

ScholarBase: Towards a Cross-Domain Knowledgebase for Linked Scholarly Data

Mahmoud Elbattah

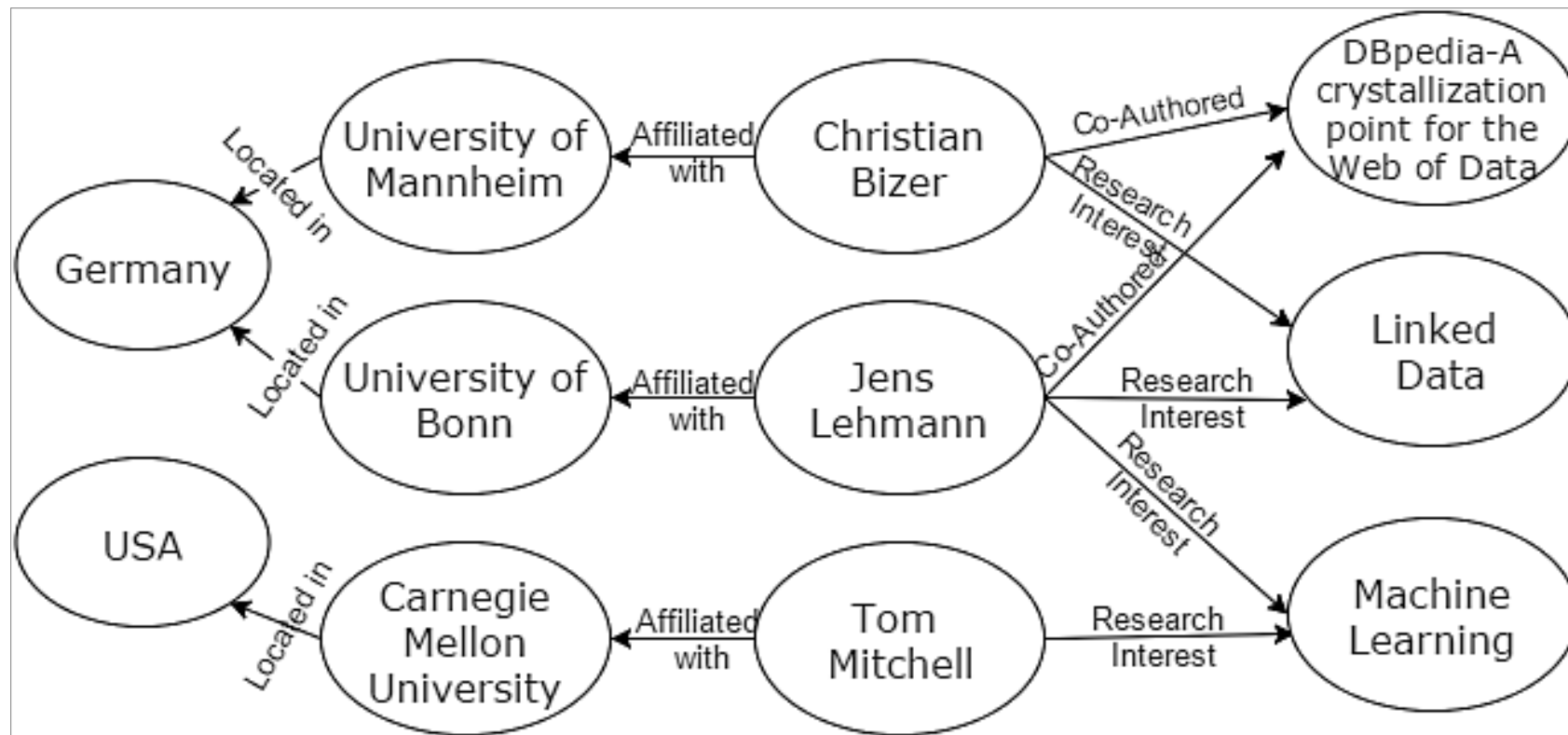


National University of Ireland, Galway

What is ScholarBase about?

