11-2010

Ontology Alignment for Linked Open Data

Prateek Jain
Wright State University - Main Campus

Pascal Hitzler
pascal.hitzler@wright.edu

Amit P. Sheth
Wright State University - Main Campus, amit@sc.edu

Kunal Verma
Wright State University - Main Campus

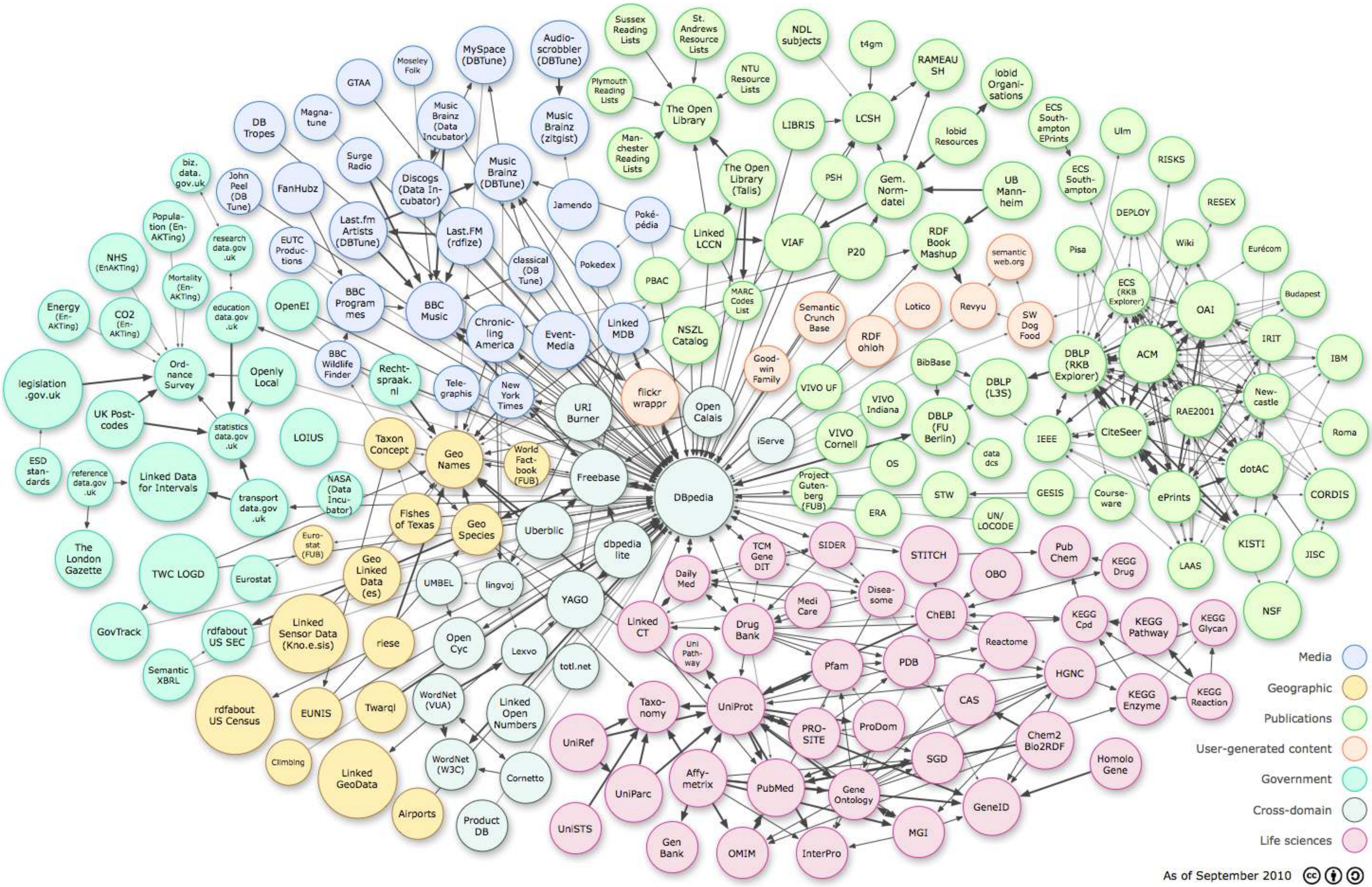
Peter Z. Yeh

Ontology Alignment for Linked Open Data
– ISWC2010 research track

Prateek Jain
Pascal Hitzler
Amit Sheth
Kno.e.sis Center
Wright State University, Dayton, OH

