.eurecom.fr/def/eip#moviesandtvshow s>,<http://data.eurecom.fr/def/eip#parents>,<http://data.eurecom.fr/def/eip#peoplea lsosearchfor>,<http://data.eurecom.fr/def/eip#seeresultsabout>,<http://data. eurecom.fr/def/eip#siblings>,<http://data.eurecom.fr/def/eip#spouse>,<http://data. eurecom.fr/def/eip#tvshows>;ns2:wasDerivedFrom <http://www.google.com/insidesearch/fe atures/search/knowledge.html>.<http://data.eurecom.fr/id/fresnel/aircraftGKPDef aultLens> ans1:Lens;ns1:classLensDomain<http://dbpedia.org/ontology/Aircraft>; ns1:group <http://data.eurecom.fr/id/fresnel/group/aircraftGroup>;ns1:purposens1:def aultLens;ns1:showProperties<http://data.eurecom.fr/def/eip#cruisespeed>,<http:// data.eurecom.fr/def/eip#designer>,<http://data.eurecom.fr/def/eip#enginetype>,<htt p://data.eurecom.fr/def/eip#firstflight>,<http://data.eurecom.fr/def/eip#introduced> ,<http://data.eurecom.fr/def/eip#length>,<http://data.eurecom.fr/def/eip#manufactur er>,<http://data.eurecom.fr/def/eip#manufacturers>,

153

<http://data.eurecom.fr/def/eip#range>,<http://data.eurecom.fr/def/eip#retired>,<http: //data.eurecom.fr/def/eip#seeresultsabout>,<http://data.eurecom.fr/def/eip#topspee d>,<http://data.eurecom.fr/def/eip#unitcost>,<http://data.eurecom.fr/def/eip#weight> ,<http://data.eurecom.fr/def/eip#wingspan>;ns2:wasDerivedFrom <http://www.google.com/insi desearch/features/search/knowledge.html>.<http://data.eurecom.fr/id/fresnel/airl ineGKPDefaultLens> ans1:Lens;ns1:classLensDomain<http://dbpedia.org/ontology /Airline>;ns1:group <http://data.eurecom.fr/id/fresnel/group/airlineGroup>;ns1:purpos ens1:defaultLens;ns1:showProperties<http://data.eurecom.fr/def/eip#ceo>,<http://d ata.eurecom.fr/def/eip#founded>,<http://data.eurecom.fr/def/eip#founder>,<http://da ta.eurecom.fr/def/eip#founders>,<http://data.eurecom.fr/def/eip#headquarters>,<http ://data.eurecom.fr/def/eip#peoplealsosearchfor>,<http://data.eurecom.fr/def/eip#rec entposts>,<http://data.eurecom.fr/def/eip#seeresultsabout>,<http://data.eurecom.fr/ def/eip#stockprice>,<http://data.eurecom.fr/def/eip#trackflightstatus>;ns2: wasDerivedFrom <http://www.google.com/insidesearch/features/search/knowledge.html> .

Listing C.1: An excerpt of the Fresnel vocabulary for top properties mappings of DBpedia 3.9

Appendix D

SAP HANA Semantic Services API

D.1

XSJS API Implementation

There are four dierent XSJS les in this project: • disambiguate.xsjs: Gets most appropriate entities / categories for a searchword • describe.xsjs: Loads the properties which are available for an entity • enrich.xsjs: Retrieves the data, ltering is possible • tableMatch.xsjs: Matches two tables by sending their cell values The disambiguate.xsjs and tableMatch.xsjs use an XSJS lib called DbpediaLib.xsjslib. You can import it with this code: $.import(harubixMatch.WebContent , DBpediaLib); This library contains the core functionalities of our services. You can use the functions in this library to build your own API extension. Just add another XSJS le, import with the command above, and you can use one of those functions:

$.harubixMatch.WebContent.DbpediaLib.queryEntitiesWithTypes(query,fuzzyStr, limit)

