

Ontology matching tutorial

Jérôme Euzenat



Thanks to Pavel Shvaiko & Natasha Noy

Goals of the tutorial

- ▶ Provide an introduction to ontology matching;
- ▶ ... and eventually the semantic web;
- ▶ Start the discussion on links with formal concept analysis

Outline

- 1 Problem
- 2 Applications
- 3 Methods
- 4 Ontology matching and FCA
- 5 Conclusions

The semantic web?

The semantic web is an effort for publishing formal knowledge on the web.

It has developed various languages:

RDF Expressing data as graphs;

OWL, RDFS Expressing the ontologies governing such graphs;

SPARQL Query language for such graph

GRDDL, RDFa Embedding knowledge on the web

There are many tools for dealing with such languages and many resources expressed through it.

The semantic web is a success!

Such technologies are used every day (by yourself).

- ▶ Tens of billions of RDF triples and thousands of ontologies on the web;
- ▶ Governments and their agencies publish their data in RDF;
- ▶ Facebook (OG), Google (GKG), Yandex, Yahoo, Microsoft (schema.org) produce and consume semantic markup.

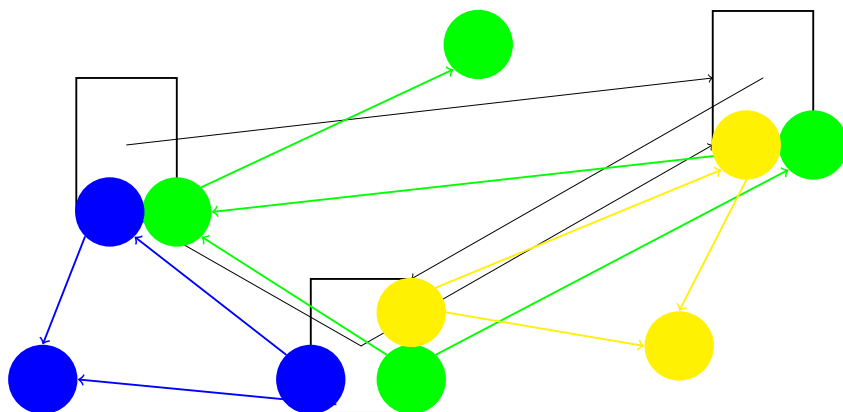
- ▶ And you do not even have to notice it.

What is an ontology?

An ontology typically provides a **vocabulary** that describes a domain of interest and a **specification of the meaning** of terms used in the vocabulary.

Depending on the precision of this specification, the notion of ontology encompasses several data and conceptual models, including, sets of terms, classifications, thesauri, database schemas, or fully axiomatized theories.

Semantic webs



Being serious about the semantic web

- ▶ It is not one guy's ontology.
- ▶ It is not several guys' common ontology.
- ▶ It is many guys and girls' many ontologies.
- ▶ So it is a mess, but a meaningful mess.

Living with heterogeneity

The semantic web will be:

- ▶ huge,
- ▶ dynamic,
- ▶ heterogeneous.

These are not bugs, these are features.

We must learn to live with them and master them.

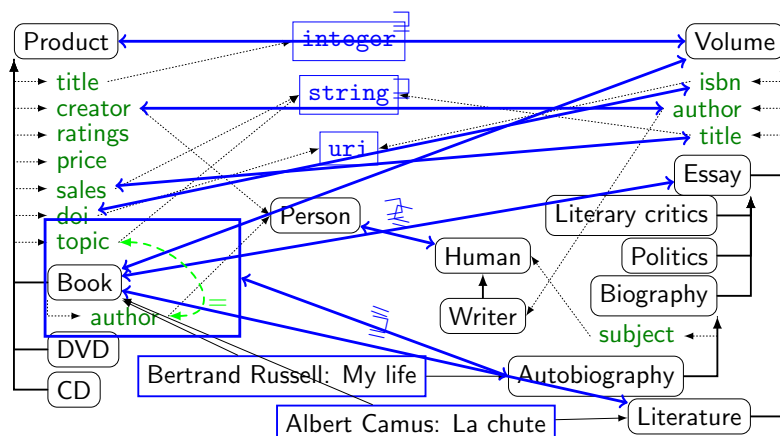
The heterogeneity problem

Resources being expressed in different ways must be reconciled before being used.

