Open Data Heterogeneity, Quality and Scale

Presentation of the Open Data Research Group

Zohra Bellahsene, Anne Laurent, François Scharffe, Konstantin Todorov

{firstname.lastname}@lirmm.fr

LIRMM / UM2

November 2013

Outline

1. The Big Picture
2. Ontology Matching
3. Data Linking
4. Dataset Recommendation
5. Warehousing and Datamining

Outline

1. The Big Picture
2. Ontology Matching
3. Data Linking
4. Dataset Recommendation
5. Warehousing and Datamining

The Big Picture

What is Open Data?

Any piece of data that is available free of cost to any individual or organization for use and re-use:

– *processing, mining, knowledge extraction, reasoning, statistical inference, etc.*

Testimonies of growing importance

- Release of the Open Data Charter by the G8
- EU digital agenda
- Open Data France
- Etalab (data.gouv.fr)

The Big Picture

The Semantic Web Context

The evolving vision of the World Wide Web:

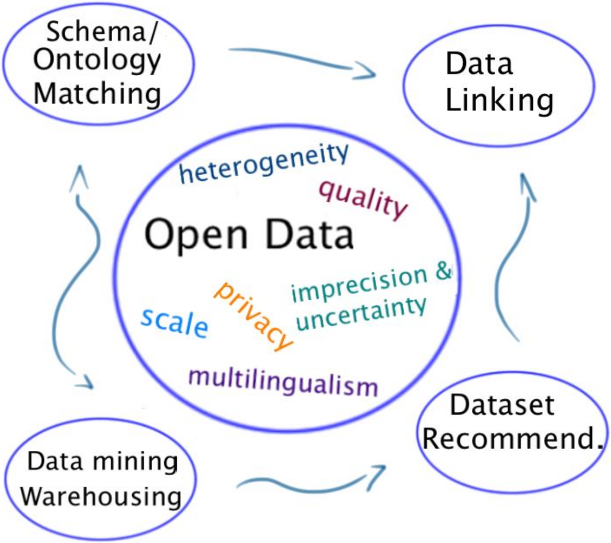
Web of Documents –> Semantic Web –> Web of Data

Linked data principles: [Bizer, Heath, Bernes-Lee. IJWS 2009]

1. Use URIs as names for things.
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).
4. Include links to other URIs, so that they can discover more things

The Big Picture

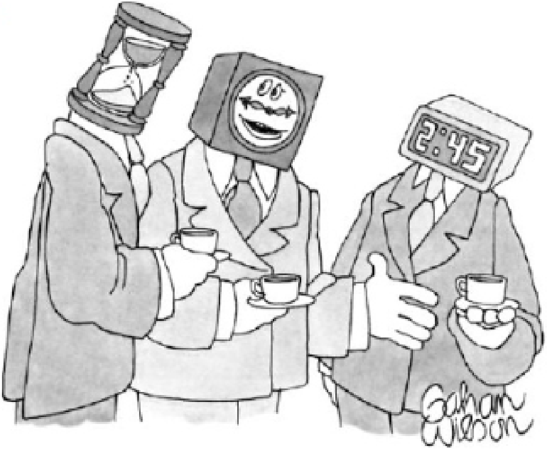
The Open Data Research Group



Outline

1. The Big Picture
2. Ontology Matching
3. Data Linking
4. Dataset Recommendation
5. Warehousing and Datamining

Ontology Matching



"Basically, we're all trying to say the same thing."

Borrowed from a tutorial by S. Staab and A. Hotho.

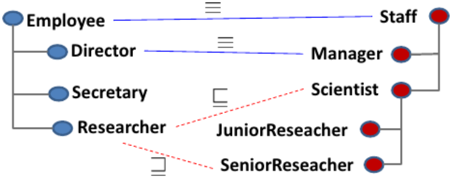
Ontology Matching

A Generic Framework for Ontology Matching and Evaluation

Ontologies are created in a **decentralized**, strongly **human biased** manner. Many ontologies describing the same domain of interest