Kunal Verma
Peter Z. Yeh
Accenture Technology Labs
San Jose, CA

Linked Open Data



Outline

- Introduction
- Motivation
- Existing Approaches
- BLOOMS Approach
- Evaluation
- Applications
- Conclusion & Future Work
- References

BLOOMS

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		
M,D	0	0	0	0	err	err	0.08	0.30	0.45	0.01		
F,D	0	0	0	0	err	err	0.11	0.40	0.33	0.04		
G,D	0	0	0	0	err	err	0.23	1	0	0		
S,F	0	0	0	0	0.3	0.2	0.52	0.11	0.30	0.20		
W,A	0.12	0.05	0.16	0.03	err	err	0.06	0.4	0.38	0.03		
W,D	0	0	0	0	err	err	0.15	0.50	0.27	0.01		
Avg.	0.07	0.01	0.17	0	NA	NA	0.17	0.43	0.25	0.04		

BLOOMS

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

When I was 6 years old…

1990

May

Same proposal recirculated

September

Mike Sendall, Tim's boss, Oks the purchase of a NeXT cube, and allows Tim to go ahead and write a global hypert
system.

October

Tim starts work on a hypertext GUI browser+editor using the NeXTStep development environment. He makes up "WorldWideWeb" as a name for the program. (See the first browser screenshot) "World Wide Web" as a name fo project (over Information Mesh, Mine of Information, and Information Mine).

Project original proposal reformulated with encouragement from CN and ECP divisional management. Robert Cailli (ECP) joins and is co-author of new version.

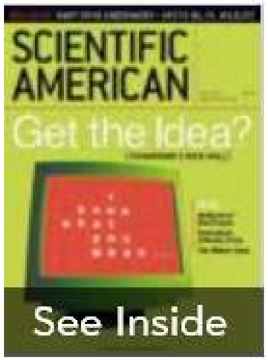
November

Initial WorldWideWeb program development continues on the NeXT (TBL) . This was a "what you see is what yo get" (wysiwyg) browser/editor with direct inline creation of links. The first web server was nxoc01.cern.ch, later ca info.cern.ch, and the **first web page** http://nxoc01.cern.ch/hypertext/WWW/TheProject.html Unfortunately CERN longer supports the historical site. Note from this era too, the least recently modified web page we know of, last cha Tue, 13 Nov 1990 15:17:00 GMT (though the URL changed.)