This will query for a search word in the TYPES table. It returns a JSON object. More specically, it will return an array containing all found entities (called entities). For each element of entities the text search score (txtScore) and the number of incoming associations (incomingNo) is returned, as well as an array of the associated types (types). Each element in types has a name and an order. Each element of entities also has a finalScore, but this is set to zero and will be calculated in the method getEntitiesWithTypes. Generally, you should not use this method at all, as it is a helper method for

getEntitiesWithTypes

D.1. XSJS API Implementation

155

$.harubixMatch.WebContent.DbpediaLib.getEntitiesWithTypes(query, limit, incomingNoWeight, fuzzy, multilang)

This method can be used externally. First of all it processes the fuzzy parameter (which should be double) into a string. Here textSearch=compare is added for all queries except for the INTERLANGUAGE table. This is done to achieve similar text search scores for both search with and without fuzzy. The INTERLANGUAGE table currently does not support textSearch=compare because of its column type (has to be SHORTTEXT or TEXT for that) If multilang is set to true, the INTERLANGUAGE table is searched (fuzzy, if specied). The top rated result is then used for further processing (replaces the query parameter). Then the queryEntitiesWithTypes method is called. If it returns no results, the RAWPROPERTIES are searched (example: AAPL). The top rated result will replace query parameter and once again queryEntitiesWithTypes is

called. Afterwards, the biggest number of incoming associations is determined from the returned ratingArray. The nalScore is calculated and saved, the ratingArray is sorted by that nalScore. The nalScore is currently calculated like this:

$.harubixMatch.WebContent.DBpediaLib.getEntityTypeWithContext(query,context,limit)

This returns one entity which ts most likely in the context given (JSON object). For each entity in the context the method getCategories is called. From this collection of types, a type vector is created a list of types, and if a type occurs more than once, its score is added to the existing entry. Then the list is sorted by score. Then the possible entities for the query are retrieved by calling getEntityWithTypes. Now the types of the entities are compared (helper method compareQueryWithContext) to the created type vector, starting with the most likely entity and the highest rated score. As soon as there is a match, the corresponding entity/type is returned.

$.harubixMatch.WebContent.DBpediaLib.getEntities(query, limit, incomingNoWeight, fuzzy, multilang, verbose)

This is a small method to only return entities and no types.

$.harubixMatch.WebContent.DBpediaLib.getCategories(query, limit, incomingNoWeight, orderWeight, fuzzy, multilang, verbose)

Here, only categories are returned. The order of the categories is determined by the finalScore of the entity they belong to and their order. The higher the order, the less specic the category is. The parameter orderWeight inuences how much is subtracted from the nalScore in case the value of order exceeds 1. Table match for example uses this function for every cell value:

$.harubixMatch.WebContent.DBpediaLib.getCategories(tables[i].getColumn(j).getCell(k).getValue(), 20, 0.15, 0.1, fuzzy, multilang, false);

156

Appendix D. SAP HANA Semantic Services API

D.2

D.2.1

D.2.1.1

{

API Documentation

Entity Enrichment

Describe Entity (describe.xsjs)

abstract : true, thumbnail : true, attributes : [{ type : http://dbpedia.org/property/name }, ...], o u t g o i n g A s s o c i a t i o n s : [] }

Listing D.1: Entity description API call return values sample

D.2.1.2

{

Enrich Entity (enrich.xsjs)

abstract: Give a query, // Mandatory filter: Filter the query with a JSON array /* { a b s t r a c t :true, a t t r i b u t e s :[{ n a m e },..],..} default: {} [no filter] */ }

type

:

Listing D.2: API call parameters for entity enrichment

{ a b s t r a c t : A p p l e is a company... a t t r i b u t e s : [{ type : name , v a l u e : A p p l e Inc. }, ...], o u t g o i n g A s s o c i a t i o n s : [...] } ,

Listing D.3: Entity enrichment API call return values sample

D.2.2

Entity Disambiguation (disambiguate.xsjs)

D.2. API Documentation

157

{ query: Give a query , entityMode: Switch between returning entities or categories { values: true or false default: true }, limit: Limit for the algorithm(bigger = more results) { values: 1 - inf[int] default: 75 }, incomingNoWeight: Weight of the number of incoming associations of an entity(bigger means more weight) { values: +-inf[float] default: 0.15 }, orderWeight: Weight of more specific types(bigger means specific types are favored more) { values: +-inf[float] default: 0.1 }, fuzzy: Parameter for fuzzy search { values: 0 - 1.0[float] default: 1.0[no fuzzy search] }, multilang: Search in INTERLANGUAGE table { values: true or false default: false }, context: Give context as a JSON array { values: { context: [{ entity : }] } default: {} }, verbose: Make the output more detailed { values: true or false default: false }, openSearch: Used for SAPUI5 search suggestion { values: true or false default: false } }

