Semantische Datenintegration

Mapping und Matching von Ontologien und Schemata

Jakob Voß

Hochschule Hannover

2017-06-10

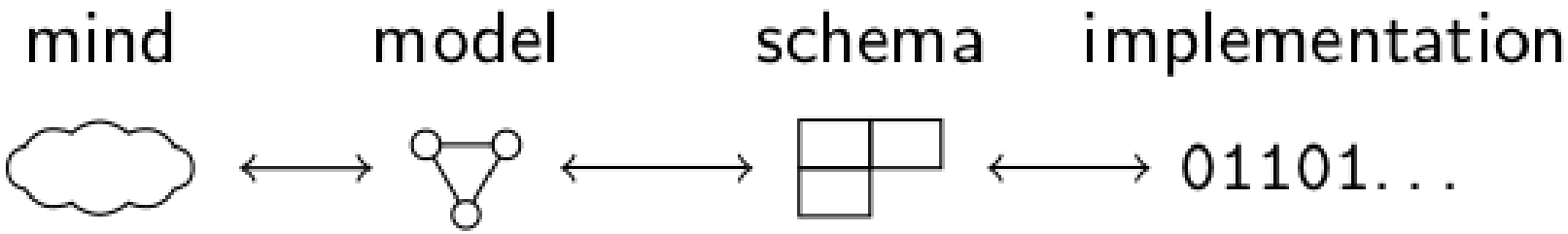
Was ist Mapping und Matching?

Mapping und Matching von Ontologien und Schemata

Schema	$\leftrightarrow$	Mapping
Ontology		Matching
Terminology		Crosswalk
Vocabulary		Concordance
...		...

- $\geq 4 \times 4$ mögliche Deskriptoren bei Prä-Kombination
- $\geq 4 + 4$ mögliche Deskriptoren bei Post-Kombination
- Reduktion durch (Quasi-)Synonyme Benennungen

Was kann gemappt werden?



Ebene	Form des Mappings
Mind	Begriffsklärung
Model	Ontology oder Terminology Mapping
Schema	Ontology oder Schema Mapping
Implementation	Schema Mapping oder Konvertierung

Was soll gemappt werden?

- Terminology Mapping
  - Knowlede Organization Systems (KOS): Normdateien, Thesauri, Klassifikationen...
- Ontology Mapping
  - Formallogische Systeme
- Schema Mapping
  - Datenformate

Sprachgebrauch uneinheitlich!

Beispiele

- Model Mapping
  - *Wir müssen reden. . .*
- Terminology Mapping
  - Wikidata Q15303972 = ORCID 0000-0002-7613-4123
  - RVK AN 65800 ≈ MeSH D007998
- Schema Mapping
  - Datenfelder
- Konvertierung
  - Zeichenkodierungen, Dateiformate. . .

Wofür brauchen wir Mappings?

- Integration verschiedener Datenquellen
  Export und Import in anderen Formaten
- Transformation und Migration eines Schemas
- Anfrageübersetzung

Terminology Mappings

Ausgangsfrage

- Anfrage in Vokabular A
- $\rightarrow$ Übersetzung in Vokabular B
- Wie gut ist die Übersetzung?

Beispiel

Vokabular A	Vokabular B
Aircraft	Aircraft – Airplane – Helicopter
Military Aircraft	Aircraft AND Military
Pest control – Pesticides	Pest control

Mögliche Äquivalenzen

- Ein Deskriptor
  - Gleiche Benennung (Ship = Ship)
  - Unterschiedliche Benennung (Ship = Vessel)
  - Weiter Begriffsumfang (Pesticides < Pest control)
  - Keine Entsprechung
- Mehrere Deskriptoren
  - OR-Kombination
    (Aircraft = Aircraft OR Airplane OR Helicopter)
  - AND-Kombination
    (Military Aircraft = Aircraft AND Military)
  - Komplexere Kombination
    (Animal food = Animals + (hunting OR husbandry))

Vorstellung: coli-conc

`http://coli-conc.gbv.de/`

- Sammlung von (Quellen für) KOS und Konkordanzen
- Software zur Verwaltung von KOS
- Bereitstellung von Konkordanzen
- Tool zur Erstellung und Bewertung von Konkordanzen

Übung: Europeana Fashion Vocabulary/Thesaurus

Kommen wir auf 100%?