11 years later…

Home » Scientific American Magazine » May 2001

Feature Articles |



The Semantic Web

A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities

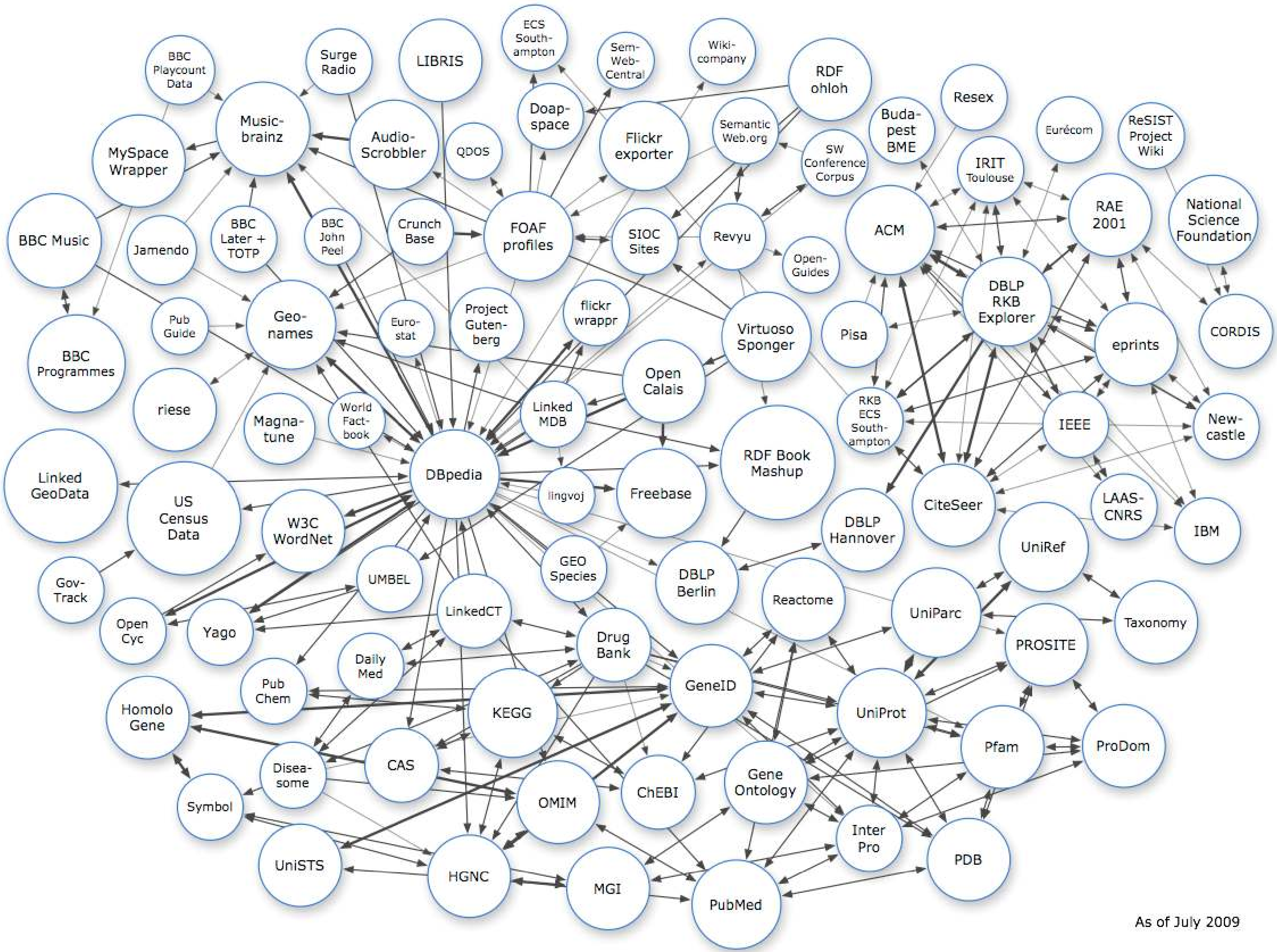
By Tim Berners-Lee, James Hendler and Ora Lassila | May 17, 2001 | 10

Share Email Print

Happy 20th Birthday, World Wide Web

CERN on March 13 celebrates the 20th anniversary of a proposal entitled, "Information Management: A Proposal," by Tim Berners-Lee, which would become the blueprint for the World Wide Web »

In 2006 Web of Data



Is it really mainstream Semantic Web?

- **What is the relationship between the models whose instances are being linked?**
- **How to do querying on LOD without knowing individual datasets?**
- **How to perform schema level reasoning over LOD cloud?**

What can be done?

- **Relationships are at the heart of Semantics.**
- **LOD captures instance level relationships, but lacks class level relationships.**
  - Superclass
  - Subclass
  - Equivalence
- **How to find these relationships?**
  - Perform a matching of the LOD Ontology's using state of the art ontology matching tools.
- **Desirable**
  - Considering the size of LOD, at least have results which a human can curate.

