Introduction to Ontology Matching

at

Méthodes et outils pour l’open data

Konstantin Todorov
todorov@lirmm.fr

LIRMM / UM2
Open Data Research Group

December 2014
Outline

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
Outline

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
Ontologies and the Semantic Web

The Semantic Web

The web of documents
Ontologies and the Semantic Web

The Semantic Web

Linking Data

RDF
Ontologies and the Semantic Web

The Semantic Web

More semantics: the ontologies
Ontologies and the Semantic Web

Vocabularies, ontologies

Best practices on the Web of Data:

- Use terms from widely developed vocabularies to name things
  - Vocabularies describing common things (people, places, projects) have emerged on the WoD.

- Align heterogeneous vocabularies
  - State that terms in different vocabularies are equivalent, or related: ontology matching
  - Make data as self descriptive as possible
Best practices on the Web of Data:

- Use terms from widely developed vocabularies to name things
  - Vocabularies describing common things (people, places, projects) have emerged on the WoD.

- **Align heterogeneous vocabularies**
  - State that terms in different vocabularies are equivalent, or related: ontology matching
  - $\rightarrow$ Make data as self descriptive as possible
Definition (Ontological Elements)

- $C$ a finite set of concepts
- $is\_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $I$ a set of instances
- $g : C \rightarrow 2^I$ a function that assigns subsets of instances from $I$ to each concept in $C$
- $l_L : C \rightarrow 2^{\Sigma^*}_L$ a function that assigns to each concept a set of labels from a set of labels $\Sigma^*_L$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, is\_a, R, I, g, l)$ forms an ontology.
Ontologies and the Semantic Web

Ontology – a formal definition

Definition (Ontological Elements)

- $C$ a finite set of concepts
- $is\_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $\\mathcal{I}$ a set of instances
- $g : C \rightarrow 2^\\mathcal{I}$ a function that assigns subsets of instances from $\\mathcal{I}$ to each concept in $C$
- $l_L : C \rightarrow 2^{\Sigma^*_L}$ a function that assigns to each concept a set of labels from a set of labels $\Sigma^*_L$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, is\_a, R, \mathcal{I}, g, l)$ forms an ontology.
Ontologies and the Semantic Web
Ontology – a formal definition

Definition (Ontological Elements)

- $C$ a finite set of concepts
- $is_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $\mathcal{I}$ a set of instances
- $g : C \rightarrow 2^\mathcal{I}$ a function that assigns subsets of instances from $\mathcal{I}$ to each concept in $C$
- $l_L : C \rightarrow 2^{\Sigma^*_L}$ a function that assigns to each concept a set of labels from a set of labels $\Sigma^*_L$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, is_a, R, \mathcal{I}, g, l)$ forms an ontology.
Ontologies and the Semantic Web

Ontology – a formal definition

**Definition (Ontological Elements)**

- **C** a finite set of concepts
- **is_a ⊆ C × C** a partial order on concepts
- **R** a set of relations on **C**
- **I** a set of instances
- **g : C → 2^I** a function that assigns subsets of instances from **I** to each concept in **C**
- **l_L : C → 2^Σ_L** a function that assigns to each concept a set of labels from a set of labels **Σ_L** coming from some alphabet **Σ_L** specific for a language **L**

**Definition (Ontology)**

\[ O = (C, is_a, R, I, g, l) \] forms an ontology.
Ontologies and the Semantic Web

Ontology – a formal definition

Definition (Ontological Elements)

- $C$ a finite set of concepts
- $is\_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $I$ a set of instances
- $g: C \rightarrow 2^I$ a function that assigns subsets of instances from $I$ to each concept in $C$
- $l_L: C \rightarrow 2^{\Sigma_L}$ a function that assigns to each concept a set of labels from a set of labels $\Sigma_L^*$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, is\_a, R, I, g, l_L)$ forms an ontology.
Ontologies and the Semantic Web

Ontology – a formal definition

Definition (Ontological Elements)

