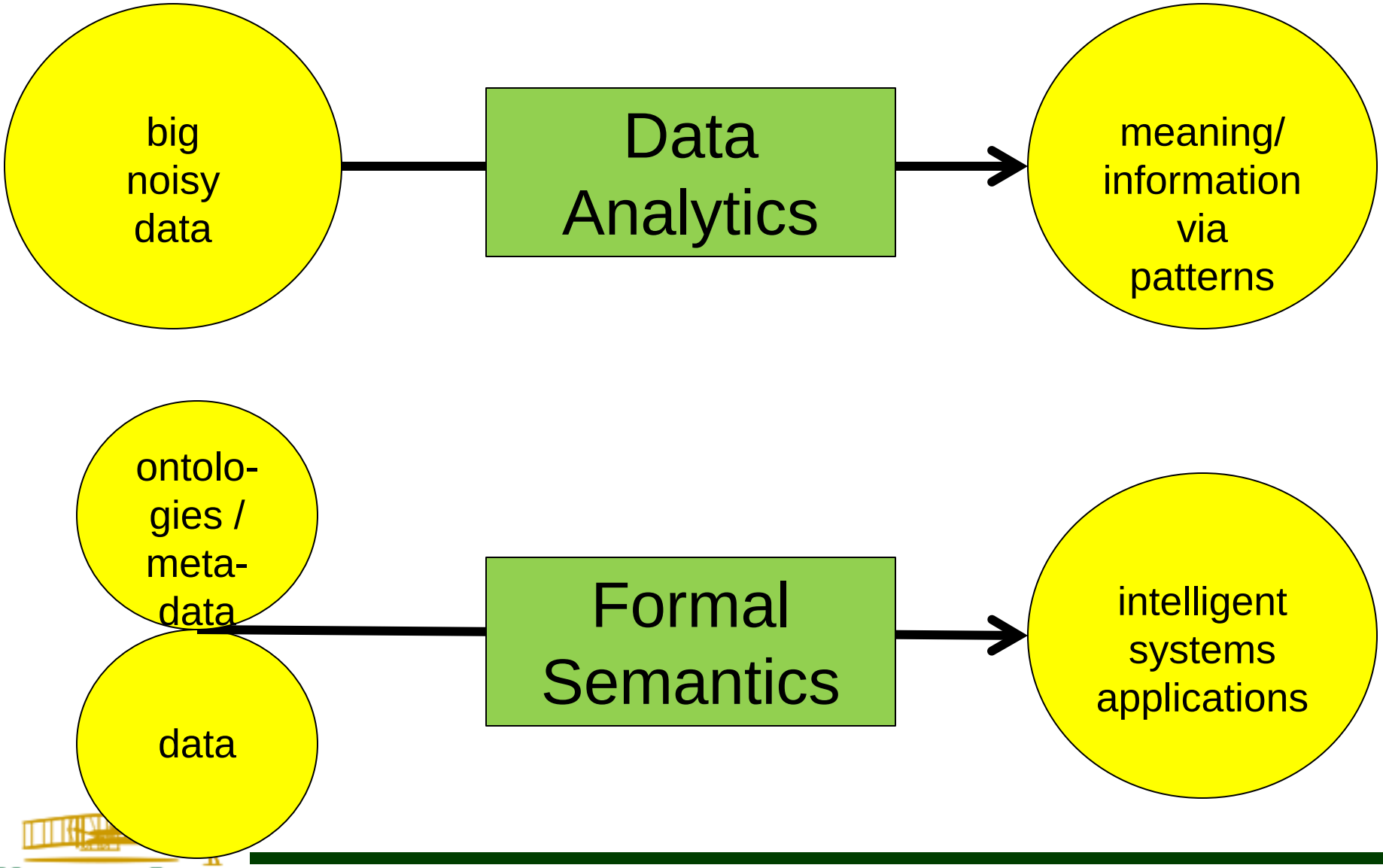
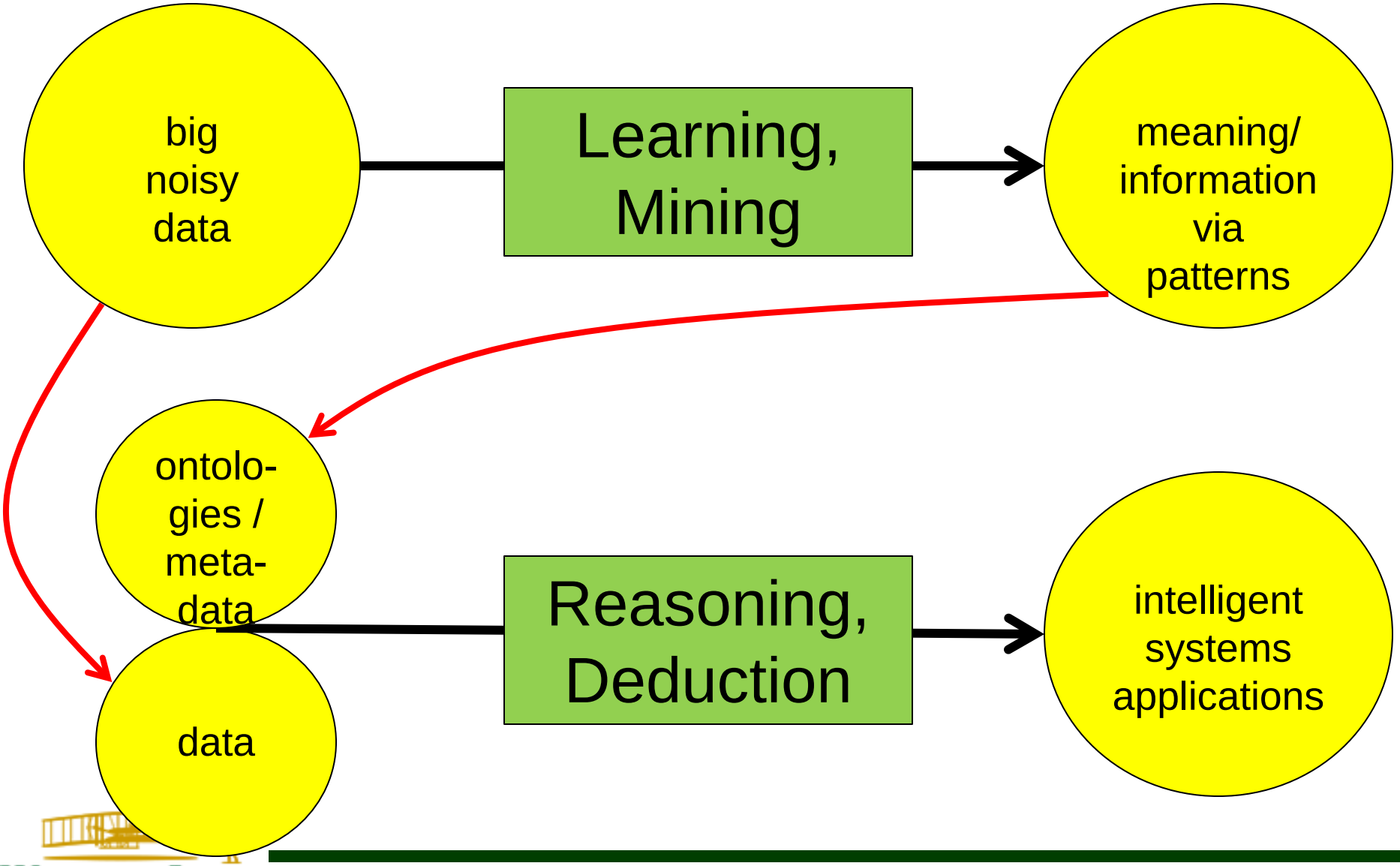
Combining Learning and Reasoning for Big Data

Pascal Hitzler
DaSe Lab for Data Semantics
Wright State University
http://www.pascal-hitzler.de

The Big Data Added Value Pipeline



The Big Data Added Value Pipeline



Structured data on the Web

DaSe Lab

e.g. schema.org, introduced 2011 by Bing, Google, Yahoo, Yandex:

“[In 2014, in] a sample of over 12 billion web pages, 21 percent, or 2.5 billion pages, use [schema.org] to mark up HTML pages, to the tune of more than 15 billion entities and more than 65 billion triples”