Listing D.4: API call parameters for entity disambiguation

158

Appendix D. SAP HANA Semantic Services API

// Entity mode { entities : [{ name : Microsoft , score :0.9681761690228368},...]} // Category mode { types : [{ name : Company , score :0.9681761690228368},...]}

Listing D.5: Entity disambiguation API call return values sample

D.2.3

{

Schema Matching (matchTables.xsjs)

fuzzy: Parameters for fuzzy search { values: 0-1.0 [float], default: 1.0 [fuzzy disabled] }, multilang: Search in INTERLANGUAGE table { values: true or false, default: false }, debug: Returns information useful for debugging i.e. weighted types per cell }

Listing D.6: API call parameters for schema matching

{ matches : [{ column1: 0, column2: 0, score: 0.627861, suggestedName: Plant },{ column1 :1, column2 :1, score: 0.868278, suggestedName: Company },{ column1: 2, column2: 2, score: 0.952602, suggestedName: Country }] }

D.2. API Documentation

159

Listing D.7: Schema matching API call return values sample

Appendix E

Source Code for Mappings

E.1

{

Open Licenses Mappings

license_id : [ODC-PDDL-1.0], disambiguations : [Open Data Commons Public Domain Dedication and License (PDDL)] }, { license_id : [CC-BY-SA-4.0 , CC-BY-SA-3.0], disambiguations : [cc-by-sa , CC BY-SA , Creative Commons Attribution Share-Alike] }, { license_id : [CC-BY-NC-4.0], disambiguations : [Creative Commons Non-Commercial (Any)] }, { license_id : [ODC-BY-1.0], disambiguations : [Open Data Commons Attribution License] }, { license_id : [CC-BY-4.0], disambiguations : [Creative Commons Attribution , CC-BY , CreativeCommonsAttributionCCBY25] }, { license_id : [geogratis], disambiguations : [Geogratis] }, { license_id : [CC0-1.0], disambiguations : [Creative Commons CCZero , CC0] }, { license_id : [ODbL-1.0], disambiguations : [Open Data Commons Open Database License (ODbL) , ODBL] }, { license_id : [OGL-UK-1.0 , OGL-UK-2.0 , OGL-UK-3.0], disambiguations : [UK Open Government Licence (OGL) , OGL]

E.1. Open Licenses Mappings

161

}, { license_id : [GPL-3.0 , GPL-2.0], disambiguations : [GNU General Public License , gpl-2.0] }, { license_id : [ukclickusepsi], disambiguations : [UK PSI (Public Sector Information) Click-Use Licence , ukclickusepsi , UK Click Use PSI] }, { license_id : [GFDL-1.3-no-cover-texts-no-invariantsections], disambiguations : [GNU Free Documentation License] }, { license_id : [MIT], disambiguations : [The MIT License (MIT) , mit-license , MIT License (MIT) , MIT] }, { license_id : [ukcrown-withrights], disambiguations : [UK Crown Copyright with data.gov.uk rights , ukcrown-withrights] }, { license_id : [canadacrown], disambiguations : [Canada Crown Copyright , canada-crown] }, { license_id : [BSD-2-Clause , BSD-3-Clause], disambiguations : [bsd-license] }, { license_id : [LGPL-2.1 , LGPL-3.0], disambiguations : [GNU Lesser General Public License , lgpl-2.1] }, { license_id : [SPL-1.0], disambiguations : [sunpublic , Sun Public License , SPL] }, { license_id : [GPL-3.0 , GPL-2.0], disambiguations : [GNU General Public License , gpl-3.0] }, { license_id : [Apache-2.0 , Apache-1.1], disambiguations : [Apache License , apache] }