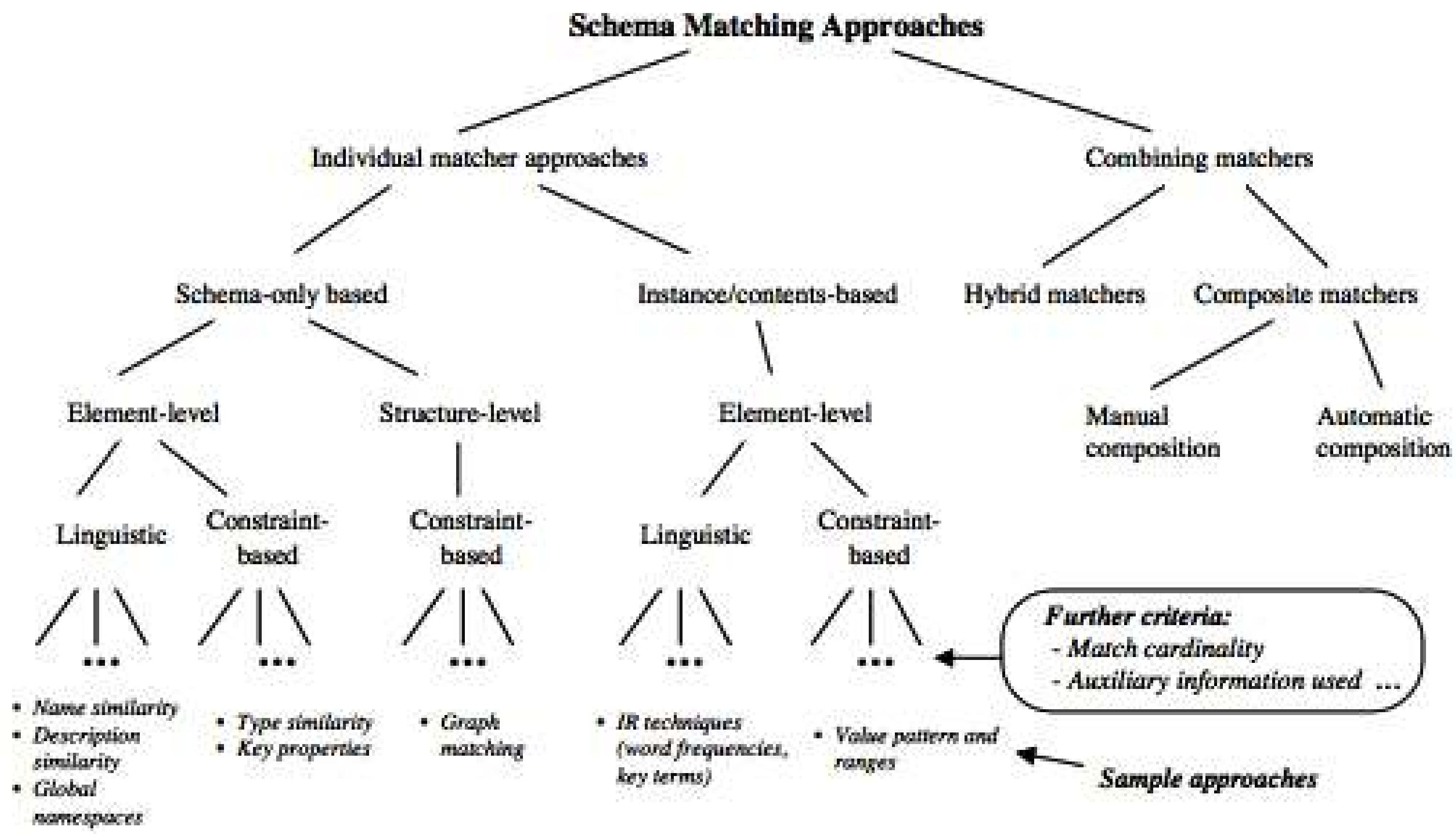
Outline

Kno.e.sis

- Introduction
- Motivation
- Existing Approaches
- BLOOMS Approach
- Evaluation
- Applications
- Conclusion & Future Work
- References

Wright State University

Existing Approaches



A survey of approaches to automatic Ontology matching by Erhard Rahm, Philip A. Bernstein in the VLDB Journal 10: 334–350 (2001)