Mismatch between formalized knowledge can occur when:

- ▶ different languages are used (OWL vs. Topic maps);
- ▶ **different terminologies are used:**
 - ▶ English vs. Chinese;
 - ▶ Book vs. Volume.
- ▶ **different models are used:**
 - ▶ different classes: Autobiography vs. Paperback;
 - ▶ classes vs. property: Essay vs. literarygenre;
 - ▶ classes vs. instances: One physical book as an instance vs. one work as an instance.
- ▶ **different scopes and granularity are used.**
 - ▶ Only books vs. cultural items vs. any product;
 - ▶ Books detailed to the print and translation level vs. books as works.

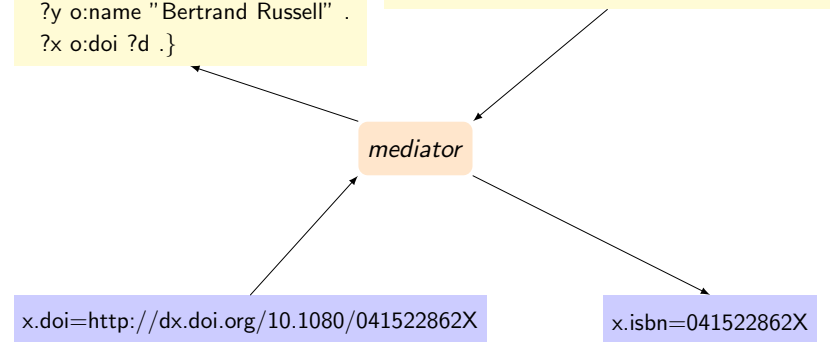
Ontology matching



Transformation and mediation

```
SELECT ?d
WHERE {?x rdf:type o:Book .
?x o:creator ?y .
?x o:topic ?y .
?y o:name "Bertrand Russell" .
?x o:doi ?d .}
```

```
SELECT ?i
WHERE { ?x rdf:type o':Autobiography .
?x o':author/o':name "Bertrand Russell" .
?x o':isbn ?i .}
```



Correspondences and alignments

Definition (Correspondence)

Given two ontologies o and o' , a **correspondence** between o and o' is a 3-uple: $\langle e, e', r \rangle$ such that:

- ▶ e and e' are **entities** of o and o' , for instance, classes, XML elements;
- ▶ r is a **relation**, for instance, **equivalence** ($=$), **more general** (\sqsupseteq), **disjointness** (\perp).

Definition (Alignment)

Given two ontologies o and o' , an **alignment (A)** between o and o' :

- ▶ is a set of correspondences between o and o'
- ▶ with some additional metadata (multiplicity: 1-1, 1-*, method, date, ...)

Terminology: a summary

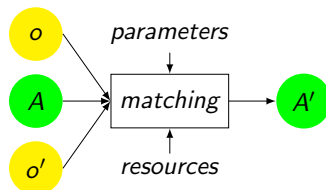
Matching is the process of finding relationships or correspondences between entities of different ontologies.

Alignment is a set of correspondences between two or more (in case of multiple matching) ontologies. The alignment is the output of the matching process.

Correspondence is the relation supposed to hold according to a particular matching algorithm or individual, between entities of different ontologies.

Mapping is the oriented version of an alignment.

The matching process

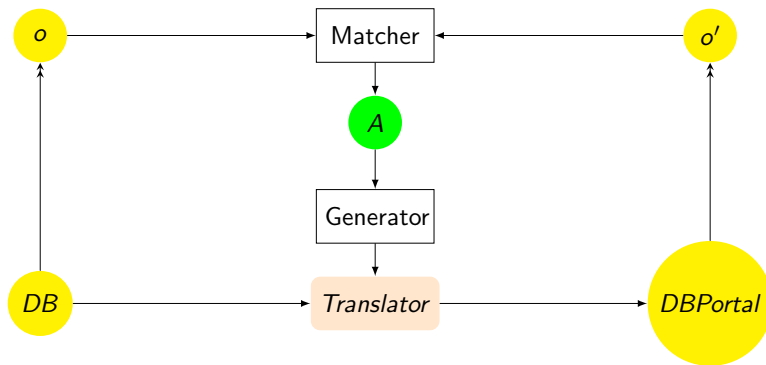


Why should we deal with this?

