Matching: KG-to-KG and CSV-to-KG

Ernesto Jiménez-Ruiz
Lecturer in Artificial Intelligence

Before we start...

Agenda

- Four sessions split into two days
- **Morning sessions**:
  - Theory: 9:00-10:30
  - Break 15 min
  - Hands-on: 10:45-12:15
- Lunch break (1 hour)
- **Afternoon sessions**:
  - Theory: 13:15-14:45
  - Break 15 min
  - Hands-on: 15:00-16:30

Course Organization

- ✓ Introduction to Knowledge Graphs
  - ✓ Lab: Creation of a small knowledge graph and ontology.
- ✓ Reasoning and Querying with Knowledge Graphs
  - ✓ Lab: First steps with the SPARQL query language.
- 3. Matching: KG-to-KG and CSV-to-KG
  - – Lab: Creation of a (simple) matching system.
- 4. Knowledge Graphs and Language Models
  - – Lab: Ontology Embeddings with OWL2Vec*.

FAIR principles and 5-star data

Why Ontologies and KGs?

5-star Data

- ⋆ **OL**: make your data available on the Web (in any format) under an open license.
- ⋆⋆ **RE**: make the data machine readable (excel instead of an scanned image).

Tim Berners-Lee. **5 ⋆ (open) data:** `https://5stardata.info/en/`

5-star Data

- ⋆ **OL**: make your data available on the Web (in any format) under an open license.
- ⋆⋆ **RE**: make the data machine readable (excel instead of an scanned image).
- ⋆⋆⋆ **OF**: use a non proprietary open format (*e.g.*, CSV).
- ⋆⋆⋆⋆ **URI**: use URIs instead of strings (RDF).

Tim Berners-Lee. **5 ⋆ (open) data:** `https://5stardata.info/en/`

5-star Data

- ⋆ **OL**: make your data available on the Web (in any format) under an open license.
- ⋆⋆ **RE**: make the data machine readable (excel instead of an scanned image).
- ⋆⋆⋆ **OF**: use a non proprietary open format (*e.g.*, CSV).
- ⋆⋆⋆⋆ **URI**: use URIs instead of strings (RDF).
- ⋆⋆⋆⋆⋆ **LOD**: link your data to other data to provide extended context.

Tim Berners-Lee. **5 ⋆ (open) data:** `https://5stardata.info/en/`

5-star Data

- ⋆ **OL**: make your data available on the Web (in any format) under an open license.
- ⋆⋆ **RE**: make the data machine readable (excel instead of an scanned image).
- ⋆⋆⋆ **OF**: use a non proprietary open format (*e.g.*, CSV).
- ⋆⋆⋆⋆ **URI**: use URIs instead of strings (RDF).
- ⋆⋆⋆⋆⋆ **LOD**: link your data to other data to provide extended context.

♠ This could be applied **within an organisation** (intranet), not only for the Web. Ideally with an OL, but at least data accessible by everyone in the organisation.

Tim Berners-Lee. **5 ⋆ (open) data:** `https://5stardata.info/en/`

5-star Data: Technical challenges:

- How to **expose** data (*e.g.*, databases, csv files) as knowledge graphs?
- How to **create** (or reuse) and use (abstract) **knowledge** ?
- How to **align** different knowledge graphs? ♠
- How to check **consistency and trust** of the data and knowledge? ♠

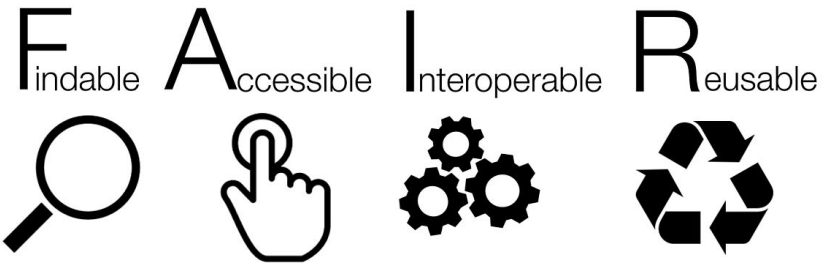
♠ Better with things than with strings

5-star Data: Technical challenges:

- How to **expose** data (*e.g.*, databases, csv files) as knowledge graphs?
  - *Matching CSV to KG - 4-5 $\star$ data*
- How to **create** (or reuse) and use (abstract) **knowledge** ?
  - *OWL Ontologies*
- How to **align** different knowledge graphs? ♠
  - *Matching KG to KG with Ontology Alignment - 5 $\star$ data*
- How to check **consistency and trust** of the data and knowledge? ♠
  - *Reasoning with OWL - 6 $\star$ data?*

♠ Better with things than with strings

FAIR Data Principles (i)



♠ For for machines and people.