=> **ontology heterogeneity:**

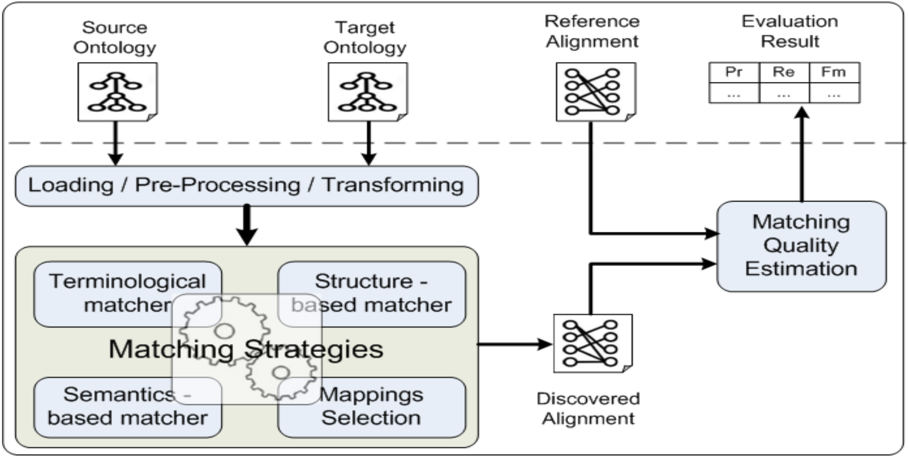
- syntactic
- terminological
- conceptual / structural



=> **Ontology Matching:** detect the semantic correspondences between the elements of two ontologies.

Ontology Matching

A Generic Framework for Ontology Matching and Evaluation



[Ngo, Bellahsene, Todorov. ESWC 2013]

Ontology Matching

Previous and Ongoing Work

Previous and ongoing work

- Schema matching / The system YAM
- The system YAM++
- Multimedia ontology matching for semantic information retrieval
- Fuzzy ontology matching with background knowledge
- Matching cross-lingual ontologies
- Large-scale matching (large ontologies vs. large number of ontologies)

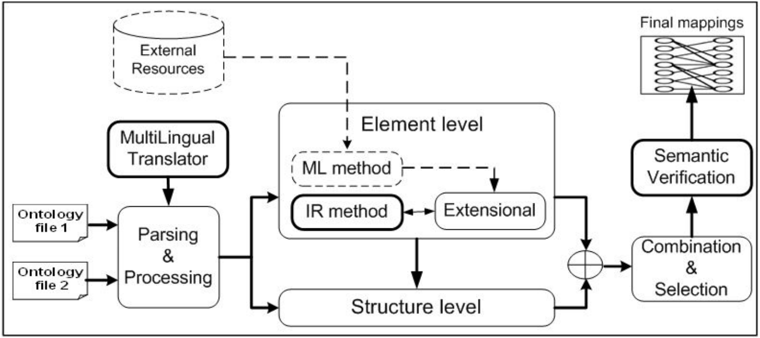
Ontology Matching

YAM++ (not) Yet Another Matcher

Many matching systems are out there. Here are some of the pluses of YAM++:

- Automatic configuration: similarity measures selection, tuning, and combination
- A novel terminological measure based on Tversky's similarity
- Able to deal with large ontologies

Among the best performing systems in the current state-of-the-art (cf. OAEI reports)



[Ngo, Bellahsene, EKAW 2012], [http://oaei.ontologymatching.org]

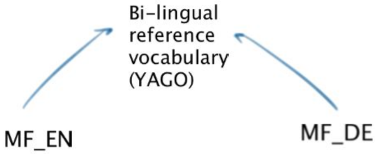
Ontology Matching

Cross-lingual Ontology Matching

Motivation

- No one-to-one correspondence between the majority of terms across different languages
- Machine translation still tolerates low precision levels
- No large training corpora with OM data

Use of background knowledge



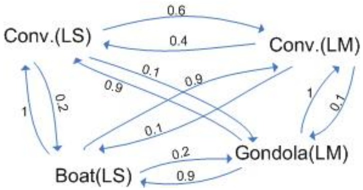
- Implicit alignment of cross-lingual ontologies (mediated by a YAGO/Wordnet taxonomy with French/German WordNet labels)
- No use of automatic translation
- Allows to capture various aspects of the similarity of concepts given in different languages

Ontology Matching

A Fuzzy Framework for Ontology Matching

Consider the (inherently) vague nature of concepts and their alignemnts

- Provide the missing implicit background knowledge
- Most matching procedures produce 1:1 mappings: often we will not be interested in the best (exact) match, but would like to find related yet not equivalent concepts
- A fuzzy set representation of the concepts, construction of a fuzzy common ontology
- Infer (fuzzy) relations between cross-ontology concepts



[Todorov, Hudelot, Popescu, Geibel. IJUFKS 2014 (in print)]

Outline

1. The Big Picture
2. Ontology Matching
3. Data Linking
4. Dataset Recommendation
5. Warehousing and Datamining

Data Linking

The fourth principle of linked data:

–> Include links to other URIs, so that they can discover more things

- Data Linking: The processes of finding equivalent resources on the web of data, or, more generally, connecting things that are somehow related
- Take two datasets as input -> produce a set of links between entities of these collections
- Establishing typed links (classes, properties, instances)
- Key-entity: a measure of similarity between instances distributed among heterogeneous data sources.