- https://bartoc.org/en/node/1819
- https://www.wikidata.org/wiki/Property:P3832
- https://tools.wmflabs.org/mix-n-match/#/catalog/409
- https://www.wikidata.org/wiki/Wikidata:WikiProject_Fashion/Taxonomy/Europeana_Fashion_Vocabulary

Schema/Ontology Mapping

Beispiel: Ontology-Mapping mit Wikidata

- equivalent class P1709
  - Book Q571 → `http://schema.org/Book`
- exact match P2888
  - Comic Q1004 → `http://schema.org/ComicStory`
  - Erdreich Q36133 → `http://aims.fao.org/aos/agrovoc/c_7156`
- narrower external class P3950
  - Sammler Q3243461 → `http://comicmeta.org/cbo/Collector`

Unterschied zwischen P1709 und P2888 etwas unklar.
Warum kein *broader external class?*

Beispiel: Ontology-Mapping mit Wikidata

- equivalent property P1628
  - Teil von P361 → `http://schema.org/isPartOf`
- external subproperty P2235
  - Ausgabe oder Übersetzung von P629 → `http://comicmeta.org/cbo/translationOf`
- external superproperty P2236
  - Vater P22 → `http://schema.org/parent`

Außerdem spezifische Properties für ausgewählte Vokabulare!

Schema-Heterogenität

XML A

```
<article>
  <title>...</title>
  <url>...</url>
  <author>
    <name>...</name>
  </author>
</article>
```

XML B

```
<publication>
  <title>...</title>
  <creator>...</title>
</publication>
```

Demo: unAPI-Server der VZG

- `http://unapi.gbv.de/`
- Beispiel
  - `http://unapi.gbv.de/?id=gvk:ppn:786718889`
- `https://github.com/gbv/transformers`

Schema-Heterogenität

ARTICLE	PUBLICATION
– ID	– ID
– title	– title
– URL	– date
AUTHORSHIP	– author
– articleID	
– personID	
PERSON	
– ID	
– name	

Beispiel basiert auf Beispiel von Naumann und Leser (2006)

Schema/Ontology-Matching

Verfahren zur Erstellung von Schema-Mappings

- Umfangreiche Schemas
- Zahlreiche Schemas
- Unbekannte Schemas (fehlende Dokumentation)

Schema-Matching-Verfahren

Schema-Matching basiert auf

- Labels
- Instanzen
- Strukturen
- Mischformen

Label-basiertes Matching

- Gleiche Namen
- Ähnliche Namen
- Übersetzungen
- ...

`author`, `authors`, `Autor`, `Urheber`...

Instanz-basiertes Matching

- Idee
  - Gleiche oder ähnliche Werte(verteilungen)
- Annahmen
  - Beide Schemas müssen mit Werten gefüllt sein
  - Beide Datebasen müssen Duplikate enthalten
  - Duplikate müssen gleiche Attribute enthalten
- Beispiel
  - coli-conc Mapping-Algorithmus
  - Ein Datensatz hat Notationen mehrerer KOS
  - Kookkurenz $\Rightarrow$ Semantische Ähnlichkeit

Beispiel für Instanz-basiertes Matching

```
{
  "AAA": "Emma Goldman",
  "BBB": "2014",
  "CCC": "978-3-89401-810-8"
}
```

```
{
  "XXX": [ "Goldman, Emma" ],
  "YYY": "2014",
  "ZZZ": "9783894018108"
}
```

Beispiel für Instanz-basiertes Matching

```
{
  "author": "Emma Goldman",
  "year": "2014",
  "isbn": "978-3-89401-810-8"
}
```

```
{
  "author": [ "Goldman, Emma" ],
  "date": "2014",
  "isbn": "9783894018108"
}
```

Strukturbasiertes Matching

- Datentypen
- Nachbarschaftsbeziehungen
- Hierarchien
- Constraints
- ...