- ScholarBase is aimed to serve as a Linked Data repository for cross-domain scholarly data.
- ScholarBase can be conceived as a knowledgebase that weaves links among:
 - scholars,
 - institutions,
 - research areas,
 - publications, and
 - geographical locations .

What is ScholarBase about?



Exemplary Queries

1. Who are the scholars that co-authored publications in relation to both of ML and Bioinformatics?
2. Who are the top-cited scholars in the field of ML, and are affiliated with institutions located in UK?
3. Who are the scholars contributed to ML, and are affiliated with institutions located in UK, and co-authored publications with scholars affiliated with institutions located outside the UK?
4. What are the institutions that are associated with the top-cited scholars in ML, and are located outside USA?
5. What are the inter-disciplinary research areas that bring together scholars from different backgrounds?

Data Source: Google Scholar Profiles

1 **Christian Bizer** Follow

Professor of Information Systems, University of Mannheim, Germany
 Linked Data, Web Science, Data Integration, Web Data Management
 Verified email at informatik.uni-mannheim.de - Homepage

2

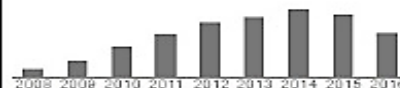
Title	1–20	Cited by	Year
Linked data-the story so far	C Bizer, T Heath, T Berners-Lee Semantic Services, Interoperability and Web Applications: Emerging Concepts ...	3991	2009
Dbpedia: A nucleus for a web of open data	S Auer, C Bizer, G Kobilarov, J Lehmann, R Cyganiak, Z Ives The semantic web, 722-735	2293	2007
Linked data: Evolving the web into a global data space	T Heath, C Bizer Synthesis lectures on the semantic web: theory and technology 1 (1), 1-136	1795	2011
DBpedia-A crystallization point for the Web of Data	C Bizer, J Lehmann, G Kobilarov, S Auer, C Becker, R Cyganiak, ... Web Semantics: science, services and agents on the world wide web 7 (3), 154-165	1594	2009
DBpedia spotlight: shedding light on the web of documents	PN Mendes, M Jakob, A Garcia-Silva, C Bizer Proceedings of the 7th international conference on semantic systems, 1-8	606	2011
Named graphs, provenance and trust:	JJ Carroll, C Bizer, P Hayes, P Stickler Proceedings of the 14th international conference on World Wide Web, 613-622	581	2005
DBpedia—a large-scale, multilingual knowledge base extracted from Wikipedia	J Lehmann, R Isele, M Jakob, A Jentzsch, D Konlokostas, PN Mendes, ... Semantic Web 6 (2), 167-195	528	2015
D2RQ-treating non-RDF databases as virtual RDF graphs	C Bizer, A Seaborne Proceedings of the 3rd international semantic web conference (ISWC2004) 2004	443	2004
The berlin sparql benchmark	C Bizer, A Schultze	437	2009
Linked data on the web (LDOW2008)	C Bizer, T Heath, K Idehen, T Berners-Lee	387	2008

3

Google Scholar

Impact Metrics

Citation indices	All	Since 2011
Citations	19607	16142
h-index	50	45
i10-index	103	91