- $C$ a finite set of concepts
- $is\_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $\mathcal{I}$ a set of instances
- $g : C \rightarrow 2^{\mathcal{I}}$ a function that assigns subsets of instances from $\mathcal{I}$ to each concept in $C$
- $l_L : C \rightarrow 2^{\Sigma_L^*}$ a function that assigns to each concept a set of labels from a set of labels $\Sigma_L^*$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, is\_a, R, \mathcal{I}, g, l)$ forms an ontology.
Ontologies and the Semantic Web

Ontology – a formal definition

Definition (Ontological Elements)

- $C$ a finite set of concepts
- $\text{is}_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $\mathcal{I}$ a set of instances
- $g : C \rightarrow 2^{\mathcal{I}}$ a function that assigns subsets of instances from $\mathcal{I}$ to each concept in $C$
- $l_L : C \rightarrow 2^{\Sigma_L^*}$ a function that assigns to each concept a set of labels from a set of labels $\Sigma_L^*$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, \text{is}_a, R, \mathcal{I}, g, l)$ forms an ontology.
Definition (Ontological Elements)

- $C$ a finite set of concepts
- $\text{is}_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $\mathcal{I}$ a set of instances
- $g : C \rightarrow 2^\mathcal{I}$ a function that assigns subsets of instances from $\mathcal{I}$ to each concept in $C$
- $l_L : C \rightarrow 2^{\Sigma^*_L}$ a function that assigns to each concept a set of labels from a set of labels $\Sigma^*_L$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, \text{is}_a, R, \mathcal{I}, g, l)$ forms an ontology.
Ontologies and the Semantic Web

Ontology – a formal definition

Definition (Ontological Elements)

- $C$ a finite set of concepts
- $is_a \subseteq C \times C$ a partial order on concepts
- $R$ a set of relations on $C$
- $\mathcal{I}$ a set of instances
- $g : C \rightarrow 2^\mathcal{I}$ a function that assigns subsets of instances from $\mathcal{I}$ to each concept in $C$
- $l_L : C \rightarrow 2^{\Sigma^*_L}$ a function that assigns to each concept a set of labels from a set of labels $\Sigma^*_L$ coming from some alphabet $\Sigma_L$ specific for a language $L$

Definition (Ontology)

$O = (C, is_a, R, \mathcal{I}, g, l)$ forms an ontology.
Ontologies and the Semantic Web

Ontology – an example

- A set of concepts: EMPLOYEE, DIRECTOR, SECRETARY, RESEARCHER
- A set of labels: "employee", "director", "secretary", "researcher"
- A subsumption relation (is_a) on the set of concepts

Note: often a set of labels is assigned to a single concept (e.g., a set of synonyms, translations).
Ontologies are created in a **decentralized**, strongly **human biased** manner.

Ontologies describing the same domain of interest => **ontology heterogeneity**.

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

Ontology Matching

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

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

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
Heterogeneity Types

- **Syntactic**
  
  *about the formal expression of ontologies*
  
  example: OWL vs. F-logic

- **Terminological**
  
  *about the choice of labels*
  
  example: "director" vs. "manager"

- **Structural / Conceptual**
  
  *about the relations between elements*
  
  example: "is_a(director, person)" vs. "is_a(director, employee)"

  - granularity
  - coverage
  - scope
Heterogeneity Types

- **Syntactic**
  - *about the formal expression of ontologies*
  - example: OWL vs. F-logic

- **Terminological**
  - *about the choice of labels*
  - example: "director" vs. "manager"

- **Structural / Conceptual**
  - *about the relations between elements*
  - example: "is_a(director, person)" vs. "is_a(director, employee)"
    - granularity
    - coverage
    - scope
Heterogeneity Types

- **Syntactic**
  *about the formal expression of ontologies*
  example: OWL vs. F-logic
- **Terminological**
  *about the choice of labels*
  example: "director" vs. "manager"
- **Structural / Conceptual**
  *about the relations between elements*
  example: "is_a(director, person)" vs. "is_a(director, employee)"
  - granularity
  - coverage
  - scope
Ontology Alignment

The process of ontology matching results in an alignment.

An alignment:
a set of correspondances between the elements of two heterogeneous ontologies,
derived by resolving the different heterogeneities that they manifest.

Similarity measures on element level or global level are applied for every
heterogeneity type (e.g., terminological measures, etc.).