“Just about every major site in every major category, from news to e-commerce (with the exception of Amazon.com), uses it”

Source: http://semanticweb.com/schema-org-fires-lit_b44380

Linked Data

DaSe Lab

A bit older but somewhat more expressive: Linked Data on the Web

Number of Datasets

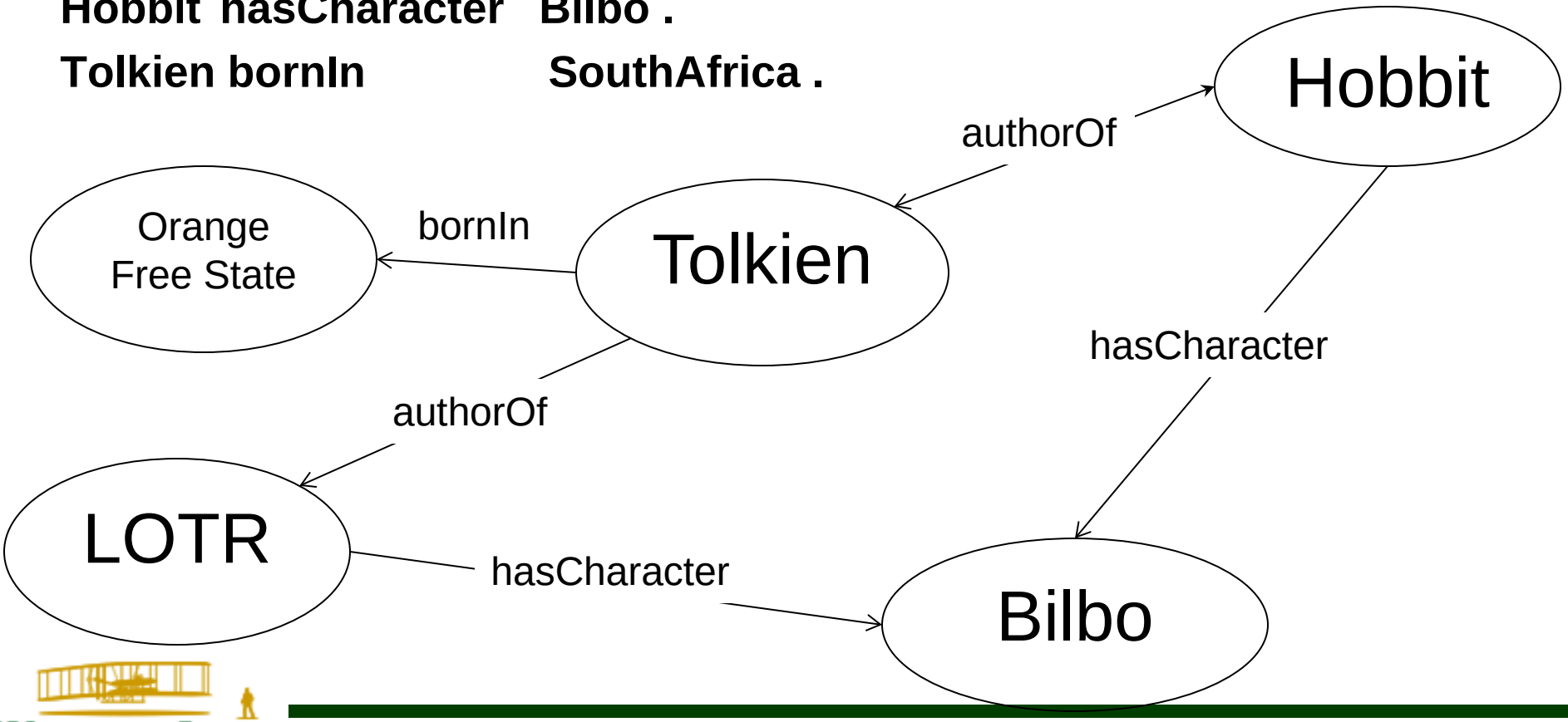
2014-08-30	**570**
2011-09-19	**295**
2010-09-22	**203**
2009-07-14	**95**
2008-09-18	**45**
2007-10-08	**25**
2007-05-01	**12**

from ca. 200 crawlable datasets alone:
ca. 200 million triples

Information as RDF graph

DaSe Lab

LOTR hasAuthor Tolkien .
Hobbit hasAuthor Tolkien .
LOTR hasCharacter Bilbo .
Hobbit hasCharacter Bilbo .
Tolkien bornIn SouthAfrica .



DBpedia: LOTR page

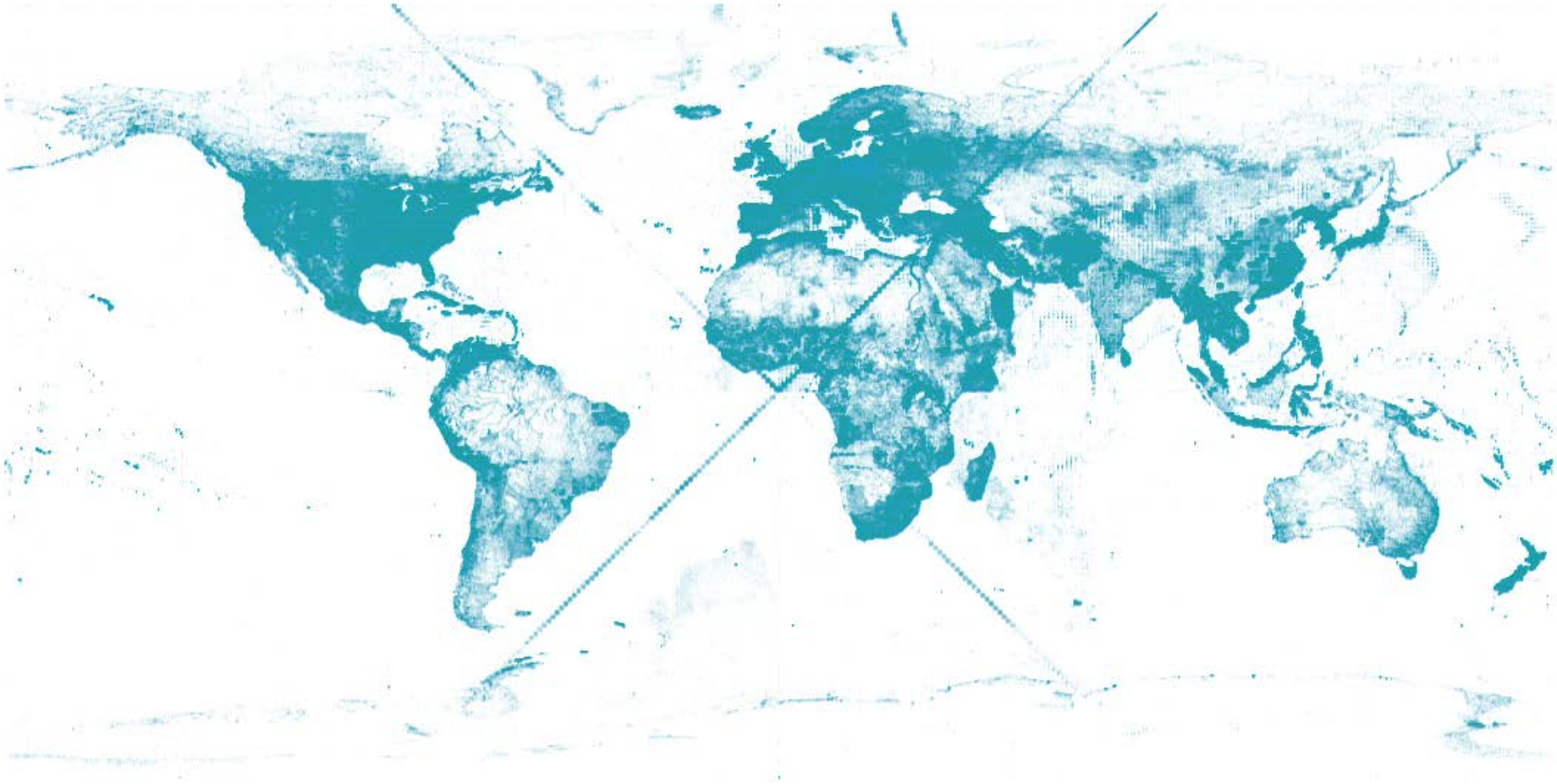
dbpedia-owl:thumbnail	▪ http://upload.wikimedia.org/wikipedia/commons/thumb/6/62/Jrrt_lotr_cover_design.jpg/200px-Jrrt_lotr_cover_design.jpg
dbpedia-owl:wikiPageExternalLink	▪ http://lotr.wikia.com ▪ http://www.glyphweb.com/arda/ ▪ http://www.tolkienlibrary.com/ ▪ http://www.tolkien.co.uk/ ▪ http://www.houghtonmifflinbooks.com/features/lordoftheringstrilogy/
dbpprop:author	▪ dbpedia:J._R._R._Tolkien
dbpprop:books	▪ dbpedia:The_Two_Towers ▪ dbpedia:The_Return_of_the_King ▪ dbpedia:The_Fellowship_of_the_Ring ▪ "Volumes:"
dbpprop:country	▪ England
dbpprop:expiry	▪ 20 (xsd:integer)
dbpprop:genre	▪ dbpedia:Adventure_novel ▪ dbpedia:High_fantasy
dbpprop:hasPhotoCollection	▪ http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/The_Lord_of_the_Rings
dbpprop:imageCaption	▪ Tolkien's own cover designs for the three volumes
dbpprop:language	▪ English
dbpprop:mediaType	▪ Print
dbpprop:name	▪ The Lord of the Rings
dbpprop:pages	▪ 1216 (xsd:integer)
dbpprop:precededBy	▪ dbpedia:The_Hobbit
dbpprop:pubDate	▪ 21 (xsd:integer)
dbpprop:publisher	▪ dbpedia:Allen_&_Unwin
dbpprop:small	▪ yes
dbpprop:wikiPageUsesTemplate	▪ dbpedia:Template:Infobox_book_series ▪ dbpedia:Template:Pp-vandalism
dcterms:subject	▪ category:Monomyths ▪ category:High_fantasy_novels ▪ category:Middle-earth_books ▪ category:British_fantasy_novels ▪ category:Fantasy_books_by_series ▪ category:1950s_fantasy_novels ▪ category:Sequel_novels ▪ category:The_Lord_of_the_Rings ▪ category:English_novels