4

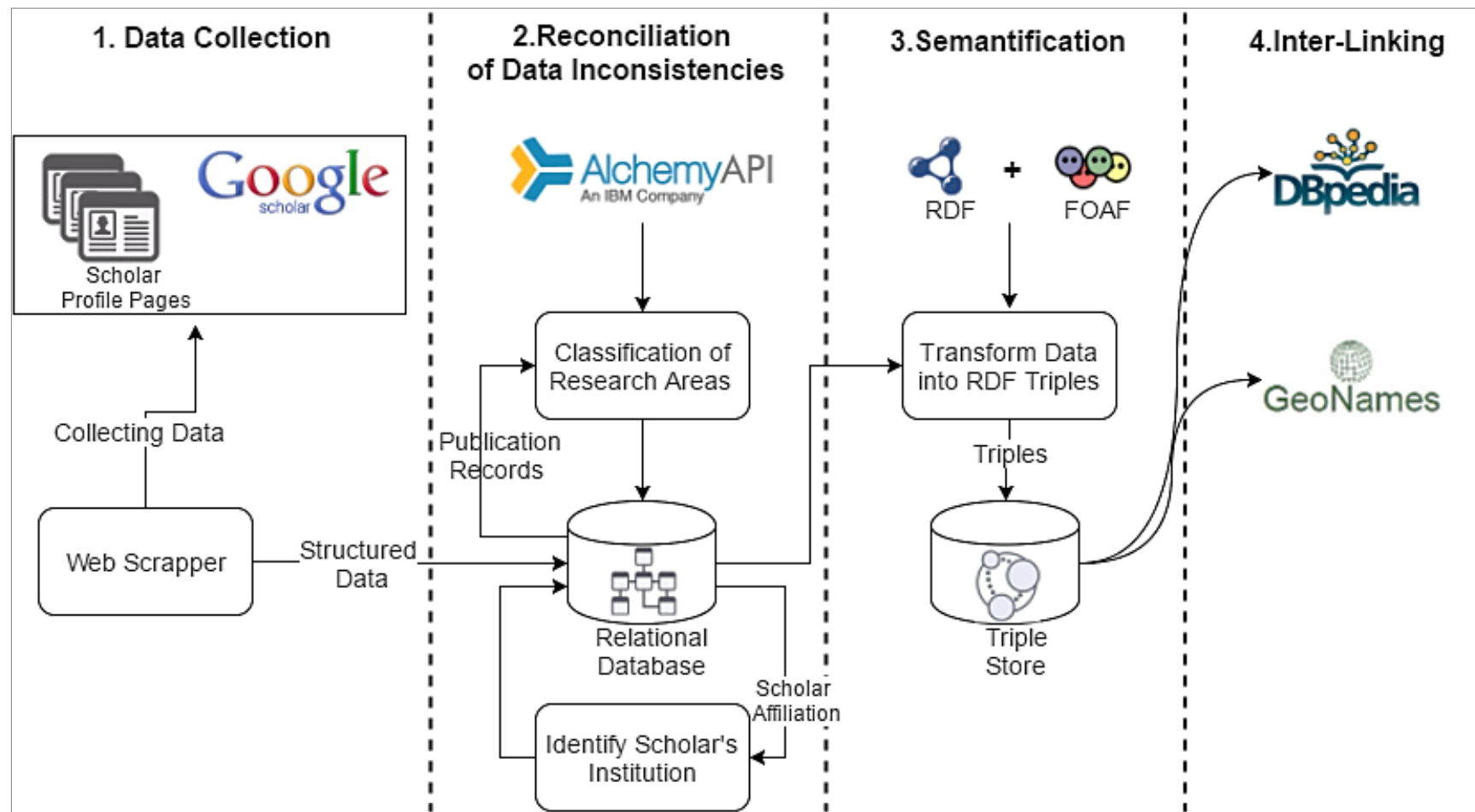
Co-authors View all...

- Richard Cyganiak
- Sören Auer
- Jens Lehmann
- Pablo N. Mendes
- Anja Jentzsch
- Heiko Paulheim
- Robert Meusel
- Oliver Lehmborg
- Hannes Mühleisen
- Olaf Hartig
- Dimitris Kontokostas
- Emmanuel Pietriga
- Dominique Ritze
- Volha Bryl
- Petar Ristoski
- David Karger
- Peter Boncz
- Oktie Hassanzadeh

Implementation Challenges

- Absence of Google Scholar APIs.
- Data inconsistencies and ambiguities.
- Missing data.