A function $\sigma : o \times o \rightarrow \mathbb{R}$ with some properties:

\[
\forall x, y \in o, \quad \sigma(x, y) \geq 0
\]
\[
\forall x, y, z \in o, \quad \sigma(x, x) \geq \sigma(y, z)
\]
\[
\forall x, y \in o, \quad \sigma(x, y) = \sigma(y, x)
\]
Ontology Alignment

The process of ontology matching results in an alignment.

An alignment:
a set of correspondences between the elements of two heterogeneous ontologies, derived by resolving the different heterogeneities that they manifest.

Similarity measures on element level or global level are applied for every heterogeneity type (e.g., terminological measures, etc.).

A function $\sigma : o \times o \rightarrow \mathbb{R}$ with some properties:

$$\forall x, y \in o, \quad \sigma(x, y) \geq 0$$
$$\forall x, y, z \in o, \quad \sigma(x, x) \geq \sigma(y, z)$$
$$\forall x, y \in o, \quad \sigma(x, y) = \sigma(y, x)$$
Ontology Alignment

The process of ontology matching results in an alignment.

An alignment:
a set of correspondances between the elements of two heterogeneous ontologies,
derived by resolving the different heterogeneities that they manifest.

Similarity measures on element level or global level are applied for every heterogeneity type (e.g., terminological measures, etc.).

A function $\sigma : o \times o \rightarrow \mathbb{R}$ with some properties:

$$\forall x, y \in o, \quad \sigma(x, y) \geq 0$$

$$\forall x, y, z \in o, \quad \sigma(x, x) \geq \sigma(y, z)$$

$$\forall x, y \in o, \quad \sigma(x, y) = \sigma(y, x)$$
Ontology Matching
Matching and Evaluation Framework

Figure: Ontology Matching: System Architecture and Evaluation Scenario
Outline

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
Outline

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
**Terminological Heterogeneity**

A Typology

**Hypothesis:** A concept = meaning of its label(s). →

**Terminological Heterogeneity:** Any difference in *spelling* between two labels which are assumed to refer to the same concept [4].

![Diagram illustrating different types of terminological heterogeneity](image)
Terminological Heterogeneity

Similarity Measures

- **String-based**
  - Edit-based
    - Levenshtein
    - ISUB
  - Token-based
    - QGrams

- **Language-based**
  - Internal linguistic properties
  - External resources

- Hybrid methods
  - Ext. Jaccard
  - SoftTFIDF
  - MongeEklan

K. Todorov
Terminological Heterogeneity

Discussion I

- **Token-based**
  - can handle compound labels
  - are less sensitive to word-swaps ("ConferenceMember" vs. "MemberConference")
  - sometimes need external resources to assign weights to the composing tokens (large corpus)

- **Edit-based**
  - can handle one-token labels with tiny variations in spelling
  - often used inside of a token-based measure
Sources of external information: dictionaries, thesauri, lexical databases (WordNet).

- **Two common problems** (for both internal and external measures)
  
  - dealing with single words and not compound ones ("PhDThesis" is not found in WN, although "PhD" and "Thesis" are)
  
  - typos or non-conventional abbreviations prevent from finding the words in dictionaries
Terminological Heterogeneity

Discussion III

- Limitations

- require large corpus for weight computation
- MongeElkan and softTFIDF are asymmetric
Terminological Heterogeneity
Measures and Heterogeneity Types
Outline

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
     - Instance-based Methods
     - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
Structural Matchers

Internal methods

Compute similarity based on the internal structure of elements (e.g., classes)

- their properties
- range
- cardinalities, etc

Usually combined with terminological techniques

Taken from [1].
Structural Matchers

External (relational) methods

Consider the relations of concepts to other concepts. Rely on already discovered terminological similarities.

- **Standard methods**
  - exploring structural relations between entities: 
    \[\rightarrow \text{descendants, ancestors, siblings, etc.}\]

- **Similarity Propagation**

![Diagram](image)
Outline

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
The similarity of two cross-ontology concepts is assessed by the help of the instances of these concepts

-> Many possible measures.
Ontology Matching
Ontology matching and machine learning

Intersection of class instance sets

-> Same instances need to be found in both ontologies.
Ontology Matching
Ontology matching and machine learning

The cosine of the prototypes

\[ \text{sim}(A, B) = s\left(\frac{1}{|A|} \sum_{j=1}^{A} i_j^A, \frac{1}{|B|} \sum_{k=1}^{B} i_k^B\right), \]

with \( s(x, y) \) the cosine similarity of \( x \) and \( y \).