...but

- The Linked Data resources: over 31 billions of triples
- Yet only 5 percent of these are links between knowledge bases

Data Linking

Data linking, link discovery...

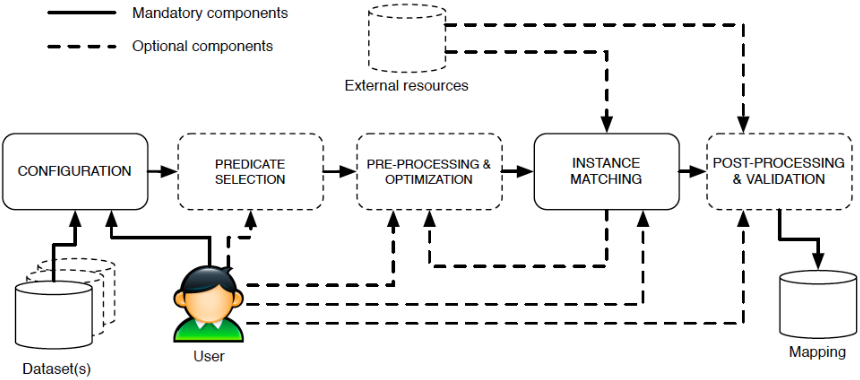
- An interdisciplinary topic on the frontiers of NLP, graph theory, statistics, RDB
- Related to the OM field, yet a separate problem
- Frameworks: RDF-AI, LIMES, SILK, zhishi.links, etc.
- Evaluation: Instance Matching track at the OAEI

Among the challenges:

- Complexity
- Link specification: choice and combination of similarity measures, which properties of the resources to take into account?
- The limitations of *owl:sameAs*: towards a more flexible definition of the relation of identity
- Multilingualism

Data Linking

A General Framework



[Ferrara, Nikolov, Scharffe. IJSW 2011]

Data Linking

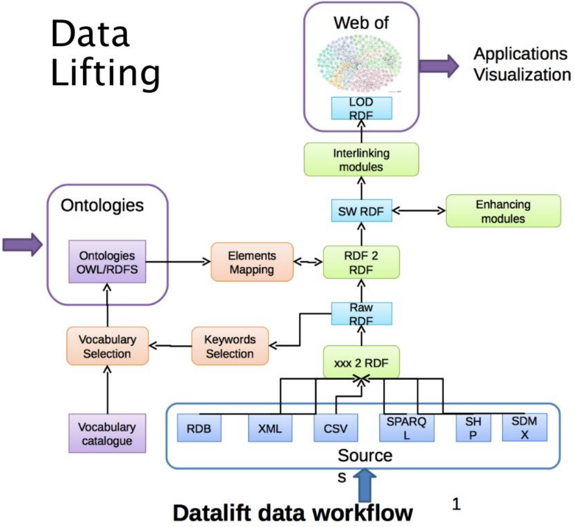
Datalift

The Datalift Project: a platform for "lifting" of data

- Different input formats are converted to RDF
- Ontologies are selected to describe the data
- Provide infrastructure for dataset publishing
- Handling licenses and access rights
- Data linking

Data Linking

Datalift



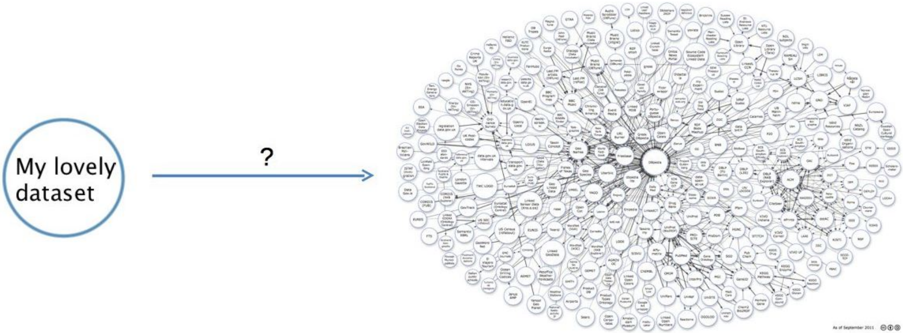
[Scharffe et al. 2012, http://datalift.org]

Outline

1. The Big Picture
2. Ontology Matching
3. Data Linking
4. Dataset Recommendation
5. Warehousing and Datamining

Dataset Recommendation for Linking

...Any candidates?