Linked Data: Volume

DaSe Lab

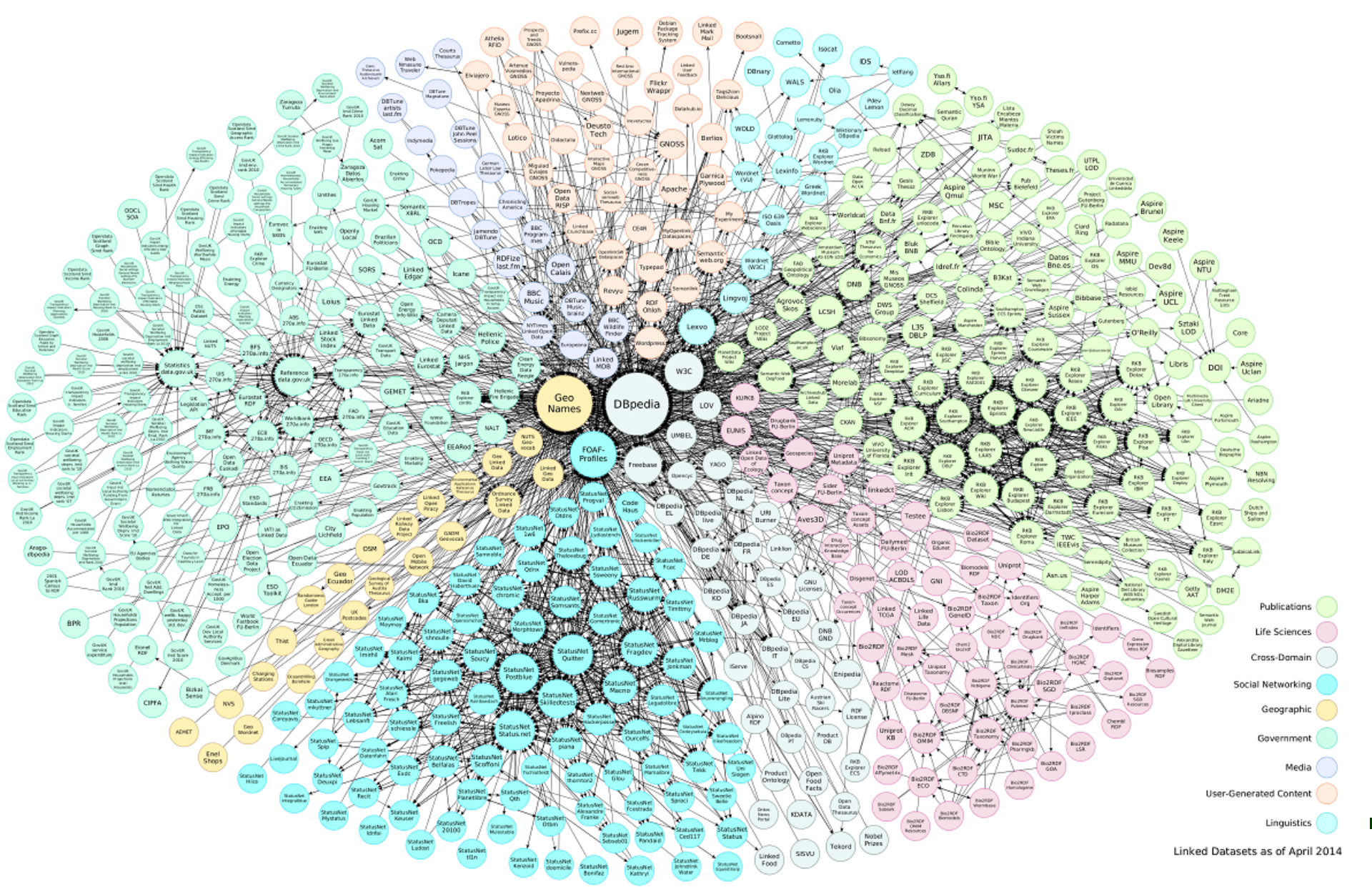
Geoindexed Linked Data – courtesy of Krzysztof Janowicz

