SOCIAM

A-posteori provenance-enabled linking of publications and datasets via crowdsourcing

Laura Drãgan, Markus Luczak-Rösch, Bettina Berendt,
Elena Simperl, Heather Packer, Luc Moreau

EPSRC
Engineering and Physical Sciences
Research Council

presented at LCPD 2014, London, 12 September 2014

Motivation

- Data driven science
- Reproducible & verifiable research

- this workshop ..

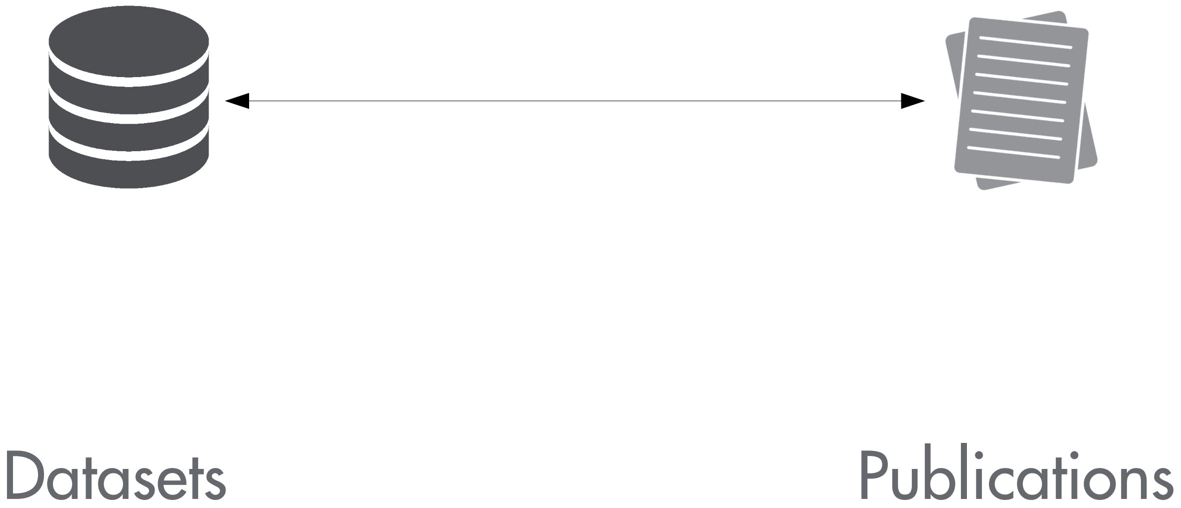
Motivation

Publications

Motivation

Datasets

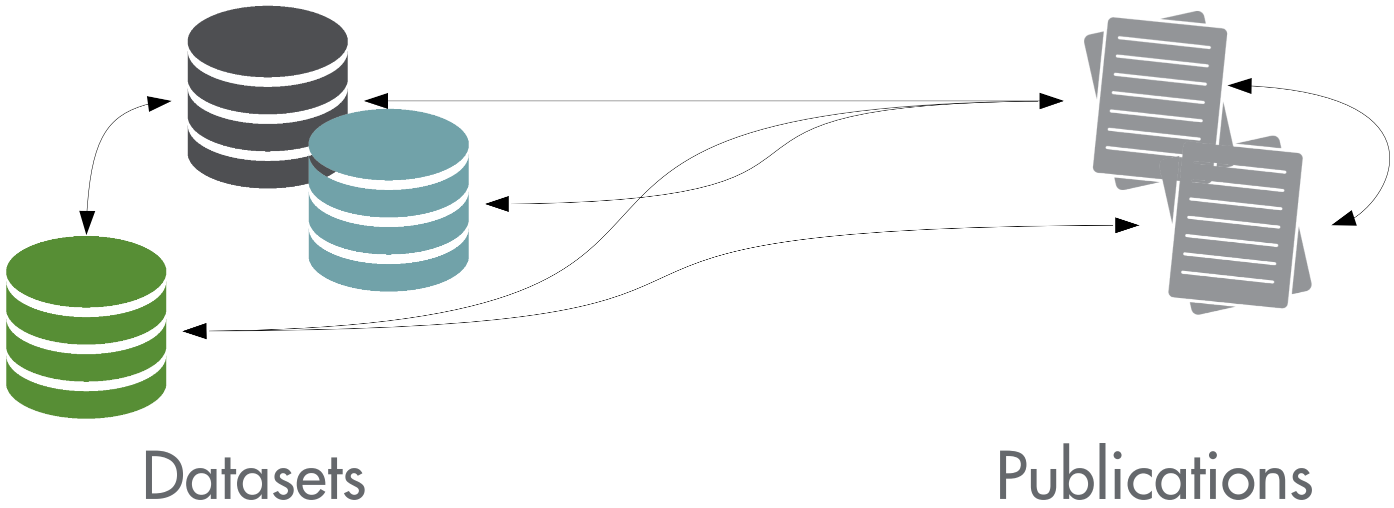
Motivation



Datasets

Publications

Motivation



Motivation

Datasets

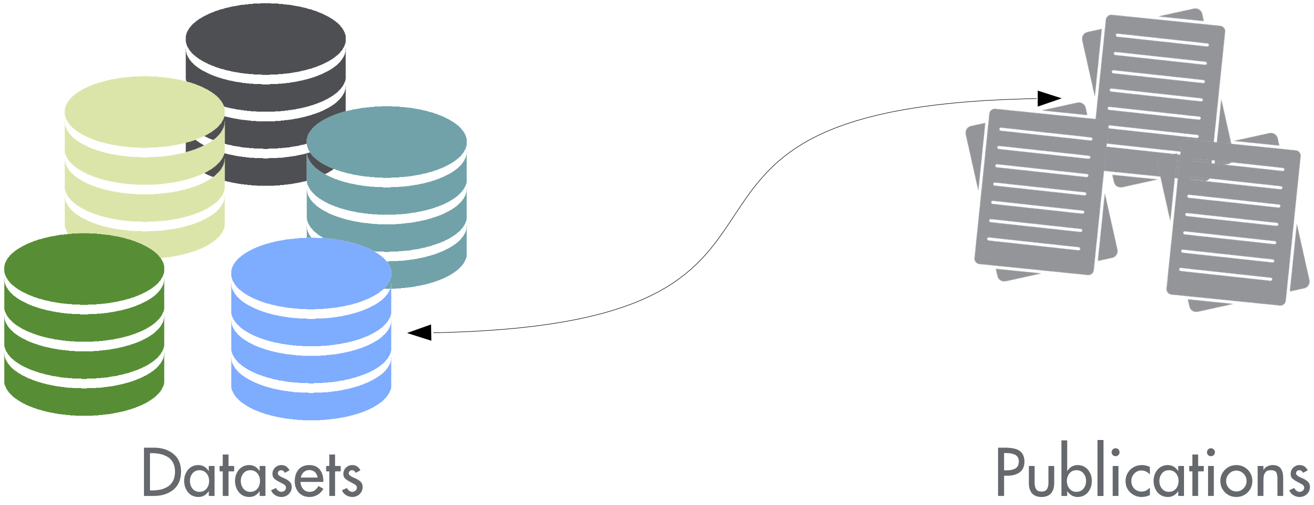
?