LOD Ontology Alignment

- **Existing systems have difficulty in matching LOD Ontologys!**
  - Nation = Menstruation, Confidence=0.9 ☺

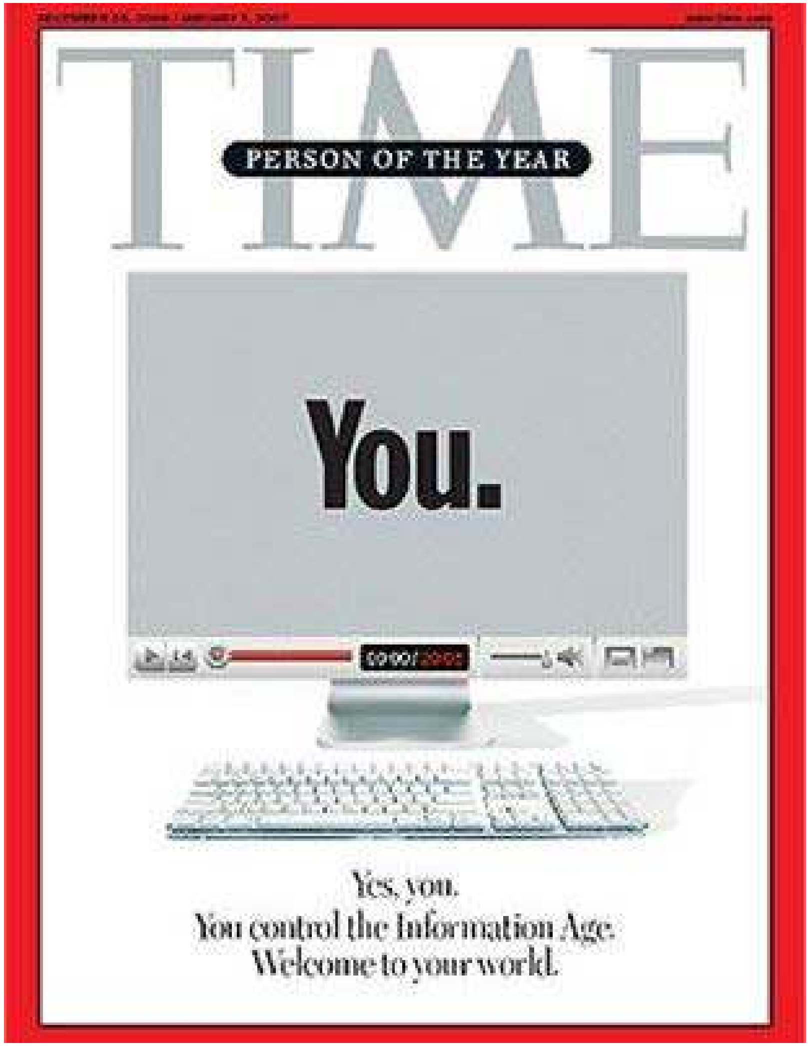
- **They perform extremely well on established benchmarks, but typically not in the wilds.**

- **LOD Ontology's are of very different nature**
  - Created by community for community.
  - Emphasis on number of instances, not number of meaningful relationships.
  - Require solutions beyond syntactic and structural matching.

Outline

- Introduction
- Motivation
- Existing Approaches
- BLOOMS Approach
- Evaluation
- Applications
- Conclusion & Future Work
- References

Something else changed..