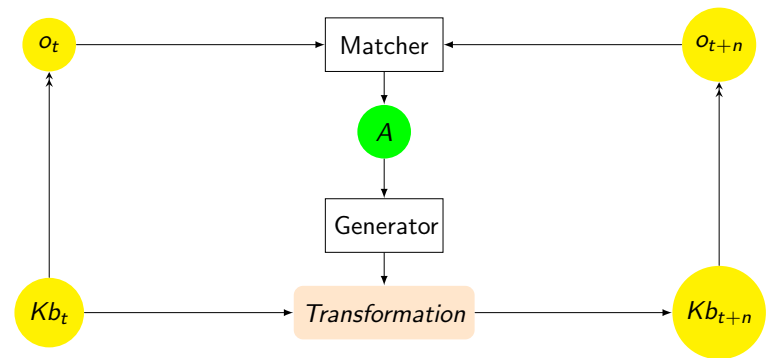
Applications of ontology matching:

- ▶ Catalogue integration
- ▶ Schema and data integration
- ▶ Query answering
- ▶ Peer-to-peer information sharing
- ▶ Web service composition
- ▶ Agent communication
- ▶ Data transformation
- ▶ Ontology evolution
- ▶ Data interlinking

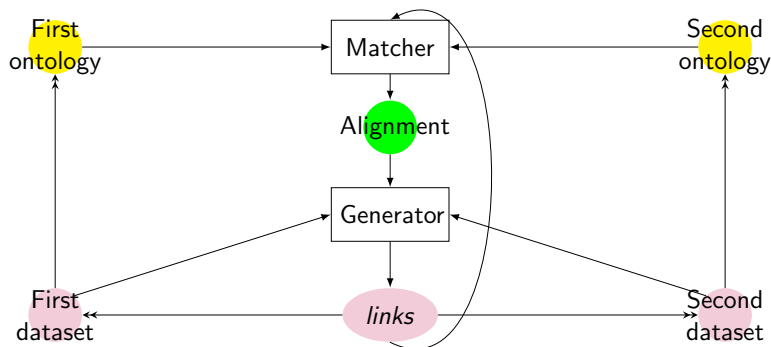
Applications: catalog integration



Applications: ontology evolution



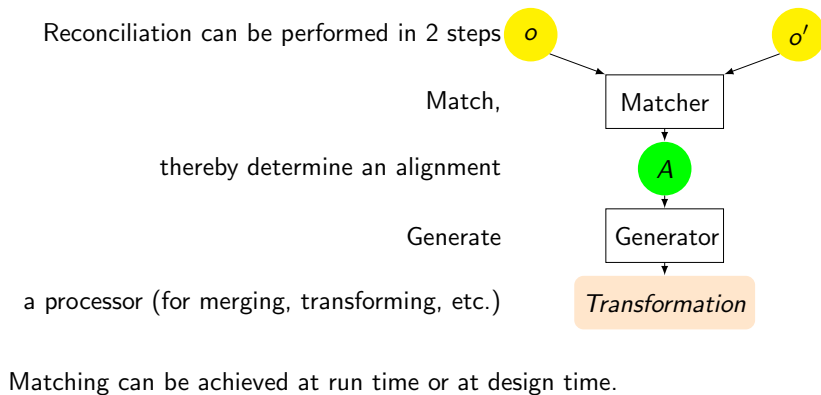
Application: Data interlinking



Applications requirements

Application	instances	run time	automatic	correct	complete	operation
Ontology evolution	✓			✓	✓	transformation
Schema integration	✓			✓	✓	merging
Catalog integration	✓			✓	✓	data translation
Data integration	✓			✓	✓	query answering
Linked data	✓			✓		data interlinking
P2P information sharing		✓				query answering
Web service composition		✓	✓	✓		data mediation
Multi agent communication		✓	✓	✓	✓	data translation
Query answering	✓	✓				query reformulation