Publications

Different versions

Ambiguous references

Motivation : Data citation



Datasets

Publications

Different versions

Ambiguous references

Creating explicit connections

- Publication – Publication

- Dataset – Dataset

- Publication – Dataset
- Dataset – Publication

Two datasets & usecases

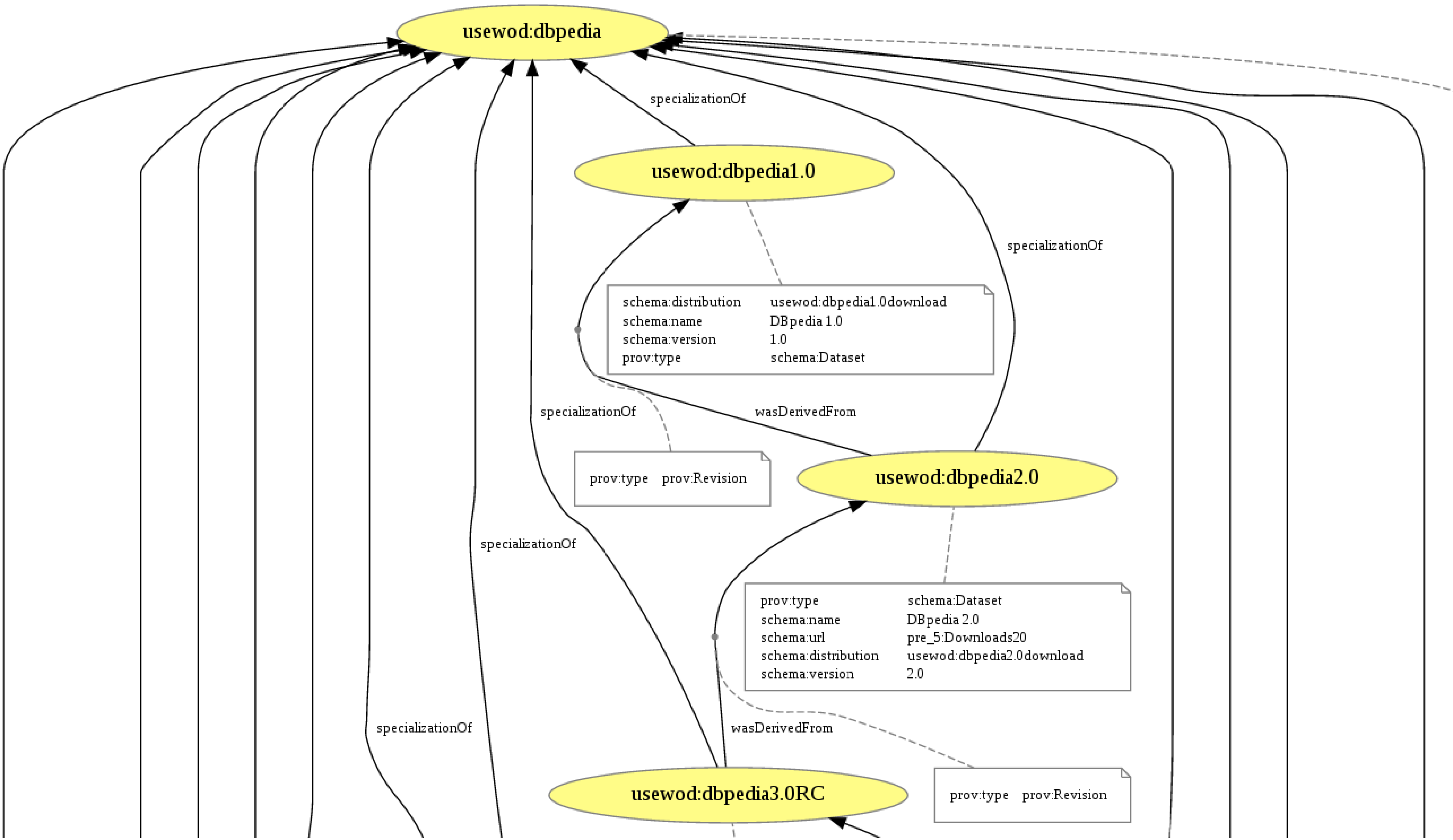
- DBpedia
- USEWOD

DBpedia

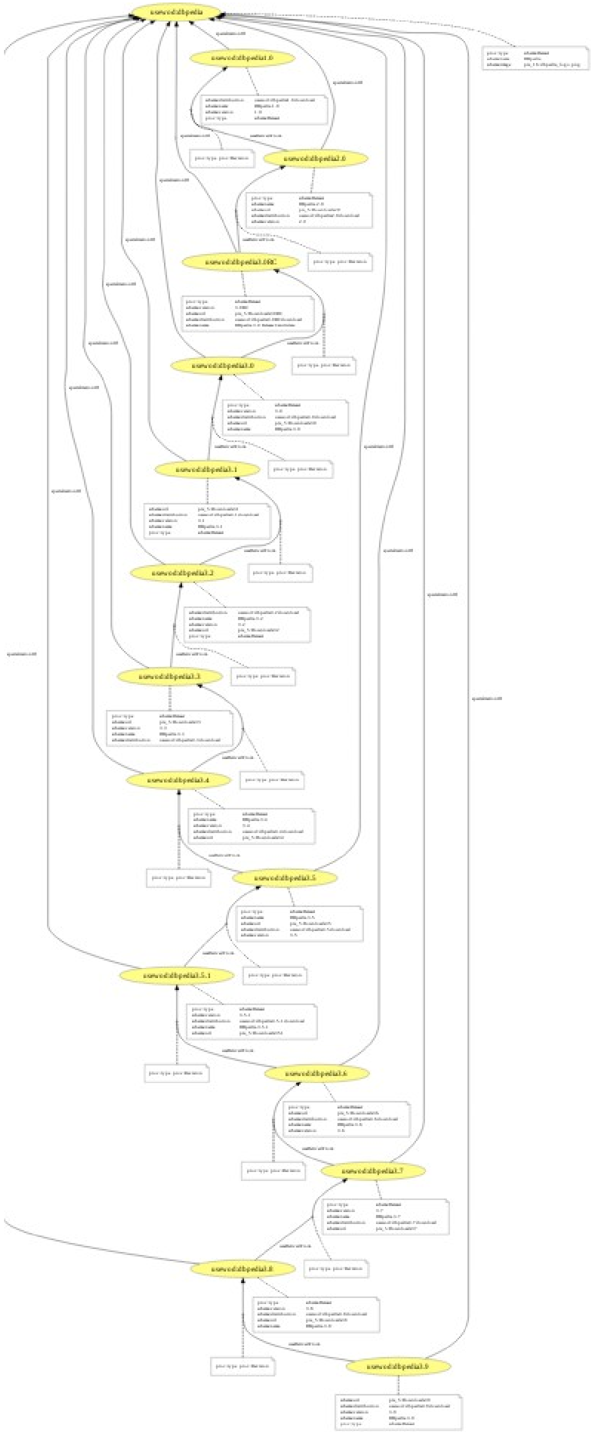
- http://dbpedia.org

- Linked Open Data dataset
- Automatically extracted from Wikipedia

DBpedia


usewod:dbpedia
specializationOf
usewod:dbpedia1.0
specializationOf
schema:distribution usewod:dbpedia1.0download
schema:name DBpedia 1.0
schema:version 1.0