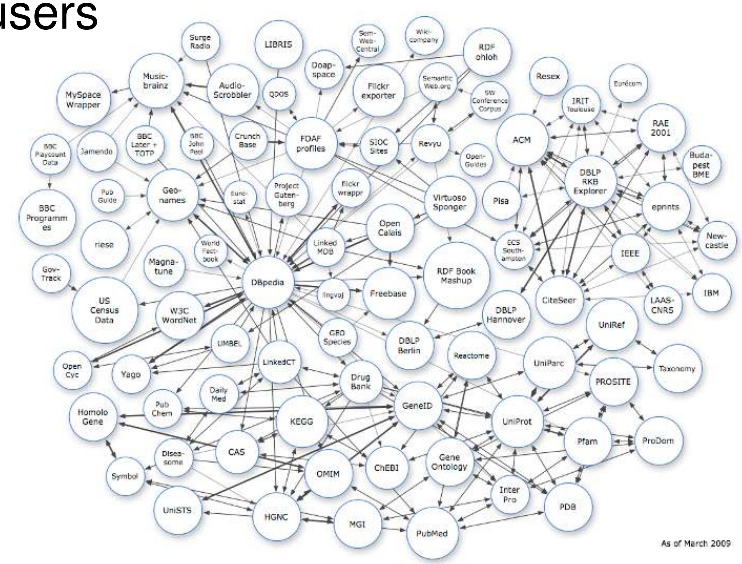

WEB 2.0 Landscape
WEB APPLICATION
Widget/ component
Aggregation/ recombination
CONTENT SHARING
RECOMMENDATIONS/ FILTERING
Rating/ Tagging
Collaborative filtering
SOCIAL NETWORK
FeedBurner
Jotspot
ZOHO
jumpcut
remember the milk
Blogger
pbwiki
Basecamp
Ning
thinkfree
Bloglines
netvibes
snap
Rojo
HYPE MACHINE
Zillow.com
trulia
PANDORA
Findory
ourmedia
edgeio
YouTube
imeem
LinkedIn
flickr
LIFE
meebo
FURL
myspace.com
del.icio.us
Frappr!
blinx.com
reddit
PLAZES
Technorati
Opinity
SHOUTWIRE
digg
last.fm
ChaCha
ROLLYO
FUTURE EXPLORATION NETWORK
www.futureexploration.net

Our Approach

Use knowledge contributed by users

To improve

Structured knowledge contributed by users



Rabbit out of a hat?

- **Traditional auxiliary data sources like (WordNet, Upper Level Ontologies) have limited coverage and are insufficient for LOD datasets.**
  - LOD datasets have diverse domains
- **Community generated data although noisy but is rich in**
  - Content
  - Structure
  - Has a “self healing property”
- **Problems like Ontology Matching have a dimension of context associated with them. Since community generated data is created by diverse set of people, hence captures diverse context.**

Wikipedia

- **The English version alone contains more than 2.9 million articles.**

- **It is continually expanded by approximately 100,000 active volunteer editors world-wide.**

- **Allows multiple points of view to be mentioned with their proper contexts.**

- **Article creation/correction is an ongoing activity with no down time.**

Ontology Matching on LOD using Wikipedia Categorization

Kno.e.sis

- On Wikipedia, categories are used to organize the entire project.
- Wikipedia's category system consists of overlapping trees.
- Simple rules for categorization
  – "If logical membership of one category implies logical membership of a second, then the first category should be made a subcategory"
  – "Pages are not placed directly into every possible category, only into the most specific one in any branch"
  – "Every Wikipedia article should belong to at least one category."

BLOOMS Approach – Step 1

- **Pre-process the input ontology**
  - Remove property restrictions
  - Remove individuals, properties

- **Tokenize the class names**
  - Remove underscores, hyphens and other delimiters
  - Breakdown complex class names
    - example: SemanticWeb => Semantic Web

BLOOMS Approach – Step 2

- **For each concept name processed in the previous step**
  - Identify article in Wikipedia corresponding to the concept.
  - Each article related to the concept indicates a sense of the usage of the word.
- **For each article found in the previous step**
  - Identify the Wikipedia category to which it belongs.
  - For each category found, find its parent categories till level 4.
- **Once the "BLOOMS tree" for each of the sense of the source concept is created (T_s), utilize it for comparison with the "BLOOMS tree" of the target concepts (T_t).**
  - BLOOMS trees are created for individual senses of the concepts.