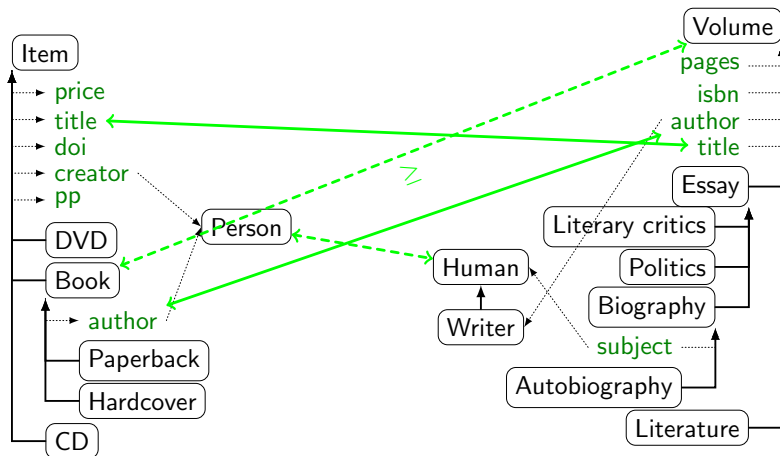
On reducing heterogeneity



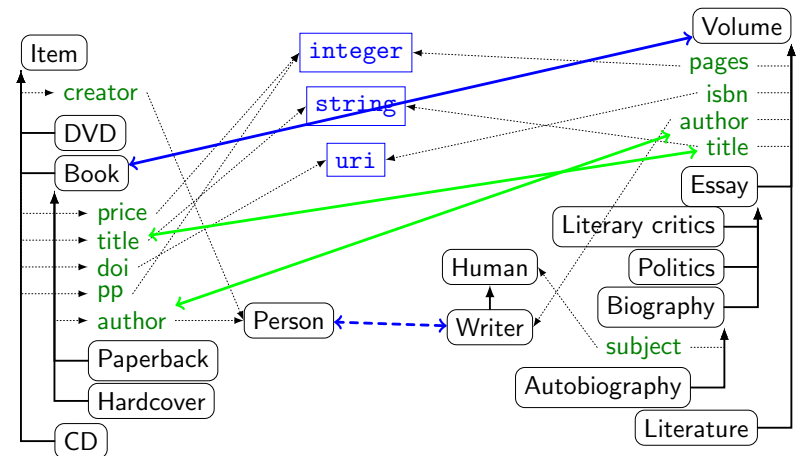
On what basis can we match?

- ▶ Content: relying on what is inside the ontology
 - ▶ **Name**, comments, alternate names, names of related entities: NLP, IR, etc.
 - ▶ **Internal structure**: constraints on relations, typing
 - ▶ **External structure**: relations between entities: data mining, discrete mathematics
 - ▶ **Extension**: statistics, data analysis, data mining, machine learning
 - ▶ **Semantics** (models): reasoning techniques
- ▶ Context: the relations of the ontology with the outside
 - ▶ **Annotated resources**:
 - ▶ The **web**
 - ▶ External **ontologies**: dbpedia, etc.
 - ▶ External **resources**: wordnet, etc.

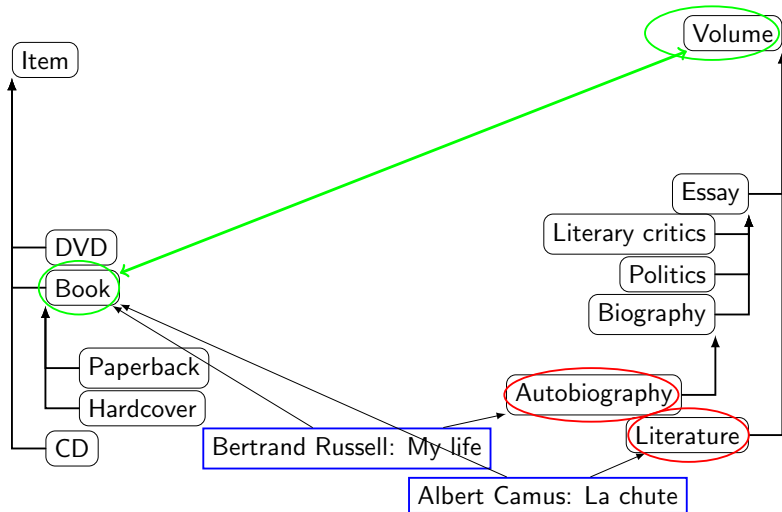
Name similarity



Structure similarity



Instance similarity



Basic methods: extensional

$$\epsilon : C \rightarrow E$$

E can be a set of instances, a set of documents which are indexed by concepts, a set of items, e.g., people, which use these concepts.