http://stko.geog.ucsb.edu/location_linked_data

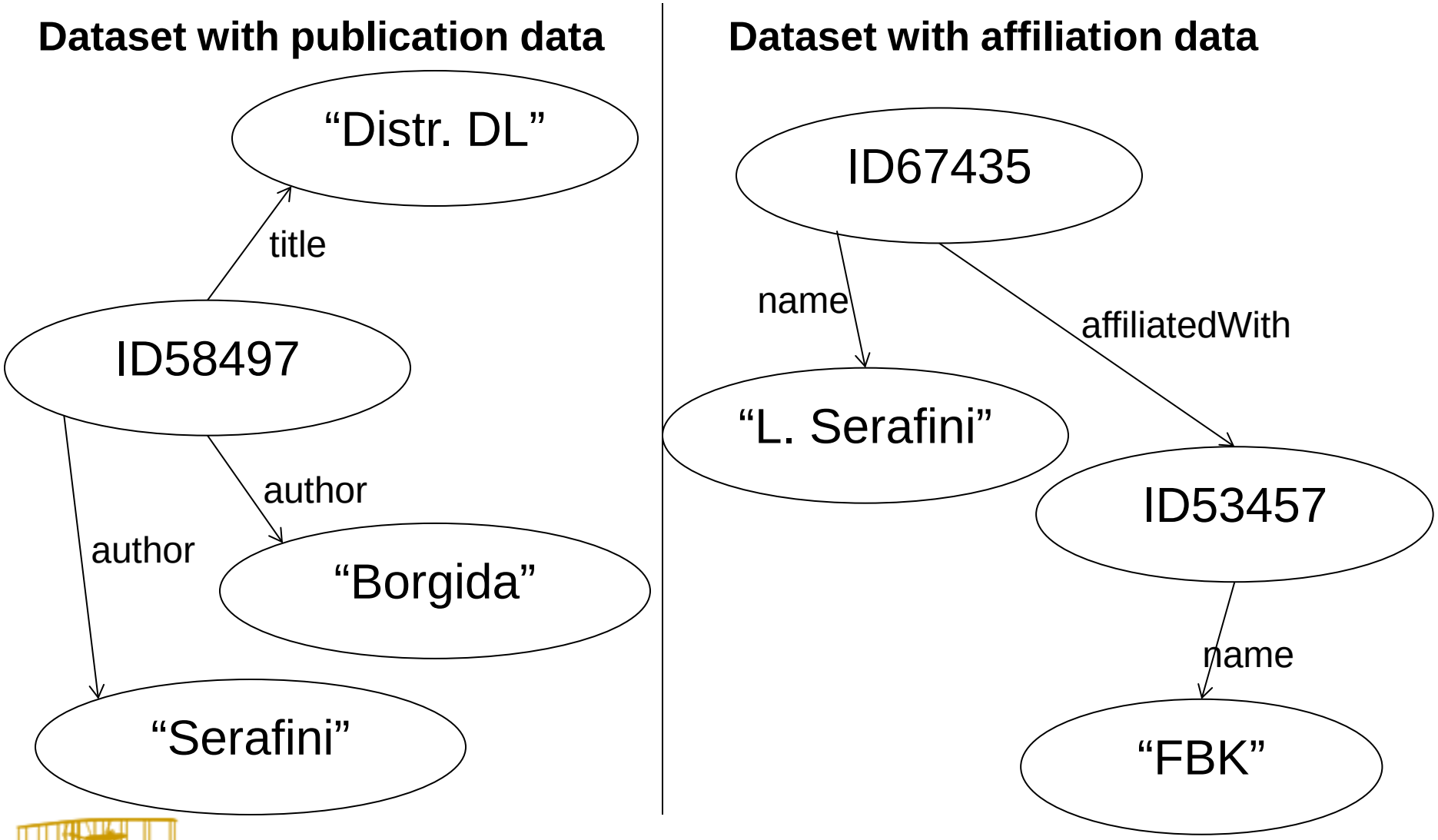


Some Linked Datasets 2014

DaSe Lab



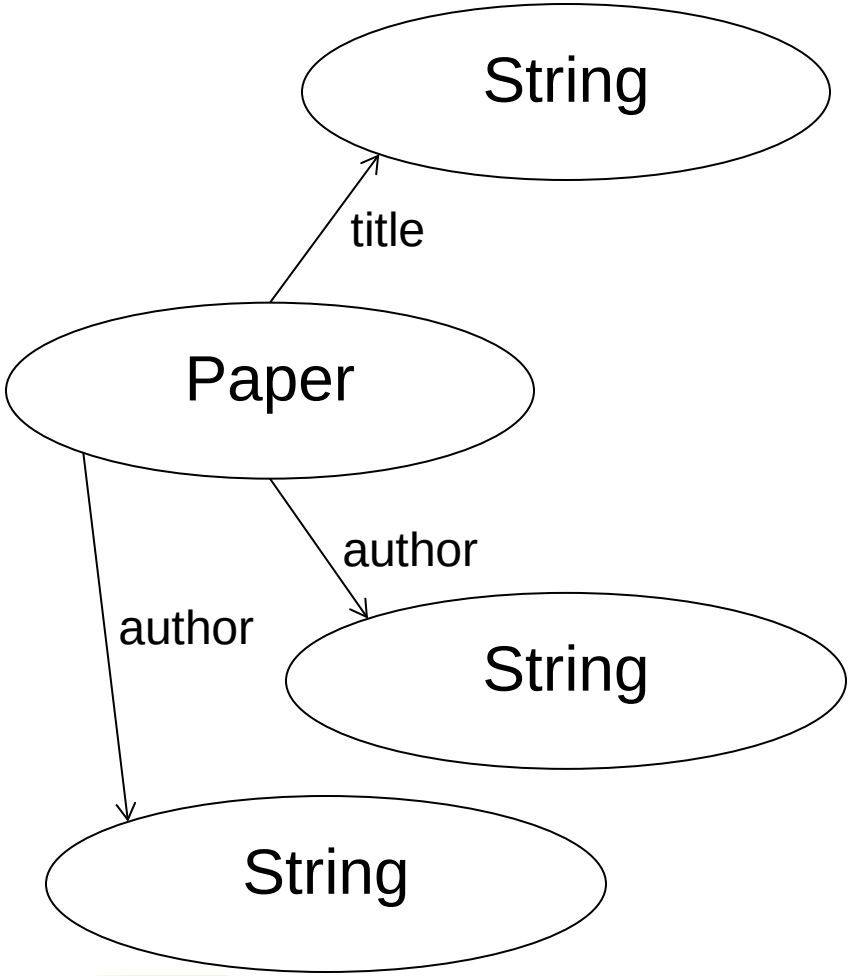
Find all FBK publications



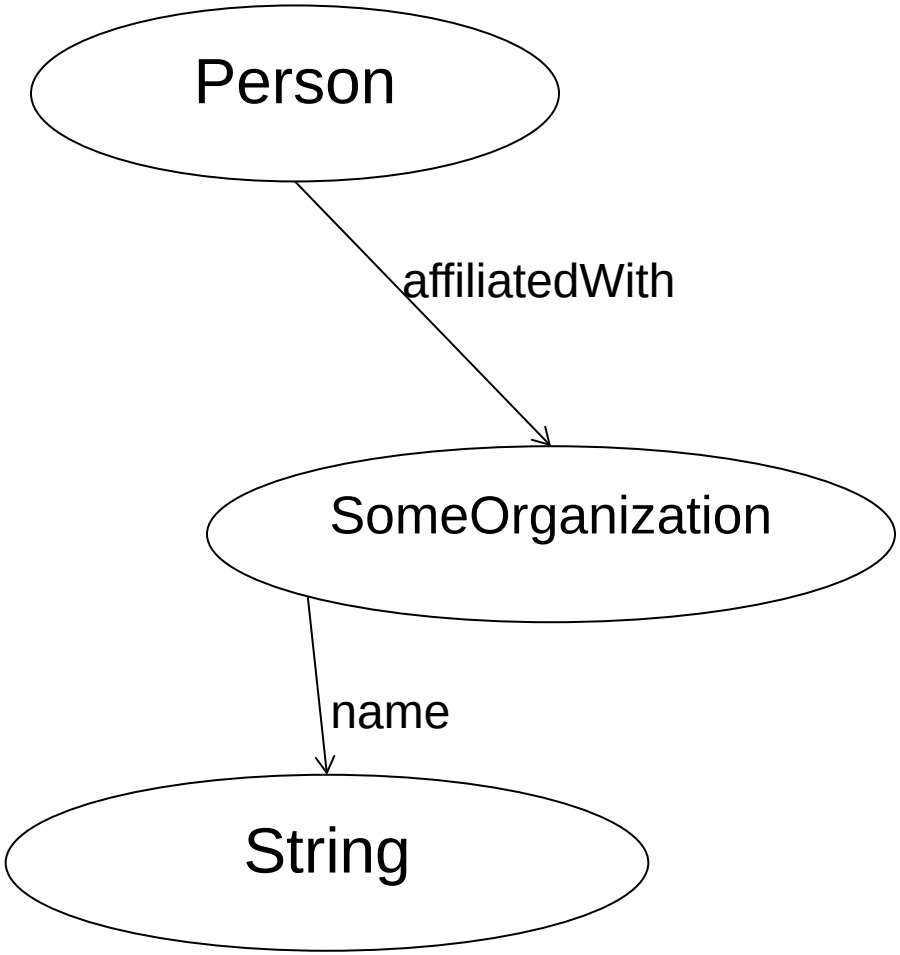
Find all FBK publications

DaSe Lab

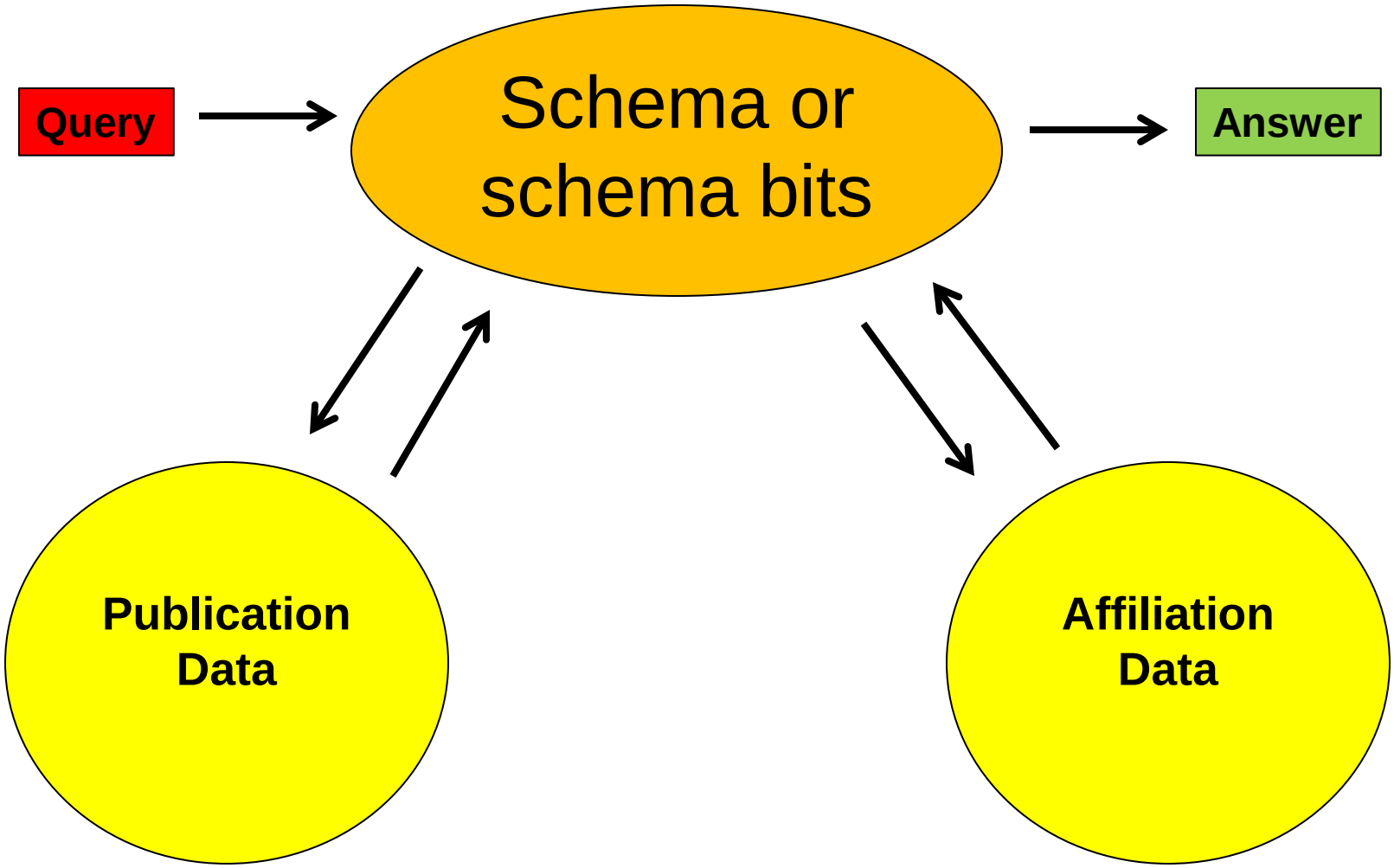
Dataset with publication data



Dataset with affiliation data



Linked Data federated querying



Joshi, Jain, Hitzler et al. ODBASE 2012

Schema Alignment

DaSe Lab

Automatic acquisition of relationships, e.g.

$$\forall x(\text{Paper}(x) \rightarrow \text{Publication}(x))$$

$$\forall x(\text{Person}(x) \rightarrow \text{Agent}(x))$$

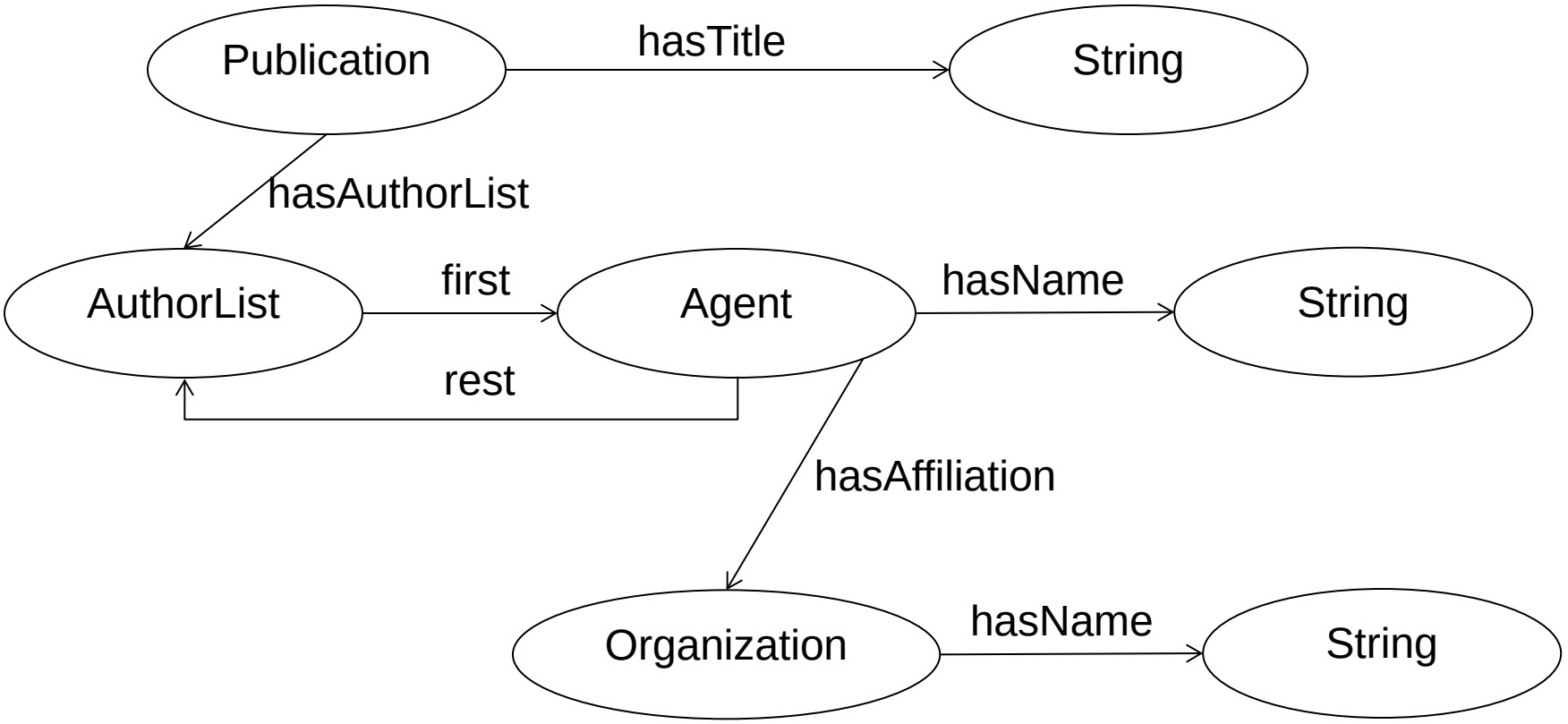
$$\forall x(\text{SomeOrganization}(x) \longleftrightarrow \text{Organization}(x))$$

$$\forall x \forall y(\text{title}(x, y) \rightarrow \text{hasName}(x, y))$$

$$\forall x \forall y(\text{name}(x, y) \rightarrow \text{hasName}(x, y))$$

$$\forall x \forall y(\text{affiliatedWith}(x, y) \longleftrightarrow \text{hasAffiliation}(x, y))$$

Overarching schema



Automated Alignments

DaSe Lab

Most systems can do only equivalences.

$$\forall x(\text{SomeOrganization}(x) \longleftrightarrow \text{Organization}(x))$$

$$\forall x \forall y(\text{affiliatedWith}(x, y) \longleftrightarrow \text{hasAffiliation}(x, y))$$

Moreover, almost all of the performance of current systems is based on string similarity metrics.