prov:type schema:Dataset
specializationOf
wasDerivedFrom
prov:type prov:Revision
usewod:dbpedia2.0
specializationOf
prov:type schema:Dataset
schema:name DBpedia 2.0
schema:url pre_5:Downloads20
schema:distribution usewod:dbpedia2.0download
schema:version 2.0
specializationOf
wasDerivedFrom
usewod:dbpedia3.0RC
prov:type prov:Revision

DBpedia

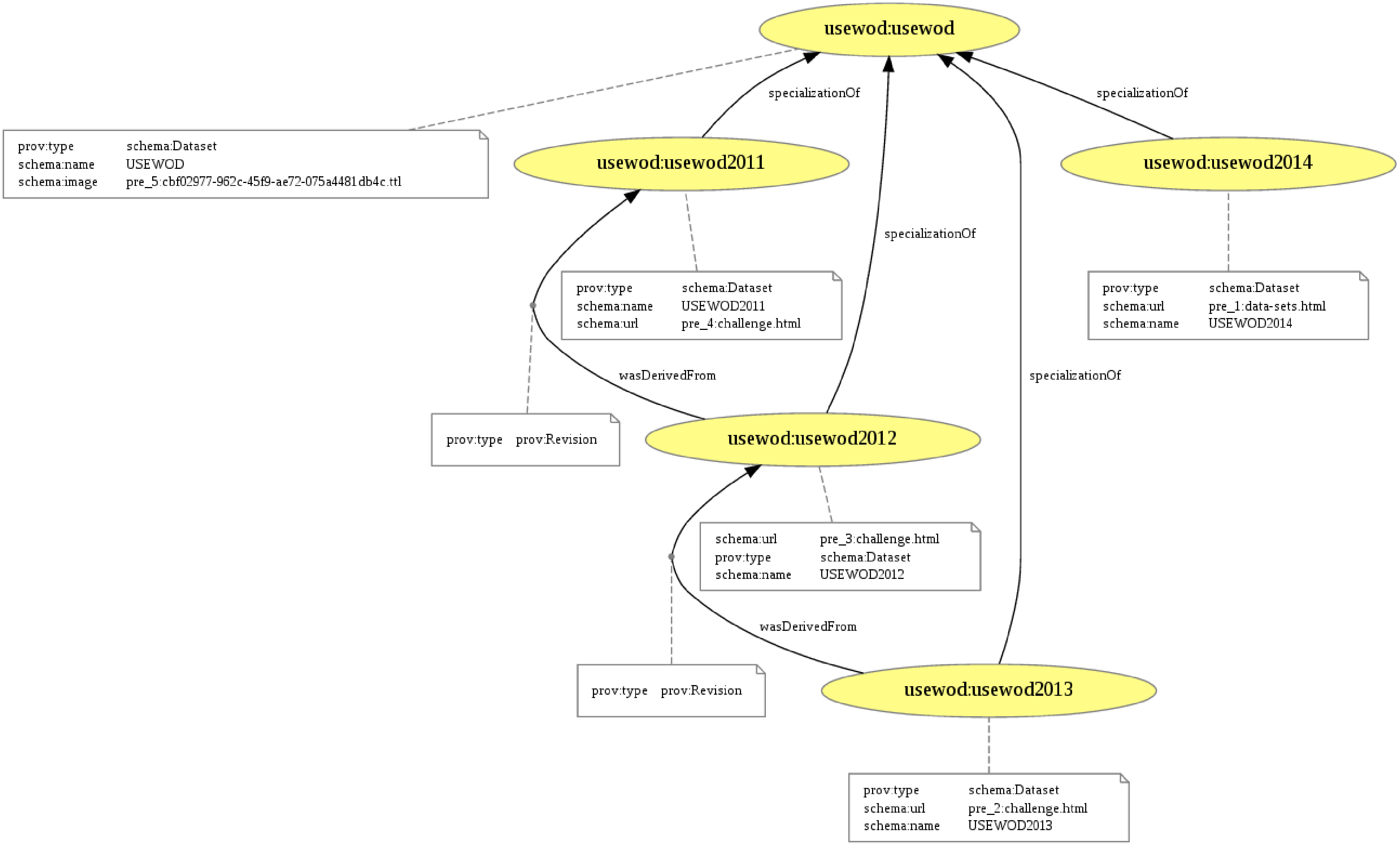

usewod:dbpedia
usewod:dbpedia1.0
usewod:dbpedia2.0
usewod:dbpedia3.0RC
usewod:dbpedia3.0
usewod:dbpedia3.1
usewod:dbpedia3.2
usewod:dbpedia3.3
usewod:dbpedia3.4
usewod:dbpedia3.5
usewod:dbpedia3.5.1
usewod:dbpedia3.6
usewod:dbpedia3.7
usewod:dbpedia3.8
usewod:dbpedia3.9