162

Appendix E. Source Code for Mappings

Listing E.1: The mappings of the Open Licenses for the LOD Cloud on the Datahub

E.2

{

Semantic Social News Aggregation Mappings

alchemy : Arts & Entertainment , alchemyCode : arts_entertainment , DMOZ : [Arts , Society], stack : [music , movies] }, { alchemy : Business , alchemyCode : business , DMOZ : [Business , News , Shopping], stack : [money , pm , answers.onstartups , patents , quant] }, { alchemy : Computers & Internet , alchemyCode : computer_internet , DMOZ : [Computers , Science], stack : [stackoverflow , serverfault , superuser] }, { alchemy : Culture & Politics , alchemyCode : culture_politics , DMOZ : [News , Society , history], stack : [politics] }, { alchemy : Gaming , alchemyCode : gaming , DMOZ : [Games], stack : [gaming] }, { alchemy : Health , alchemyCode : health , DMOZ : [Health , Society], stack : [fitness , sustainability] }, {

E.2. Semantic Social News Aggregation Mappings

163

alchemy : Law & Crime , alchemyCode : law_crime , DMOZ : [News , Society] }, { alchemy : Religion , alchemyCode : religion , DMOZ : [Reference , Society], stack : [islam , christianity] }, { alchemy : Recreation , alchemyCode : recreation , stack : [philosophy , photo], DMOZ : [Recreation , Society] }, { alchemy : Science & Technology , alchemyCode : science_technology , DMOZ : [Science , News], stack : [stats , math] }, { alchemy : Sports , alchemyCode : sports , DMOZ : [Sports , News], stack : [sports] }, { alchemy : Weather , alchemyCode : weather , DMOZ : [News] }

Listing E.2: The mappings of the StackExchange services with DMOZ and Alchemy API

{ alchemy : Arts & Entertainment , alchemyCode : arts_entertainment , DMOZ : [Arts , Society], youtube : [Entertainment , Comedy , Music , Film], youtube_edu : [{ id : 1,

164

Appendix E. Source Code for Mappings

title : Fine Arts }, { id : 17, title : Language Arts }] }, { alchemy : Business , alchemyCode : business , DMOZ : [Business , News , Shopping], youtube : [News , Videoblog], youtube_edu : [{ id : 65, title : Business }] }, { alchemy : Computers & Internet , alchemyCode : computer_internet , DMOZ : [Computers , Science], youtube : [Tech , Howto , Videoblog], youtube_edu : [{ id : 109, title : Computer Science }, { id : 47, title : Science }, { id : 39, title : Mathematics }] }, { alchemy : Culture & Politics , alchemyCode : culture_politics , DMOZ : [News , Society], youtube : [News , People], youtube_edu : [{

E.2. Semantic Social News Aggregation Mappings

165

id : 10, title : History & Social Sciences }, { id : 375, title : Social Sciences }] }, { alchemy : Gaming , alchemyCode : gaming , DMOZ : [Games], youtube : [Games , Entertainment] }, { alchemy : Health , alchemyCode : health , DMOZ : [Health , Society], youtube : [People , News , Education], youtube_edu : [{ id : 6, title : Health }] }, { alchemy : Law & Crime , alchemyCode : law_crime , DMOZ : [News , Society], youtube : [News], youtube_edu : [{ id : 193, title : Law }, { id : 10, title : History & Social Sciences }] }, { alchemy : Religion , alchemyCode : religion , DMOZ : [Reference , Society],

166

Appendix E. Source Code for Mappings

youtube : [News , Education], youtube_edu : [{ id : 178, title : Religion }, { id : 10, title : History & Social Sciences }] }, { alchemy : Recreation , alchemyCode : recreation , DMOZ : [Recreation , Society], youtube : [Entertainment , Comedy] }, { alchemy : Science & Technology , alchemyCode : science_technology , DMOZ : [Science , News], youtube : [Tech , News , Howto , Education], youtube_edu : [{ id : 47, title : Science }, { id : 56, title : University }] }, { alchemy : Sports , alchemyCode : sports , DMOZ : [Sports , News], youtube : [Sports , News , Autos] }, { alchemy : Weather , alchemyCode : weather , DMOZ : [News], youtube : [News] }

E.2. Semantic Social News Aggregation Mappings

167

Listing E.3: The mappings of YouTube categories with DMOZ and Alchemy API