Sinnvoll vor allem in Mischformen

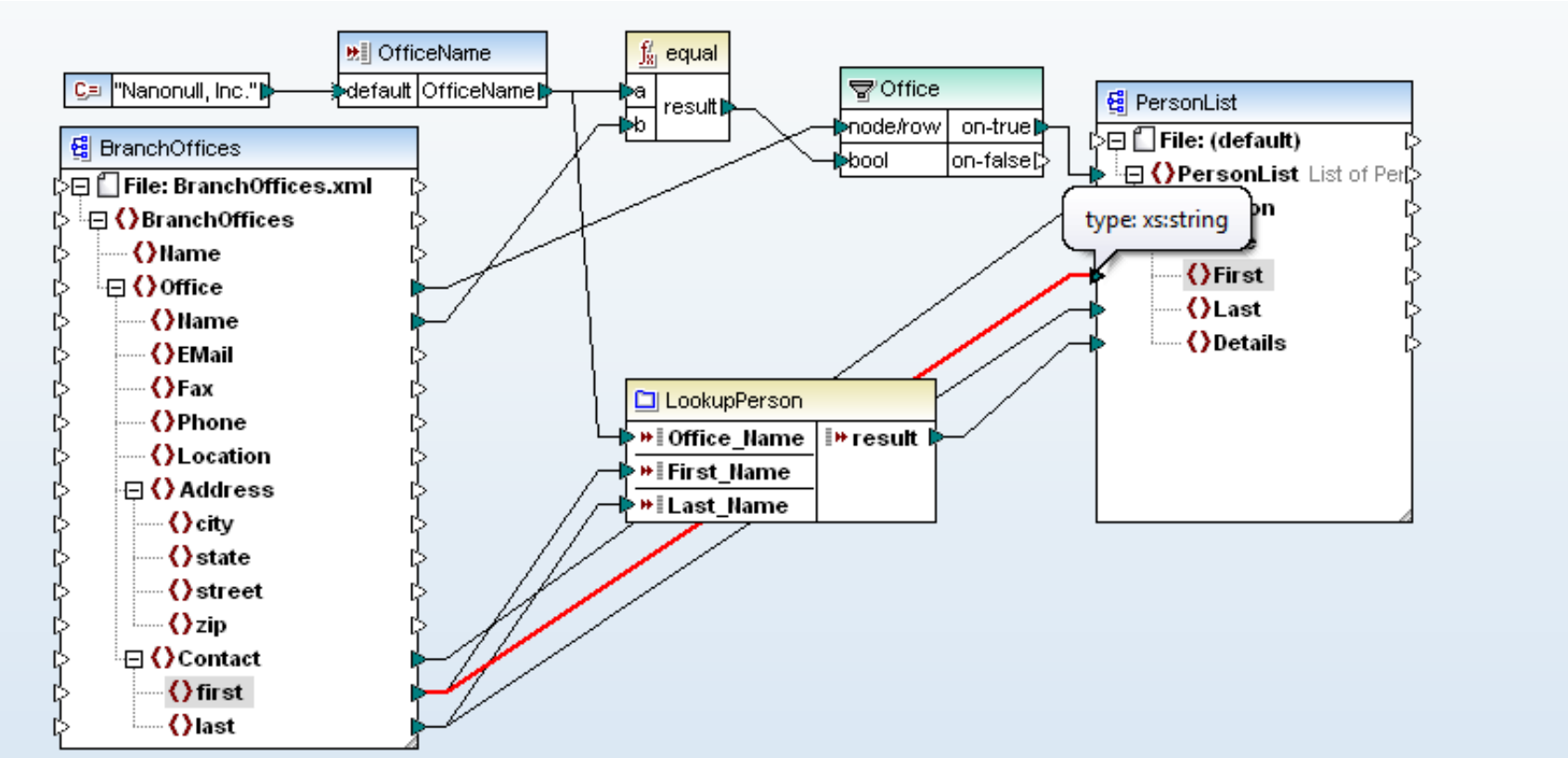
Zusammenfassung Matching-Verfahren

- Label/Instanz/Struktur + Mischformen
- State of the art nach (Otero-Cerdeira, Rodríguez-Martínez, und Gómez-Rodríguez 2015)
  - Review von 1600 bzw. 700 Fachartikeln (2003-2013)
  - Vor allem theoretische Ansätze
  - Weniger praktische Anwendungen
  - Bestehende Herausforderungen
- Kluft zwischen Automatischen und Manuellen Ansätzen (mein Eindruck)

Mapping-Tools

- Viele Forschunsgssysteme
  - `http://oaei.ontologymatching.org/` (OAEI)
  - `http://ontologymatching.org/`
- Bestandteil einiger ETL- und BD-Tools
- Einige erfolgreiche kommerzielle Systeme
  - Eher spezialisiert
  - Mehr Konvertierung und Mapping statt Matching

Beispiel: Altanova MapForce



https://www.altova.com/de/mapforce.html

Literatur und Quellen

Quellen dieser Folien: https://github.com/hshdb/MWM-317-02/
Folien zu *Terminology Mappings* basieren grob auf Unterlagen eines Tutorials von Dagobert Soergel auf der ECDL 2008 (S. 189-192)

J. Euzenat, P. Shvaiko. 2013. *Ontology matching*. 2. Aufl. Springer.
Naumann, Felix, und Ulf Leser. 2006. *Informationsintegration: Architekturen und Methoden zur Integration verteilter und heterogener Datenquellen*. Heidelberg: dpunkt-Verlag.
Otero-Cerdeira, Lorena, Francisco J. Rodríguez-Martínez, und Alma Gómez-Rodríguez. 2015. „Ontology matching: A literature review“. *Expert Systems with Applications* 42 (2): 949–71. doi:10.1016/j.eswa.2014.08.032.