USEWOD

- http://usewod.org

- Server access logs from Linked Data servers
- 4 yearly versions, starting 2011

USEWOD



prov:type	schema:Dataset
schema:name	USEWOD
schema:image	pre_5:cbf02977-962c-45f9-ae72-075a4481db4c.ttl

prov:type	schema:Dataset
schema:name	USEWOD2011
schema:url	pre_4:challenge.html

prov:type	schema:Dataset
schema:url	pre_1:data-sets.html
schema:name	USEWOD2014

prov:type	prov:Revision

schema:url	pre_3:challenge.html
prov:type	schema:Dataset
schema:name	USEWOD2012

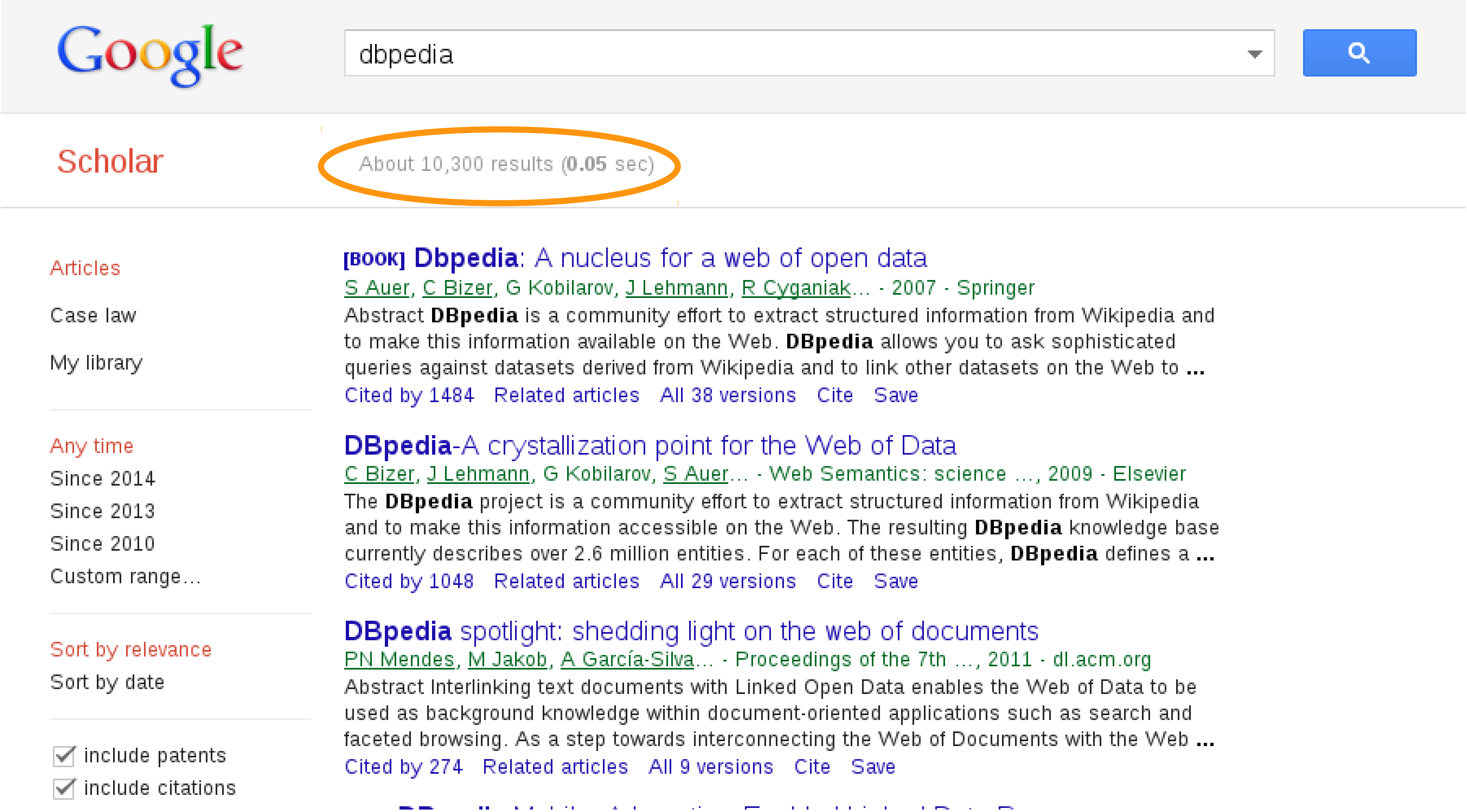
prov:type	prov:Revision

prov:type	schema:Dataset
schema:url	pre_2:challenge.html
schema:name	USEWOD2013

Dataset – Dataset links

- Inclusion
- Dependence
- Transformation
- Aggregation
- Projection
- …

Publications



dbpedia

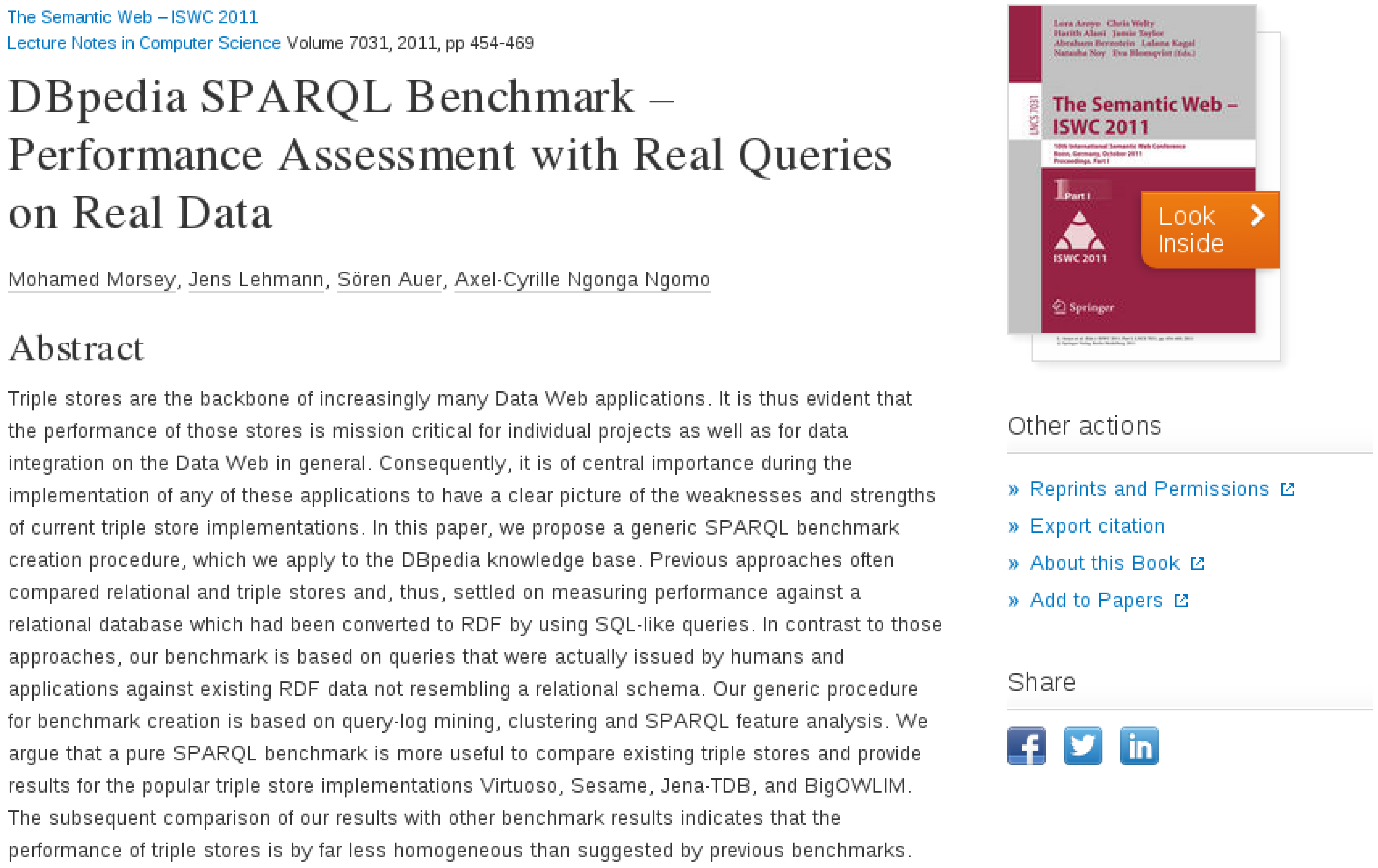
About 10,300 results (**0.05** sec)