[Cheatham and Hitzler, ISWC 2013]

Table 1. Results of strings only approaches and the competitors from the OAEI 2012 competition on the conference data set (left) and the anatomy data set (right)

Metric	Prec.	Recall	F-meas.
YAM++	0.81	0.69	0.75
LogMap	0.82	0.58	0.68
StringsOpt	**0.85**	**0.55**	**0.67**
StringsAuto	**0.79**	**0.57**	**0.66**
Optima	0.62	0.68	0.65
CODI	0.74	0.57	0.64
GOMMA	0.85	0.47	0.61
Wmatch	0.74	0.50	0.60
WeSeE	0.76	0.49	0.60
Hertuda	0.74	0.50	0.60
MaasMatch	0.63	0.57	0.60
LogMapLt	0.73	0.50	0.59
HotMatch	0.71	0.51	0.59
Baseline 2	0.79	0.47	0.59
ServOMap	0.73	0.46	0.56
Baseline 1	0.80	0.43	0.56
ServOMapLt	0.88	0.40	0.55
MEDLEY	0.54	0.50	0.52
ASE	0.63	0.43	0.51
MapSSS	0.50	0.51	0.50
AUTOMSv2	0.67	0.36	0.47
AROMA	0.33	0.48	0.39

Metric	Prec.	Recall	F-meas.
GOMMA-bk	0.92	0.93	0.92
YAM++	0.94	0.86	0.90
CODI	0.97	0.83	0.89
StringsOpt	**0.88**	**0.87**	**0.88**
LogMap	0.92	0.85	0.88
GOMMA	0.96	0.80	0.87
StringsAuto	**0.86**	**0.84**	**0.85**
MapSSS	0.94	0.75	0.83
WeSeE	0.91	0.76	0.83
LogMapLt	0.96	0.73	0.83
TOAST*	0.85	0.76	0.80
ServOMap	1.00	0.64	0.78
ServOMapLt	0.99	0.64	0.78
HotMatch	0.98	0.64	0.77
AROMA	0.87	0.69	0.77
StringEquiv	1.00	0.62	0.77
Wmatch	0.86	0.68	0.76
Optima	0.85	0.58	0.69
Hertuda	0.69	0.67	0.68
MaasMatch++	0.43	0.78	0.56

Ontology Alignment

Linked Data Alignment seems to be quite different from established benchmark problems.

Plus, we need at least subsumption mapping.

$$\forall x(\text{Paper}(x) \rightarrow \text{Publication}(x))$$

$$\forall x(\text{Person}(x) \rightarrow \text{Agent}(x))$$

$$\forall x \forall y(\text{title}(x, y) \rightarrow \text{hasName}(x, y))$$

$$\forall x \forall y(\text{name}(x, y) \rightarrow \text{hasName}(x, y))$$

[Jain, Hitzler et al. ISWC 2010]

Bootstrapping-based alignment

DaSe Lab

Table 4. Results of various systems for LOD Schema Alignment. Legends: Prec=Precision, Rec=Recall, M=Music Ontology, B=BBC Program Ontology, F=FOAF Ontology, D=DBpedia Ontology, G=Geonames Ontology, S=SIOC Ontology, W=Semantic Web Conference Ontology, A=AKT Portal Ontology, err=System Error, NA=Not Available

Linked Open Data Schema Ontology Alignment												
	Alignment API		OMViaUO		RiMoM		S-Match		AROMA		BLOOMS	
Test	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
M,B	0.4	0	1	0	err	err	0.04	0.28	0	0	0.63	0.78
M,D	0	0	0	0	err	err	0.08	0.30	0.45	0.01	0.39	0.62
F,D	0	0	0	0	err	err	0.11	0.40	0.33	0.04	0.67	0.73
G,D	0	0	0	0	err	err	0.23	1	0	0	0	0
S,F	0	0	0	0	0.3	0.2	0.52	0.11	0.30	0.20	0.55	0.64
W,A	0.12	0.05	0.16	0.03	err	err	0.06	0.4	0.38	0.03	0.42	0.59
W,D	0	0	0	0	err	err	0.15	0.50	0.27	0.01	0.70	0.40
Avg.	0.07	0.01	0.17	0	NA	NA	0.17	0.43	0.25	0.04	0.48	0.54

[Jain, Hitzler et al, ISWC2010]

Property subsumption

DaSe Lab

Detecting property subsumptions is still in infancy.

$$\forall x \forall y(\text{title}(x, y) \rightarrow \text{hasName}(x, y))$$

$$\forall x \forall y(\text{name}(x, y) \rightarrow \text{hasName}(x, y))$$

[Cheatham, Hitzler OM 2014]

System	Class Prec	Class Rec	Class Fms	Prop Prec	Prop Rec	Prop Fms
AML	0.86	0.62	0.72	1.00	0.20	0.33
AMLback	0.86	0.64	0.73	1.00	0.24	0.39
CIDER_CL	0.46	0.59	0.52	0.07	0.22	0.11
HerTUDA	0.84	0.56	0.67	0.26	0.20	0.23
HotMatch	0.81	0.57	0.67	0.24	0.20	0.22
IAMA	0.87	0.55	0.67	0.14	0.07	0.09
LogMap	0.82	0.65	0.73	0.62	0.28	0.39
			0.66	0.00	0.00	0.00
			0.67	0.32	0.26	0.29
			0.73	0.32	0.26	0.29
			0.69	0.00	0.00	0.00
			0.67	0.00	0.00	0.00
			0.66	0.50	0.02	0.04
			0.66	0.26	0.22	0.24
YAM++	0.82	0.71	0.76	0.68	0.57	0.62
Average	0.79	0.60	0.68	0.36	0.18	0.21

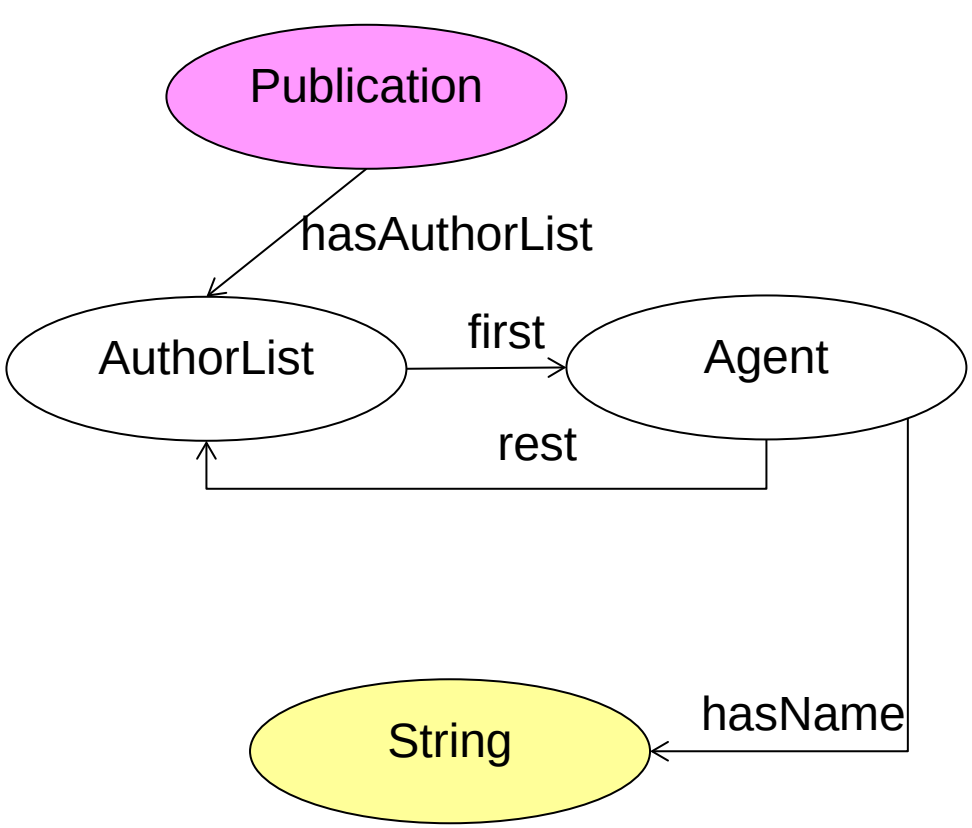
Table 1. Performance of the top 2013 OAEI competitors on classes versus prope