Towards an automatic ***discovery*** *and* ***recommendation*** *of candidate datasets for linking*

Dataset Recommendation for Linking

Given a dataset *d*, return a (possibly) ranked set of WoD datasets with respect to their relevance to the dataset *d* in view of the linking task.

Towards dataset profiling: definition of a collection of characteristics that allow to

- describe in the best possible way a dataset
- separate this dataset in the best possible way from other datasets
- many (statistical) characteristics of interest (scale, coverage, data values range, degree of connectedness, attribute entropy, etc...)

Two main approaches already proposed in the literature:

- Collect simple characteristics coming from vocabularies and meta-data [Luger. MSc Thesis 2012]
- A key-word search over an existing semantic web index [Nikolov et al. JIST 2011]

Outline

1. The Big Picture
2. Ontology Matching
3. Data Linking
4. Dataset Recommendation
5. Warehousing and Datamining

Warehousing and Datamining

Opening data requires to build data warehouses from sources that are selected and integrated.

Application of data mining techniques for selecting relevant views

Key questions:

- which cost functions are relevant in the particular context of open and linked data;
- how data mining algorithms can be used to automatically build relevant views;
- how such (multidimensional) views can be published to end-users; and
- how logical operators can be adapted to fit the web of data framework.

[A. Laurent, MJ Lesot. 2009], [Laurent et al. IJUFKS 2012]

References

[Bizer, Heath, Bernes-Lee. IJWS 2009] Christian Bizer, Tom Heath, Tim Berners-Lee: Linked Data - The Story So Far. Int. J. Semantic Web Inf. Syst. 5(3): 1-22 (2009)

[Ferrara, Nikolov, Scharffe. IJSW 2011] Alfio Ferrara, Andriy Nikolov, Franois Scharffe: Data Linking for the Semantic Web. Int. J. Semantic Web Inf. Syst. 7(3): 46-76 (2011) [Laurent, Lesot. 2009] A. Laurent, MJ Lesot. Scalable Fuzzy Algorithms for Data Management and Analysis:

Methods and Design. IGI Publishing. octobre 2009.

[Laurent et al. IJUFKS 2012] Joel Pinho Lucas, A. Laurent, M. N. Morenoa, M. Teisseire. Fuzzy Associative Classification Approach For Recommender Systems. In International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems (IJUFKS). vol. 20. n4. 2012.

[Luger. MSc Thesis 2012] Michael Luger. A Candidate Dataset Discovery and Linkage Recommendation System for Linked Data . MSc thesis, University of Innsbruck. (2012)

[Ngo, Bellahsene, EKAW 2012] DuyHoa Ngo, Zohra Bellahsene: YAM++ : A Multi-strategy Based Approach for Ontology Matching Task. EKAW 2012: 421-425

[Ngo, Bellahsene, Todorov. ESWC 2013] DuyHoa Ngo, Zohra Bellahsene, Konstantin Todorov: Opening the Black Box of Ontology Matching. ESWC 2013: 16-30

[Nikolov et al. JIST 2011] Andriy Nikolov, Mathieu d'Aquin, Enrico Motta: What Should I Link to? Identifying Relevant Sources and Classes for Data Linking. JIST 2011: 284-299

[Scharffe et al. AAAI 2012] Franois Scharffe, Ghislain Atemezing, Raphal Troncy, Fabien Gandon, Serena Villata, Bndicte Bucher, Fayal Hamdi et al. Enabling linked-data publication with the datalift platform. In Proc. AAAI workshop on semantic cities. 2012.

[Todorov, Hudelot, Popescu, Geibel. IJUFKS 2014 (in print)] Konstantin Todorov, Celine Hudelot, Adrian Popescu, Peter Geibel. Fuzzy Ontology Alignment Using Background Knowledge. Intl. Journal on Uncertainty, Fuzziness and Knowledge-Based Systems. To appear 2014.

Thank you for listening!