[BOOK] **Dbpedia**: A nucleus for a web of open data
S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak... - 2007 - Springer
Abstract **DBpedia** is a community effort to extract structured information from Wikipedia and to make this information available on the Web. **DBpedia** allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to **...**
Cited by 1484 Related articles All 38 versions Cite Save

DBpedia-A crystallization point for the Web of Data
C Bizer, J Lehmann, G Kobilarov, S Auer... - Web Semantics: science ..., 2009 - Elsevier
The **DBpedia** project is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web. The resulting **DBpedia** knowledge base currently describes over 2.6 million entities. For each of these entities, **DBpedia** defines a **...**
Cited by 1048 Related articles All 29 versions Cite Save

DBpedia spotlight: shedding light on the web of documents
PN Mendes, M Jakob, A García-Silva... - Proceedings of the 7th ..., 2011 - dl.acm.org
Abstract Interlinking text documents with Linked Open Data enables the Web of Data to be used as background knowledge within document-oriented applications such as search and faceted browsing. As a step towards interconnecting the Web of Documents with the Web **...**
Cited by 274 Related articles All 9 versions Cite Save

Publications



Lora Aroyo Chris Welty
Harith Alani Jamie Taylor
Abraham Bernstein Lalana Kagal
Natasha Noy Eva Blomqvist (Eds.)
LNCS 7031
The Semantic Web – ISWC 2011
Part I
ISWC 2011
Springer
Look Inside

The Semantic Web – ISWC 2011
Lecture Notes in Computer Science Volume 7031, 2011, pp 454-469

DBpedia SPARQL Benchmark – Performance Assessment with Real Queries on Real Data

Mohamed Morsey, Jens Lehmann, Sören Auer, Axel-Cyrille Ngonga Ngomo

Abstract

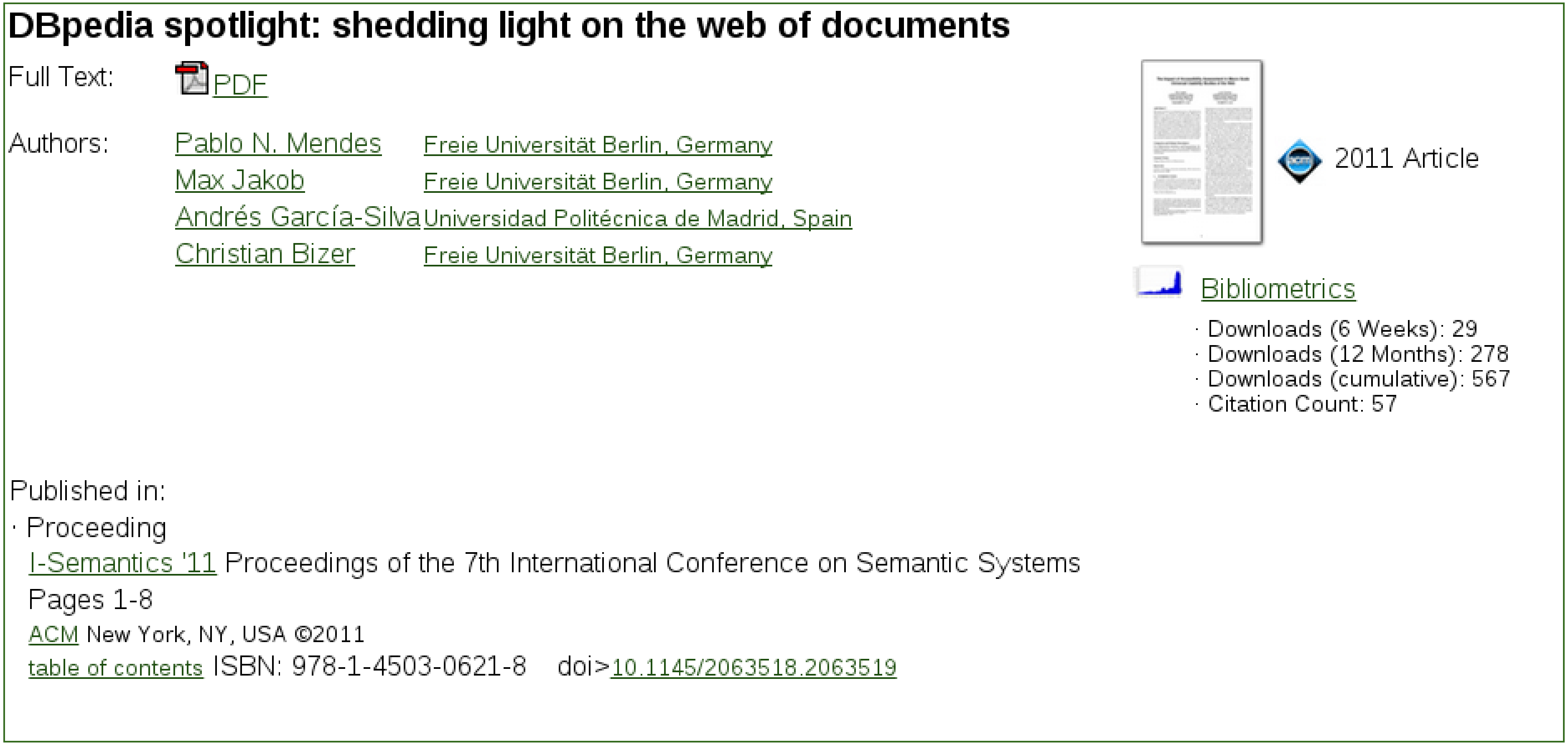
Triple stores are the backbone of increasingly many Data Web applications. It is thus evident that the performance of those stores is mission critical for individual projects as well as for data integration on the Data Web in general. Consequently, it is of central importance during the implementation of any of these applications to have a clear picture of the weaknesses and strengths of current triple store implementations. In this paper, we propose a generic SPARQL benchmark creation procedure, which we apply to the DBpedia knowledge base. Previous approaches often compared relational and triple stores and, thus, settled on measuring performance against a relational database which had been converted to RDF by using SQL-like queries. In contrast to those approaches, our benchmark is based on queries that were actually issued by humans and applications against existing RDF data not resembling a relational schema. Our generic procedure for benchmark creation is based on query-log mining, clustering and SPARQL feature analysis. We argue that a pure SPARQL benchmark is more useful to compare existing triple stores and provide results for the popular triple store implementations Virtuoso, Sesame, Jena-TDB, and BigOWLIM. The subsequent comparison of our results with other benchmark results indicates that the performance of triple stores is by far less homogeneous than suggested by previous benchmarks.