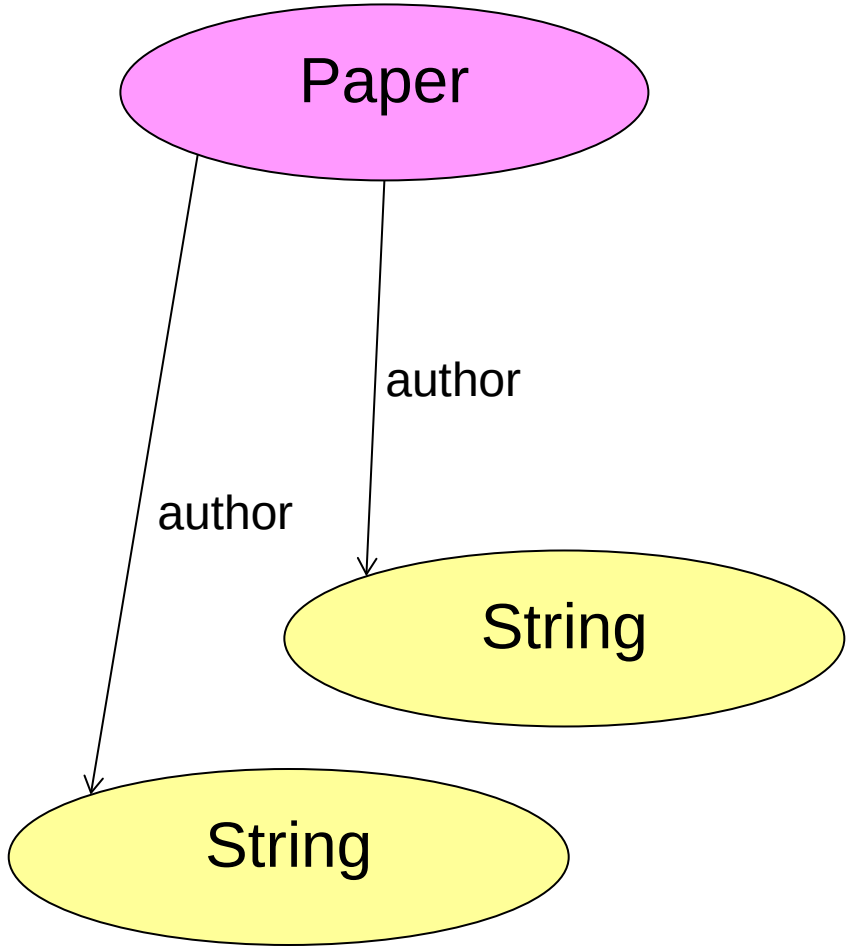
System	Precision	Recall	F-measure
PropString (prec)	1.0	0.26	0.41
PropString (rec)	0.34	0.5	0.4
Soft TF-IDF	0.2	0.24	0.22