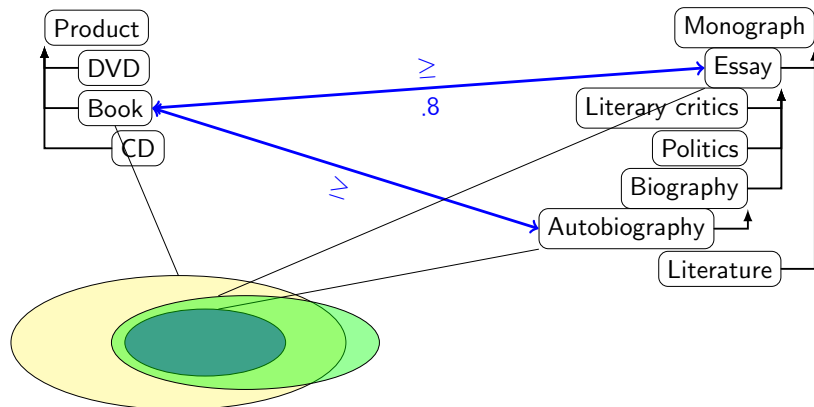
Two cases:

- ▶ E is common to both ontologies;
- ▶ E depends on the ontology. This can be reduced to the former case by identification or record linkage techniques.

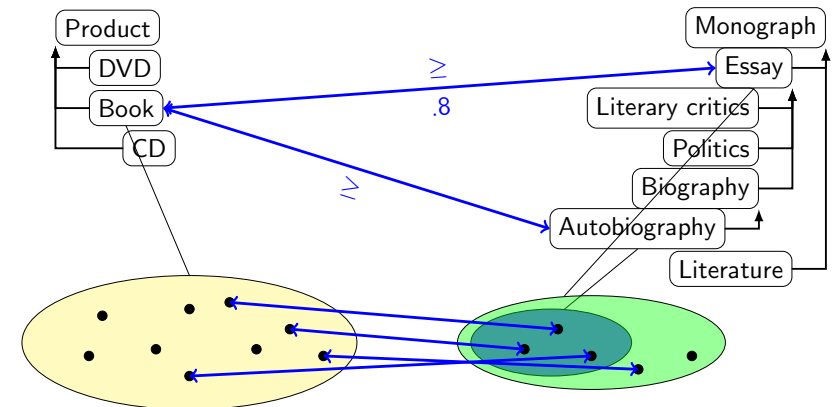
Techniques:

- ▶ statistical and machine learning techniques infer and compare the characteristics of populations;
- ▶ set-theoretic techniques compare the extensions;

Extensional techniques



Extensional techniques



Ontology matching and FCA

Ontology matching:

- ▶ From concepts, individuals and features of **two** sources
- ▶ Find equivalent concepts, features (and individuals)

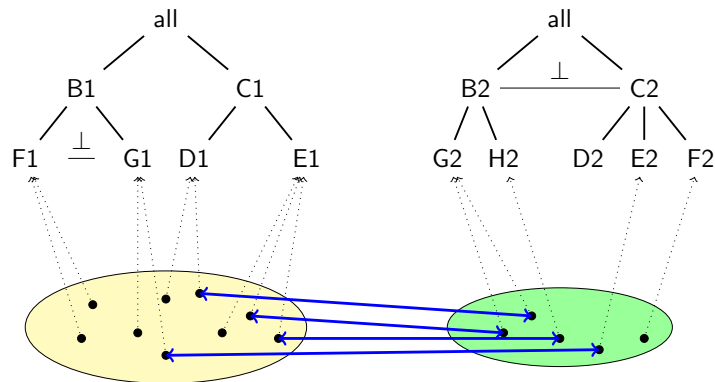
Formal concept analysis:

- ▶ Form individuals and features
- ▶ Find concepts

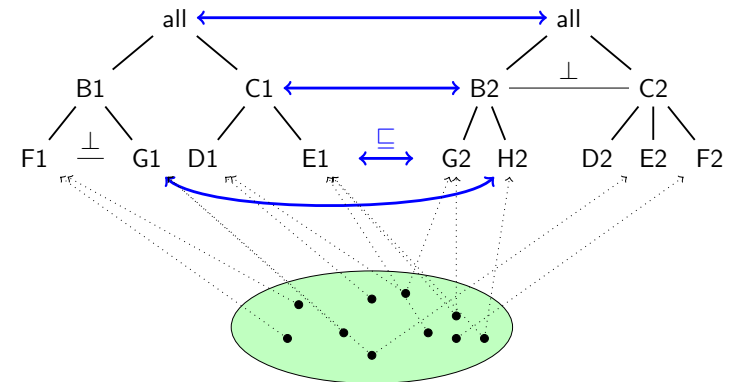
What is the same/what is different

- ▶ two sides (with no correspondences) instead of one
- ▶ the goal is not to create concepts

Rough idea



Rough idea



Encoding OM into FCA

- ▶ Really need to have common instances:
 - ▶ data interlinking (see tomorrow talk)
 - ▶ any other technique (which is equivalent)
- ▶ what can be the features:
 - ▶ classes (in both ontologies)
 - ▶ properties/relations

The problem is that the result will not be much different from cardinality analysis (concepts will be pairs of classes for which cardinality is 100%).

FCA-Merge [Stumme and Mädche, 2001]

1. instance extraction (→ create common instances);
2. compute lattice (FCA);
3. interactive merge of the ontologies (comparing classes covering concept extent and deciding to merge them).

Summary

- ▶ Heterogeneity of ontologies is in the nature of the semantic web;
- ▶ Ontology matching is part of the solution;
- ▶ It can be based on many different techniques;
- ▶ There are already numerous systems around;
- ▶ A relatively solid research field has emerged (tools, formats, evaluation, etc.) and it keeps making progress;
- ▶ But there remain serious challenges ahead.

Challenges

- ▶ Large-scale and efficient matching,
- ▶ Matching with background knowledge,
- ▶ Matcher selection, combination and tuning,
- ▶ User involvement,
- ▶ Social and collaborative matching,
- ▶ Uncertainty in matching,
- ▶ Reasoning with alignments,
- ▶ Alignment management.

and, of course, many others...

Acknowledgments

We thank all the participants of the Heterogeneity workpackage of the **Knowledge Web** network of excellence



In particular, we are grateful to Than-Le Bach, Jesus Barrasa, Paolo Bouquet, Jan De Bo, Jos De Bruijn, Rose Dieng-Kuntz, Enrico Franconi, Raúl García Castro, Manfred Hauswirth, Pascal Hitzler, Mustafa Jarrar, Markus Krötzsch, Ruben Lara, Malgorzata Mochol, Amedeo Napoli, Luciano Serafini, François Sharffe, Giorgos Stamou, Heiner Stuckenschmidt, York Sure, Vojtěch Svátek, Valentina Tamma, Sergio Tessaris, Paolo Traverso, Raphaël Troncy, Sven van Acker, Frank van Harmelen, and Ilya Zaihrayeu.

And more specifically to Marc Ehrig, Fausto Giunchiglia, Loredana Laera, Diana Maynard, Deborah McGuinness, Petko Valchev, Mikalai Yatskevich, and Antoine Zimmermann for their support and insightful comments

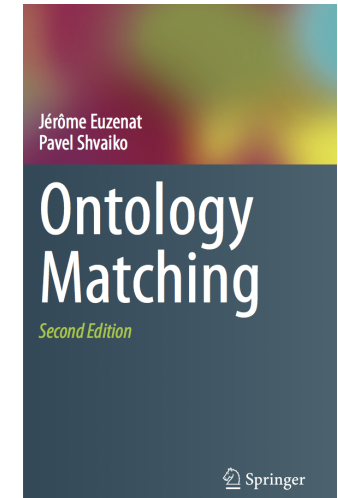
Part of this work was carried out while Pavel Shvaiko was with the University of Trento.

Ontology matching the book, 2nd edition

Jérôme Euzenat, Pavel Shvaiko

Ontology matching

1. Applications
2. The matching problem
3. Methodology
4. Classification
5. Basic similarity measures
6. Global matching methods
7. Strategies
8. Systems
9. Evaluation
10. Representation
11. User involvement
12. Processing



<http://book.ontologymatching.org>

Thank you

for your attention and interest!

Jerome.Euzenat@inria.fr

Pavel.Shvaiko@infotn.it

<http://www.ontologymatching.org>