Overview



Stage 1: Data Collection

Data Collection Strategy

Stage 1-Random Walk:

Find initial seeds (i.e. scholar profiles) based on random search queries.

Stage 2-Collect Keywords:

Collect data describing research keywords and institutions from seed profiles gathered at Stage 1.

Stage 3-Focused-Search:

Find scholars based on focused-queries using keywords gathered at Stage 2.

Stage 4-Catch All:

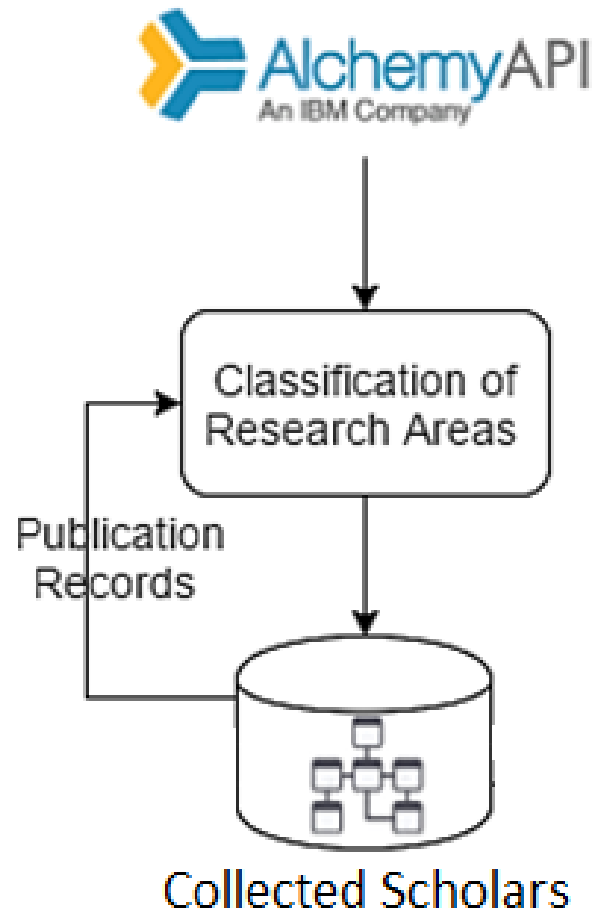
Collect scholars associated with keywords / institutions gathered at Stage 2 and Stage 3.

Stage 2: Reconciliation of Data Inconsistencies

Research Keywords Inconsistency

- Variation of keywords.
- Lack of specification.
- Excessive specification.
- Vagueness of acronyms.
- Variation of languages.
- Missing keywords.
- Misspelled keywords.

Reconciliation of Keywords Inconsistencies



Example of Keyword Reconciliation

Scholar Name	GS Keywords	Keywords Extracted by AlchemyAPI	
		Concepts	Taxonomy
Lotfi A. Zadeh	Fuzzy Logic, Soft Computing, Artificial Intelligence, Human-Level Machine Intelligence	Fuzzy Logic, Fuzzy Set	/technology and computing /science/computer science/artificial intelligence
Andrew P. Feinberg	Epigenetics, Epigenomics	Cancer, Epigenetics, DNA, DNA Methylation, Oncology	/health and fitness/disease/cancer

Affiliation Inconsistencies

Scholar	Affiliation
David Karger	MIT
N. P. Suh	M.I.T.
David Pesetsky	Massachusetts Institute of Technology

Affiliation Inconsistencies (cont'd)

Scholar	Affiliation	Verified email at
David Karger	MIT	mit.edu
N. P. Suh	M.I.T.	
David Pesetsky	Massachusetts Institute of Technology	