Other actions

- » Reprints and Permissions
- » Export citation
- » About this Book
- » Add to Papers

Share

Publications



DBpedia spotlight: shedding light on the web of documents

Full Text: PDF

Authors:	Pablo N. Mendes	Freie Universität Berlin, Germany
	Max Jakob	Freie Universität Berlin, Germany
	Andrés García-Silva	Universidad Politécnica de Madrid, Spain
	Christian Bizer	Freie Universität Berlin, Germany

2011 Article

Bibliometrics

- Downloads (6 Weeks): 29
- Downloads (12 Months): 278
- Downloads (cumulative): 567
- Citation Count: 57

Published in:

- Proceeding

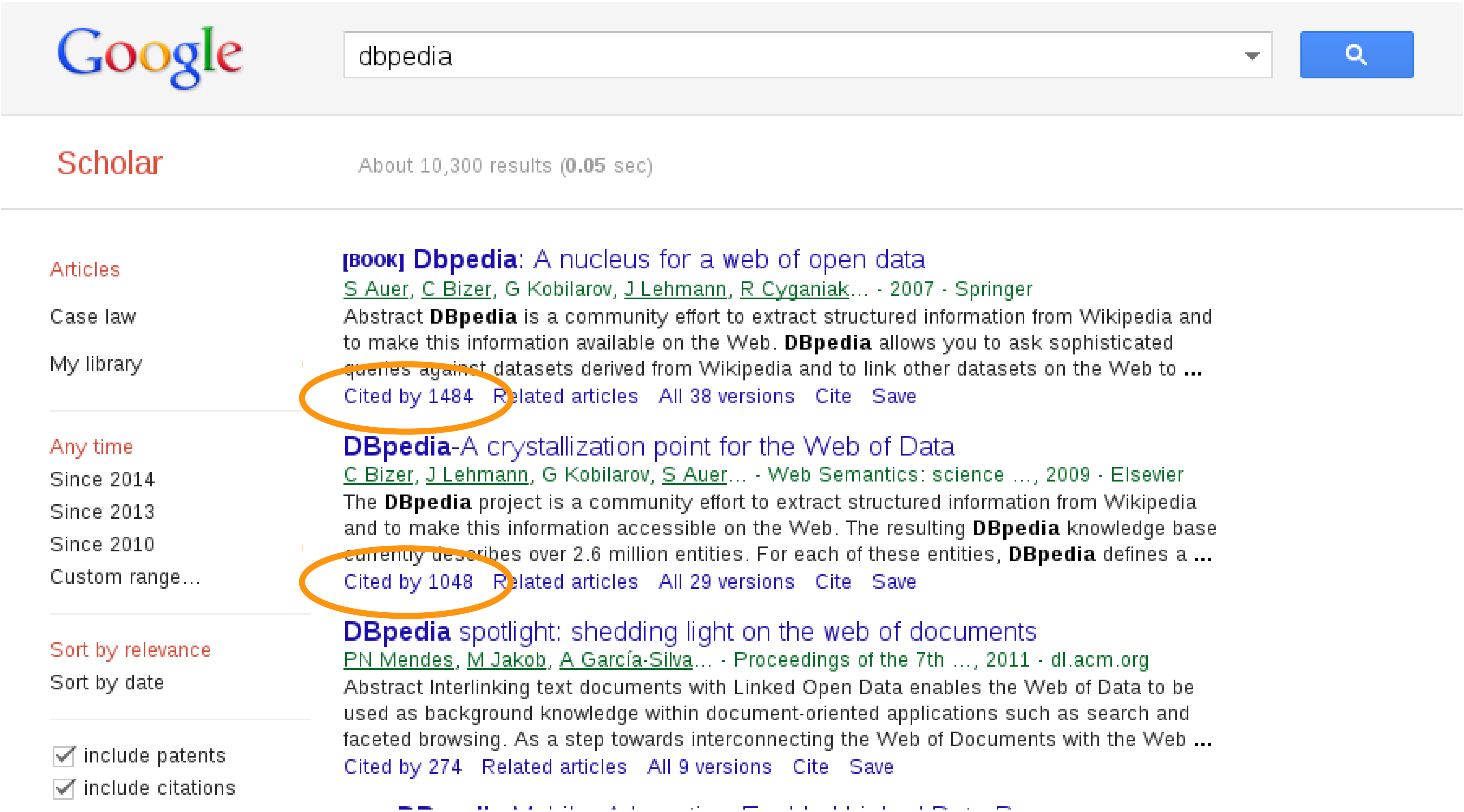
  I-Semantics '11 Proceedings of the 7th International Conference on Semantic Systems

  Pages 1-8

  ACM New York, NY, USA ©2011

  table of contents ISBN: 978-1-4503-0621-8 doi>10.1145/2063518.2063519

Publications



About 10,300 results (**0.05** sec)

[BOOK] **Dbpedia**: A nucleus for a web of open data
S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak… - 2007 - Springer
Abstract **DBpedia** is a community effort to extract structured information from Wikipedia and to make this information available on the Web. **DBpedia** allows you to ask sophisticated queries against datasets derived from Wikipedia and to link other datasets on the Web to **...**
Cited by 1484 Related articles All 38 versions Cite Save