BLOOMS Approach – Step 3

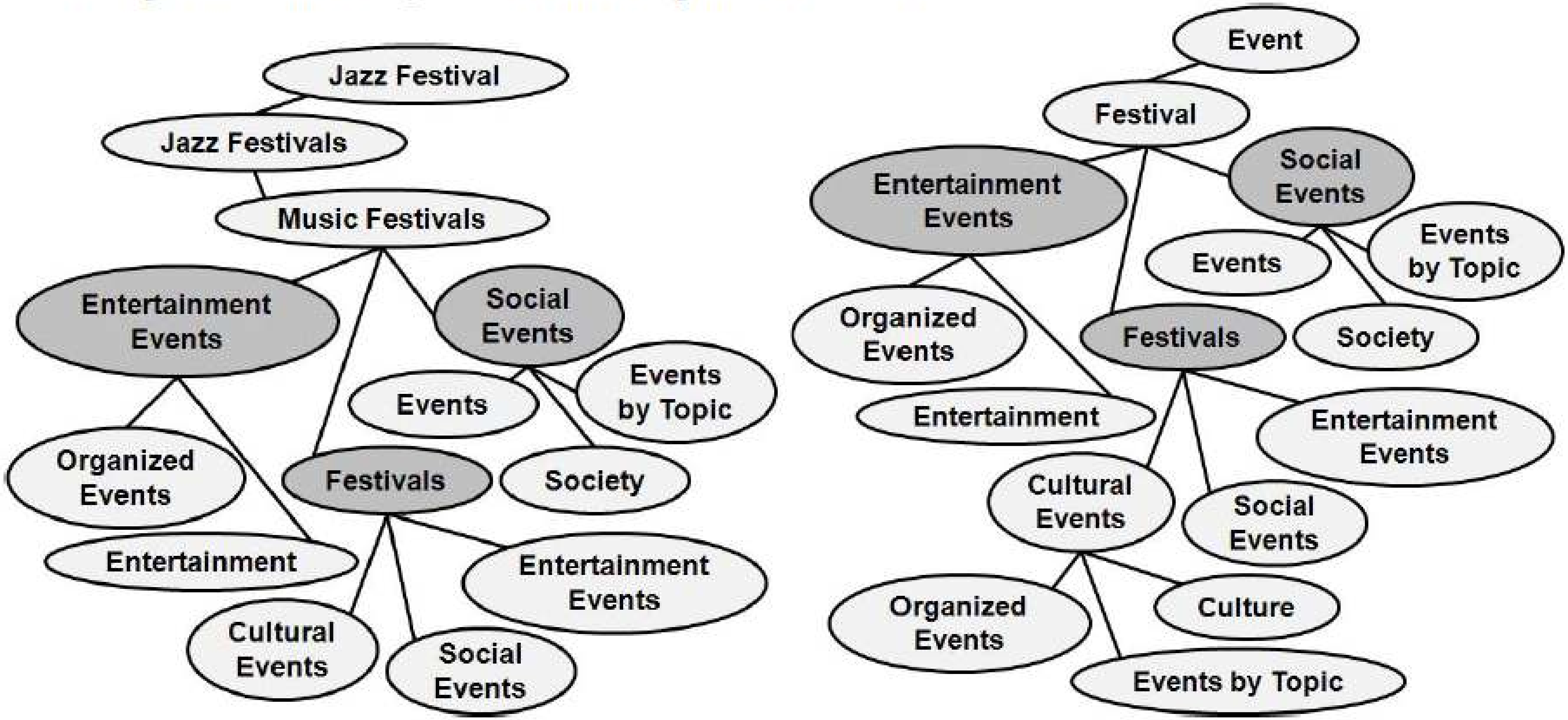
- **In the tree $T_{s,}$ remove all nodes for which the parent node which occurs in $T_{t\ to}$ create T_s'.**
  - All leaves of Ts are of level 4 or occur in Tt.
  - The pruned nodes do not contribute any additional new knowledge.