Stage 3: Semantification

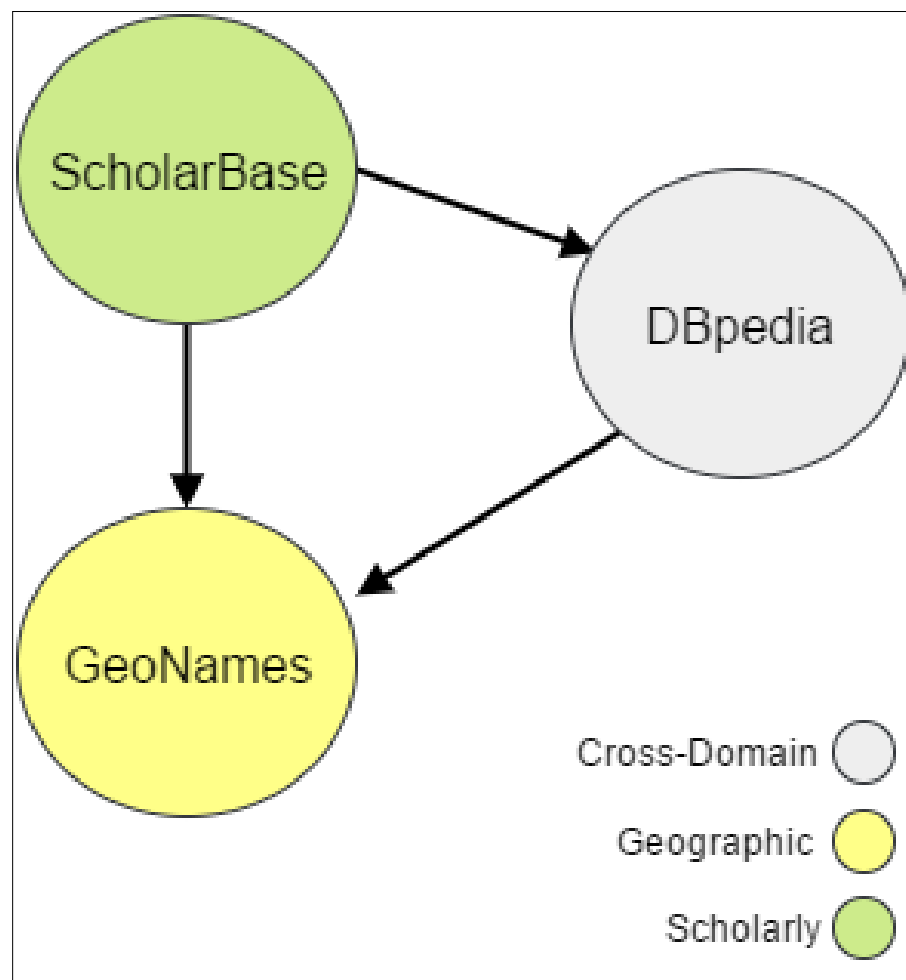
Semantification



Subject: <https://scholar.google.com/citations?user=S6H-0RAAAAAAJ>
Predicate: http://xmlns.com/foaf/spec/#term_topic_interest
Object: http://dbpedia.org/page/Fuzzy_logic

Stage 3: Linking to LOD

Linking to LOD



What is different about ScholarBase?

- ScholarBase might be the first initiative towards structuring the data of GS profiles.
- Unlike other endeavours that focused on specific domains in science (e.g semantic DBLP), or conferences (e.g. ESWC and ISWC), ScholarBase aims to be a knowledgebase of cross-domain scholarly data.
- Having consistent keywords for describing research keywords and affiliations can help to understand more about the dynamics of research areas, and answering complex queries about scholars .

Limitations

- The scholar entities within ScholarBase are tightly coupled to the presence of a GS profile.
- In other words, if a scholar does not have a GS profile, that scholar will not be included in the ScholarBase dataset.
- It is difficult to test the comprehensiveness of data collected by the web scraper, whereas we could not find any official reports from GS about the number of existing profiles.
- The web scraper cannot find GS profiles that are not set to not be publicly visible.

THANK YOU!

Mahmoud Elbattah

m.elbattah1@nuigalway.ie