DBpedia-A crystallization point for the Web of Data
C Bizer, J Lehmann, G Kobilarov, S Auer… - Web Semantics: science …, 2009 - Elsevier
The **DBpedia** project is a community effort to extract structured information from Wikipedia and to make this information accessible on the Web. The resulting **DBpedia** knowledge base currently describes over 2.6 million entities. For each of these entities, **DBpedia** defines a **...**
Cited by 1048 Related articles All 29 versions Cite Save

DBpedia spotlight: shedding light on the web of documents
PN Mendes, M Jakob, A García-Silva… - Proceedings of the 7th …, 2011 - dl.acm.org
Abstract Interlinking text documents with Linked Open Data enables the Web of Data to be used as background knowledge within document-oriented applications such as search and faceted browsing. As a step towards interconnecting the Web of Documents with the Web **...**
Cited by 274 Related articles All 9 versions Cite Save

Publication – Dataset links

- Simple usage:
  “publication P uses dataset D”

- Complex / detailed usage:
  “**how** does publication P use dataset D”

Method for link generation

- Crowdsourcing

Crowdsourcing [Howe, 2006]

“Simply defined, crowdsourcing represents the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call. This can take the form of peer-production (when the job is performed collaboratively), but is also often undertaken by sole individuals. The crucial prerequisite is the use of the open call format and the large network of potential.”

Dimensions of crowdsourcing

- What is outsourced
  - Human skills
  - Difficult for machines
  - Macrotasks vs. microtasks

Dimensions of crowdsourcing

- What is outsourced
- Who is the crowd
  - Open call
  - Target specific skills or expertise

Dimensions of crowdsourcing

- What is outsourced
- Who is the crowd
- How is the task designed
    - Explicit vs implicit participation
    - In parallel vs sequentially
    - Coordination
    - Aggregation of answers

Dimensions of crowdsourcing

- What is outsourced
- Who is the crowd
- How is the task designed
- How are the results validated
  - Solution space open vs. closed
  - Ground truth known vs. Unknown
  - Performance and reputation measurements

Dimensions of crowdsourcing

- What is outsourced
- Who is the crowd
- How is the task designed
- How are the results validated
- How can the process be optimised
  - Incentives
  - Algorithmic task assignment

Dimensions of crowdsourcing

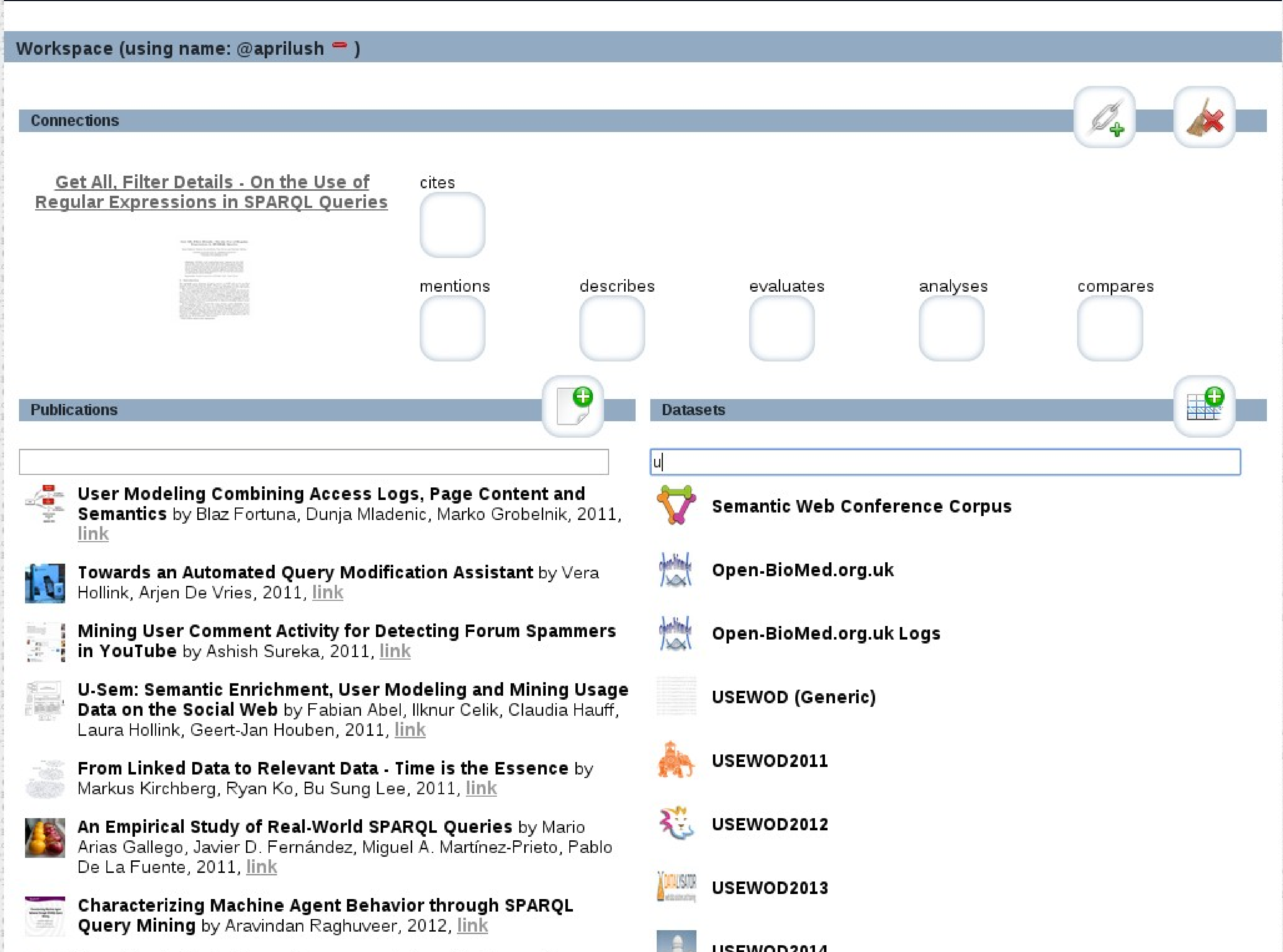
- What is outsourced
- Who is the crowd
- How is the task designed
- How are the results validated
- How can the process be optimised

[Quinn & Bederson, 2012]

USEWOD user study

- Run at USEWOD2014
- 1 hour
- 6 participants
- http://prov.usewod.org

USEWOD2014 - 4th International Workshop on Usage Analysis and the Web of Data



Workspace (using name: @aprilush −)