Table 4. Results on the OAEI Conference track

WRIGHT STA
UNIVERSIT

We need mapping *rules*

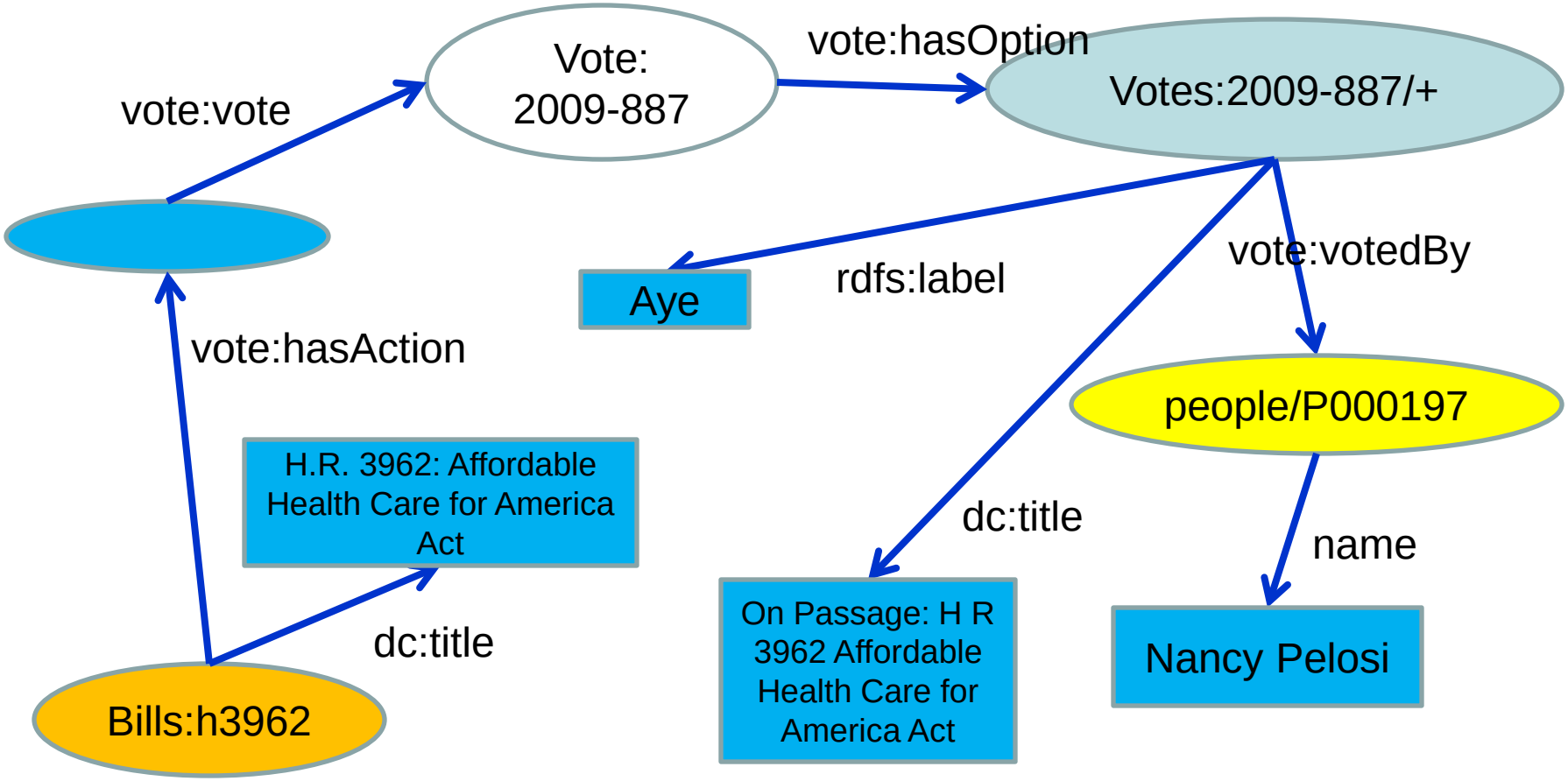




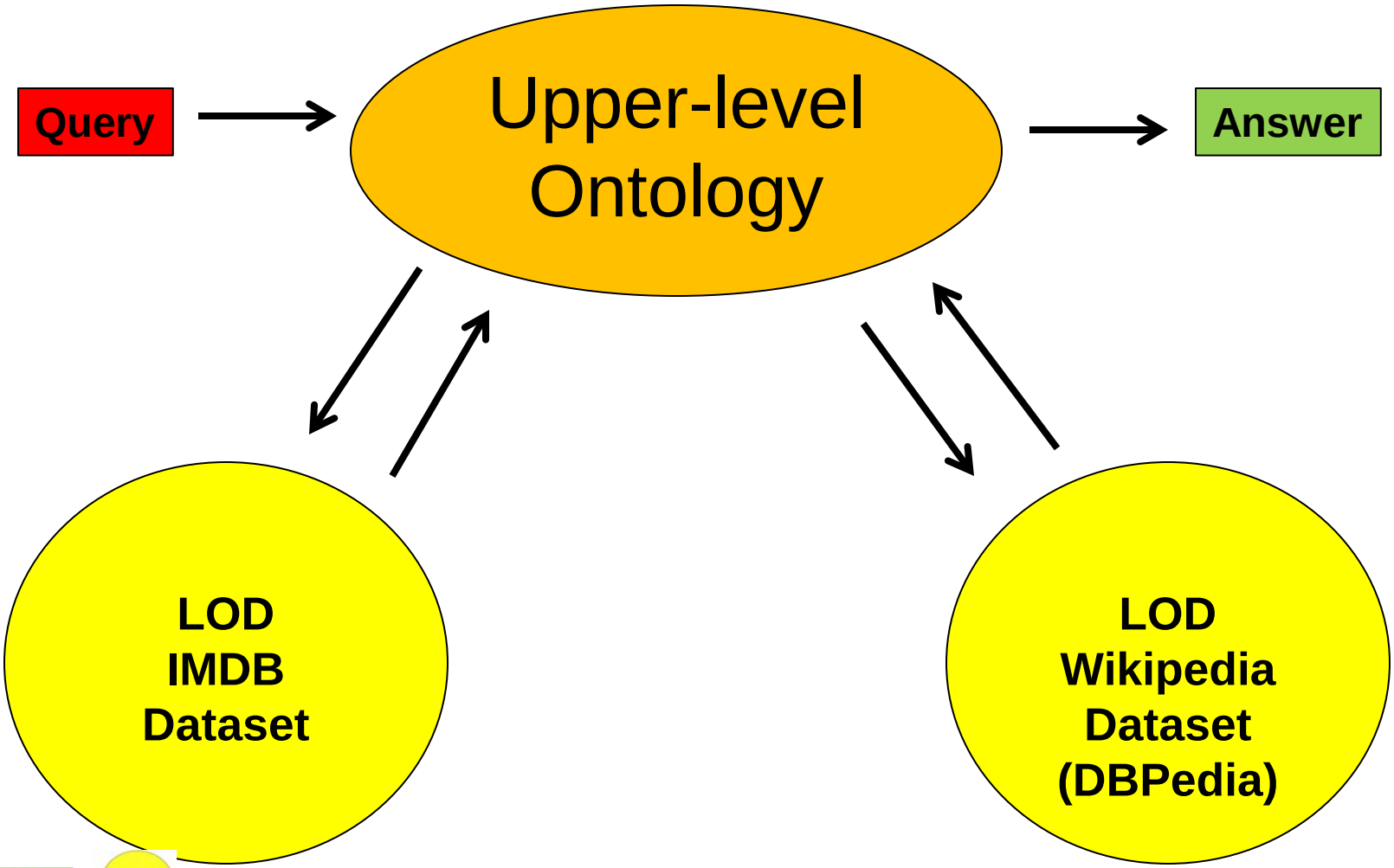
Linked Data: Variety

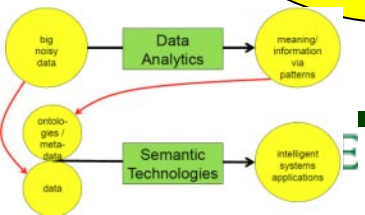
DaSe Lab

“Nancy Pelosi voted in favor of the Health Care Bill.”



Linked Data federated querying





Joshi, Jain, Hitzler et al. ODBASE 2012

Thanks!

References

- Pascal Hitzler, Frank van Harmelen, A reasonable Semantic Web. Semantic Web 1 (1-2), 39-44, 2010.
- Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, Amit P. Sheth, Linked Data is Merely More Data. In: Dan Brickley, Vinay K. Chaudhri, Harry Halpin, Deborah McGuinness: Linked Data Meets Artificial Intelligence. Technical Report SS-10-07, AAAI Press, Menlo Park, California, 2010, pp. 82-86. ISBN 978-1-57735-461-1. Proceedings of LinkedAI at the AAAI Spring Symposium, March 2010.
- Pascal Hitzler, Krzysztof Janowicz, *What's Wrong with Linked Data?* http://blog.semantic-web.at/2012/08/09/whats-wrong-with-linked-data/ , August 2012.
- Pascal Hitzler, Markus Krötzsch, Sebastian Rudolph, *Foundations of Semantic Web Technologies*. Chapman and Hall/CRC Press, 2009.

References

- Pascal Hitzler, Krzysztof Janowicz, Linked Data, Big Data, and the 4th Paradigm. Semantic Web 4 (3), 2013, 233-235.
- Krzysztof Janowicz, Pascal Hitzler, The Digital Earth as Knowledge Engine. Semantic Web 3 (3), 213-221, 2012.
- Krzysztof Janowicz, Pascal Hitzler, Thoughts on the Complex Relation Between Linked Data, Semantic Annotations, and Ontologies. In: Paul N. Bennett, Evgeniy Gabrilovich, Jaap Kamps, Jussi Karlgren (eds.), Proceedings of the 6th International Workshop on Exploiting Semantic Annotation in Information Retrieval, ESAIR 2013, ACM, San Francisco, 2013, pp. 41-44.
- Krzysztof Janowicz, Frank van Harmelen, James A. Hendler, Pascal Hitzler, Why the Data Train Needs Semantic Rails. AI Magazine. To appear.

References

- Prateek Jain, Pascal Hitzler, Amit P. Sheth, Kunal Verma, Peter Z. Yeh, Ontology Alignment for Linked Open Data. In P. Patel-Schneider, Y. Pan, P. Hitzler, P. Mika, L. Zhang, J. Pan, I. Horrocks, B. Glimm (eds.), The Semantic Web - ISWC 2010. 9th International Semantic Web Conference, ISWC 2010, Shanghai, China, November 7-11, 2010, Revised Selected Papers, Part I. Lecture Notes in Computer Science Vol. 6496. Springer, Berlin, 2010, pp. 402-417.
- Amit Krishna Joshi, Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, Amit P. Sheth, Mariana Damova, Alignment-based Querying of Linked Open Data. In: Meersman, R.; Panetto, H.; Dillon, T.; Rinderle-Ma, S.; Dadam, P.; Zhou, X.; Pearson, S.; Ferscha, A.; Bergamaschi, S.; Cruz, I.F. (eds.), On the Move to Meaningful Internet Systems: OTM 2012, Confederated International Conferences: CoopIS, DOA-SVI, and ODBASE 2012, Rome, Italy, September 10-14, 2012, Proceedings, Part II. Lecture Notes in Computer Science Vol. 7566, Springer, Heidelberg, 2012, pp. 807-824.

References

- Prateek Jain, Peter Z. Yeh, Kunal Verma, Reymonrod G. Vasquez, Mariana Damova, Pascal Hitzler, Amit P. Sheth, Contextual Ontology Alignment of LOD with an Upper Ontology: A Case Study with Proton. In: Grigoris Antoniou, Marko Grobelnik, Elena Paslaru Bontas Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter De Leenheer, Jeff Pan (Eds.): The Semantic Web: Research and Applications - 8th Extended Semantic Web Conference, ESWC 2011, Heraklion, Crete, Greece, May 29-June 2, 2011, Proceedings, Part I. Lecture Notes in Computer Science 6643, Springer, 2011, pp. 80-92.
- Prateek Jain, Pascal Hitzler, Kunal Verma, Peter Yeh, Amit Sheth, Moving beyond sameAs with PLATO: Partonomy detection for Linked Data. In: Ethan V. Munson, Markus Strohmaier (Eds.): 23rd ACM Conference on Hypertext and Social Media, HT '12, Milwaukee, WI, USA, June 25-28, 2012. ACM, 2012, pp. 33-42.

References

- Michelle Cheatham, Pascal Hitzler, String Similarity Metrics for Ontology Alignment. In: H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J.X. Parreira, L. Aroyo, N. Noy, C. Welty, K. Janowicz (eds.), The Semantic Web - ISWC 2013. 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II. Lecture Notes in Computer Science Vol. 8219, Springer, Heidelberg, 2013, pp. 294-309.
- Michelle Cheatham, Pascal Hitzler, The Properties of Property Alignment. In: Proceedings OM-2014, The Ninth International Workshop on Ontology Matching, at the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Trentino, Italy, October 2014. To appear.