- **Compute overlap O_s between the source and target tree.**
  - Os= n/(k-1)
  - n = |z|, z ε Ts' Π Tt
  - k= |s|, s ε Ts'

- **The decision of alignment is made as follows.**
  - For Ts ε Tc and Tt ε Td, we have Ts=Tt, then C=D.
  - If min{o(Ts,Tt),o(Tt,Ts)} ≥ x, then set C rdfs:subClassOf D if o(Ts,Tt) ≤ o(Tt, Ts), and set D rdfs:subClassOf C if o(Ts, Tt) ≥ o(Tt, Ts).

Example

Fig. 1. BLOOMS trees for Jazz Festival with sense Jazz Festival and for Event with sense Event. To save space, some categories are not expanded to level 4.



Outline

- Introduction
- Motivation
- Existing Approaches
- BLOOMS Approach
- Evaluation
- Applications
- Conclusion & Future Work
- References

Evaluation Objectives

- **Examine BLOOMS as a tool for the purpose of LOD ontology matching.**
- **Examine the ability of BLOOMS to serve as a general purpose ontology matching system.**

BLOOMS

Kno.e.sis

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

BLOOMS

Table 1. Results on the oriented matching track. Results for RiMOM and AROMA have been taken from the OAEI 2009 website. Legends: Prec=Precision, A-API=Alignment API, OMV=OMViaUO, NaN=division by zero, likely due to empty alignment.

Ontology Alignment Initiative—Oriented Matching Track												
	A-API		OMV		S-Match		AROMA		RiMoM		BLOOMS	
Test	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
1XX	0	0	0.02	0.06	0.01	0.71	NaN	0	1	1	1	1
2XX	0	0	0.01	0.03	0.05	0.30	0.84	0.08	0.67	0.85	0.52	0.51
3XX	0.01	0.03	0.02	0.047	0.01	0.14	0.72	0.11	0.59	0.81	1	0.84
Avg.	0.00	0.01	0.02	0.04	0.03	0.38	0.63	0.07	0.75	0.88	0.84	0.78

Outline

- Introduction
- Motivation
- Existing Approaches
- BLOOMS Approach
- Evaluation
- Applications
- Conclusion & Future Work
- References

Potential Applications

- **Schema level reasoning over LOD.**
- **Identification and rectification of contradictory/misleading assertions**
  - Population of London is X (Geonames) / Population of London is Y (DBpedia), but geonames London is same as Dbpedia London.
  - Hollywood is a country. (Really?)
- **Enabling intelligent federated querying of LOD**
  - Beyond merely crawling.
  - Terminological difference can be resolved automatically.

Outline

- Introduction
- Motivation
- Existing Approaches
- BLOOMS Approach
- Evaluation
- Applications
- Conclusion & Future Work
- References

Conclusion

- **State of the art tools fail to scale up to the requirements of LOD ontologies.**
- **There is plenty of knowledge presented in community generated data which can be harnessed for improving itself.**

Future Work

- **New ways for computing overlap**
  - Penalize nodes which match at lower levels
  - Give priority to leftmost categories over rightmost categories.
- **Context based matching**
  - Harness implicit and explicit contextual information in matching.
  - Provide user with matches and the context of matching.
- **Use “committee” of auxiliary data sources for matching.**
- **BLOOMS based smart federated querying framework of LOD.**

References

- Prateek Jain, Pascal Hitzler, Amit P. Sheth, Kunal Verma, Peter Z. Yeh: Ontology Alignment for Linked Open Data. Proceedings of the 9th International Semantic Web Conference 2010, Shanghai, China, November 7th-11th, 2010. Pages 402-417

- Prateek Jain, Pascal Hitzler, Peter Z. Yeh, Kunal Verma, and Amit P.Sheth, Linked Data Is Merely More Data. In: Dan Brickley, Vinay K. Chaudhri, Harry Halpin, and 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

- Prateek Jain, Pascal Hitzler, Amit P. ShethFlexible Bootstrapping-Based Ontology AlignmentIn: P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt, M. Mao, I. Cruz (eds.), Ontology Matching, OM-2010. Proceedings of the 5th International Workshop on Ontology Matching, at ISWC2010, Shanghai, China, November 2010, pp. 136-137.

Thank You!

Questions?