Connections

Get All, Filter Details - On the Use of Regular Expressions in SPARQL Queries

Publications

User Modeling Combining Access Logs, Page Content and Semantics by Blaz Fortuna, Dunja Mladenic, Marko Grobelnik, 2011, link

Towards an Automated Query Modification Assistant by Vera Hollink, Arjen De Vries, 2011, link

Mining User Comment Activity for Detecting Forum Spammers in YouTube by Ashish Sureka, 2011, link

U-Sem: Semantic Enrichment, User Modeling and Mining Usage Data on the Social Web by Fabian Abel, Ilknur Celik, Claudia Hauff, Laura Hollink, Geert-Jan Houben, 2011, link

From Linked Data to Relevant Data - Time is the Essence by Markus Kirchberg, Ryan Ko, Bu Sung Lee, 2011, link

An Empirical Study of Real-World SPARQL Queries by Mario Arias Gallego, Javier D. Fernández, Miguel A. Martínez-Prieto, Pablo De La Fuente, 2011, link

Characterizing Machine Agent Behavior through SPARQL Query Mining by Aravindan Raghuveer, 2012, link

Datasets

u

Semantic Web Conference Corpus

Open-BioMed.org.uk

Open-BioMed.org.uk Logs

USEWOD (Generic)

USEWOD2011

USEWOD2012

USEWOD2013

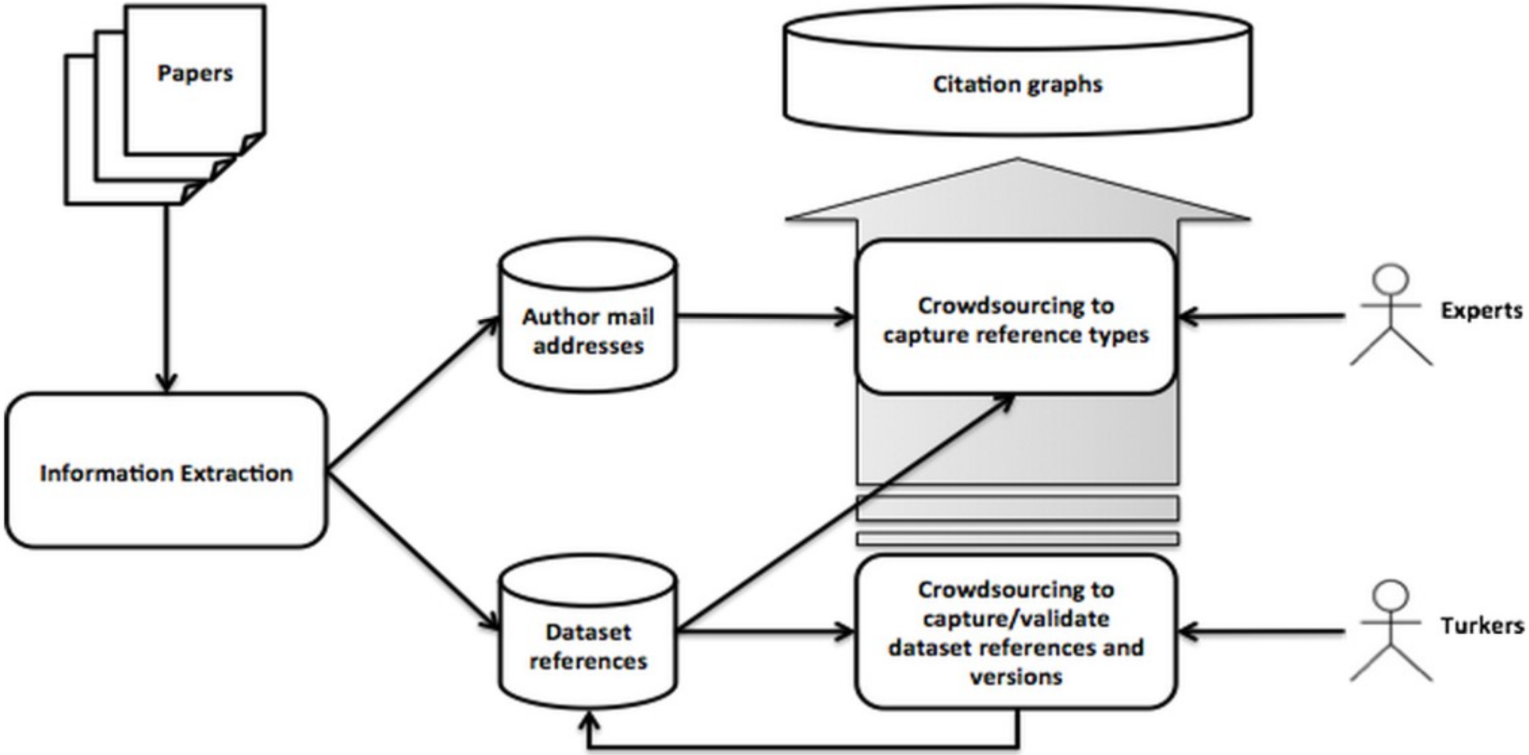
USEWOD2014

USEWOD user study results

Tasks	**81**	avg: 13.5, min: 2, max: 27
Publications	**19**	
Datasets	**2 (3)**	
Links	**95**	Inclusion: 62 Analysis: 21, Mention: 6

A generic process

- Hybrid approach
- Information extraction + crowdsourcing


Papers
Citation graphs
Author mail addresses
Crowdsourcing to capture reference types
Experts
Information Extraction
Dataset references
Crowdsourcing to capture/validate dataset references and versions
Turkers

Task definition

- Microtasks
- Annotate publications with dataset (and version) information
  - Which dataset and version is used
  - How is the dataset used

Who is the crowd

- Experts – authors of the papers, domain experts, librarians
- Non-experts – English speakers

Task description

- Non-experts validate extracted information about used datasets and versions.
- Experts and non-experts input information about used datasets and versions
- Experts input how the given datasets and versions are used.

Validation of results

- Non-experts / simple usage:
  - Algorithmic restrictions,
  - Information extraction,
  - Inter-annotator agreement
- Experts / complex usage:
  - Clustering
  - Inter-annotator agreement

Optimisation

- Gamification
- Twitter contest
- Target authors of the publications first
- Change the task
  - Find all publications that use a dataset D
- Incentivise
  - Show benefit to authors and readers
  - Pay-per-task, pay-per-time, prizes

“A-posteriori”

- After the writing of the publication

- Rich data citation network
- Incentivise the creation of data citation links at the time of writing

Conclusion

- Generate data citation graphs
- Feedback from the USEWOD study
- Hybrid approach: IE + crowdsourcing
- Participants: experts and non-experts
- Task descriptions can be tweaked