-> Flattening class structure
Ontology Matching

Ontology matching and machine learning

The Jaccard coefficient

\[ Jacc(A, B) = \frac{Pr(A \cap B)}{Pr(A \cup B)}. \]

Machine learning is used to estimate the joint probabilities.
Outline

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
Ontology Matching

Machine Learning Approach

| Labeled training examples | Machine learning algorithm | Unknown example | Prediction rules | Predicted classification |

| attribute_i | attribute_j |

<table>
<thead>
<tr>
<th><strong>Ontology matching</strong></th>
<th><strong>Supervised binary classification</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;source entity, target entity&gt;</td>
<td>Example</td>
</tr>
<tr>
<td>Similarity measures</td>
<td>Attribute names</td>
</tr>
<tr>
<td>Similarity values</td>
<td>Attribute values</td>
</tr>
<tr>
<td>Confidence value</td>
<td>Predicted class</td>
</tr>
</tbody>
</table>
Outline

1 Ontologies and the Semantic Web

2 Heterogeneities and Alignments

3 Techniques
   Terminological Methods
   Structural Methods
   Instance-based Methods
   Combination of Measures

4 An Ontology Matching System

5 Some Current Topics in OM
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, multilingual

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

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

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

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

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, multilingual

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

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

1. Ontologies and the Semantic Web
2. Heterogeneities and Alignments
3. Techniques
   - Terminological Methods
   - Structural Methods
   - Instance-based Methods
   - Combination of Measures
4. An Ontology Matching System
5. Some Current Topics in OM
Background knowledge (BK) – any piece of external information that improves or enables the alignment [7].

- Dictionaries, thesauri, previous alignments, ontologies, the web...
- Domain specific sources of knowledge
  - domain specific corpora (of schemas and mappings);
  - domain specific ontologies, e.g., in the field of anatomy, upper-level ontologies, or all the ontologies available on the semantic web;
- The web and specifically linked data, Wikipedia (DBPedia, YAGO) [7];
- The use of BK results in a transformation of the input ontologies
Current Topics in OM
Multilingualism

**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** [6]

- Implicit alignment of cross-lingual ontologies (mediated by a YAGO/Wordnet taxonomy with multilingual labels)
- No use of automatic translation
- Allows to capture various aspects of the similarity of concepts given in different languages
Fuzzy Matching with BK
Hierarchical Fuzzification of an Ontology
Fuzzy Matching with BK

Example of Fuzzy Membership functions

(a) pc.hardware

(b) alt.atheism

(c) mac.hardware

(d) religion.christian

(e) (pc.hardware, mac.hardware)

(a)–(d) single fuzzy concepts; (e) the fuzzy concept of the match of two concepts.
Current Topics in OM

...and also

- User Involvement: include the user in the matching process
- Large-scale matching (large ontologies or multiple ontologies)
- Many-to-many type alignment
- Matcher evaluation
- Imprecision and uncertainty in the matching process
Many OM techniques are used in the data linking process (instance matching [2]).
Thank you for listening.
J. Euzenat and P. Shvaiko.  
*Ontology matching.*  

Alfio Ferrara, Andriy Nikolov, Jan Noessner, and François Scharffe.  
Evaluation of instance matching tools: The experience of OAEI.  

T.R. Gruber et al.  
Toward principles for the design of ontologies used for knowledge sharing.  

DuyHoa Ngo, Zohra Bellahsene, and Konstantin Todorov.  
Extended tversky similarity for resolving terminological heterogeneities across ontologies.  

P. Shvaiko and J. Euzenat.  
Ontology matching: state of the art and future challenges.  

Konstantin Todorov, Celine Hudelot, and Peter Geibel.  
Fuzzy and cross-lingual ontology matching mediated by background knowledge.  

Konstantin Todorov, Céline Hudelot, Adrian Popescu, and Peter Geibel.  
Fuzzy ontology alignment using background knowledge.  