The FAIR Guiding Principles for scientific data management and stewardship. Nature Scientific Data 2016.
https://www.go-fair.org/fair-principles/

FAIR Data Principles (ii)

- **Findable**: (meta)data is assigned a unique identifier and data is described with rich metadata.
- **Accessible**: (meta)data can be accessed via a known protocol. $(\star)$
- **Interoperable**: meta(data) uses a formal language.
- **Reusable**: well described (meta)data with rich provenance.

$(\star)$ Not necessarily open. (FAIR $\neq$ OPEN)

FAIR Data Principles (ii)

- **Findable**: (meta)data is assigned a unique identifier and data is described with rich metadata. (URIs)
- **Accessible**: (meta)data can be accessed via a known protocol. (SPARQL) ($\star$)
- **Interoperable**: meta(data) uses a formal language. (Knowledge representation with OWL)
- **Reusable**: well described (meta)data with rich provenance. (W3C standards)

($\star$) Not necessarily open. (FAIR $\neq$ OPEN)

FAIR Data Principles (iii)

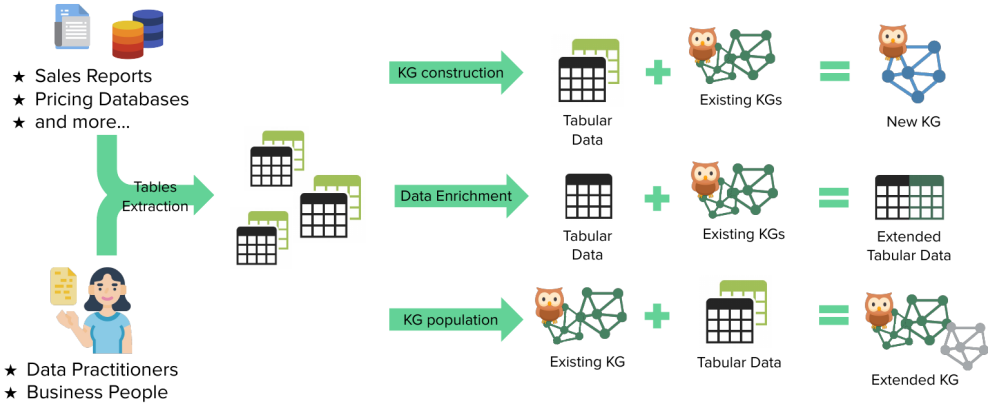
– FAIR data is more valuable, easier to find and combine thanks to unique identifiers and a formal shared knowledge representation.

– “KGs must be in want of FAIR data. And FAIR data is in want of KGs”.

Carole Golbe. FAIRy stories: the FAIR Data principles in theory and in practice.
`https://www.slideshare.net/carolegoble/fairy-stories-the-fair-data-principles-in-theory-and-in-practice`

PART I: From (Tabular) Data to Knowledge Graphs

Transformation workflows



Vincenzo Cutrona. **Why Table Understanding Matters**.

Exposing data as RDF: Ingredients

- **Ontology vocabulary**. Custom and/or given by a public KG.
- **Mappings**. Define a transformation from the tabular data to RDF data.
- **Ontology Axioms** (optional) - ♠

♠ Ernesto Jimenez-Ruiz and others. **BootOX: Practical Mapping of RDBs to OWL 2**. ISWC 2015

Exposing data as RDF: Direct Mapping Example

Automatic triples:

```
ex:row1 ex:col1 ‘‘China’’
ex:row1 ex:col2 ‘‘Beijing’’
ex:row2 ex:col1 ‘‘Indonesia’’
ex:row2 ex:col2 ‘‘Jakarta’’
...
```

China	Beijing
Indonesia	Jakarta
Congo	Kinshasa
Brazil	
Congo	Brazzaville

(*) `ex:` is the prefix defined for the namespace `http://example.org/`

Exposing data as RDF: Direct Mapping Example

Automatic triples:

```
ex:row1 ex:col1 ‘‘China’’
ex:row1 ex:col2 ‘‘Beijing’’
ex:row2 ex:col1 ‘‘Indonesia’’
ex:row2 ex:col2 ‘‘Jakarta’’
...
```

(we can probably do better)

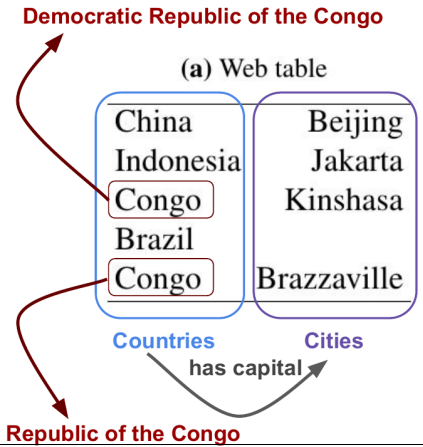
China	Beijing
Indonesia	Jakarta
Congo	Kinshasa
Brazil	
Congo	Brazzaville

(*) `ex:` is the prefix defined for the namespace `http://example.org/`

Exposing data as RDF: Enhanced Mapping/Transformation (i)

– **If** we know the **semantics** of the data.
– **Potential automatic triples:**

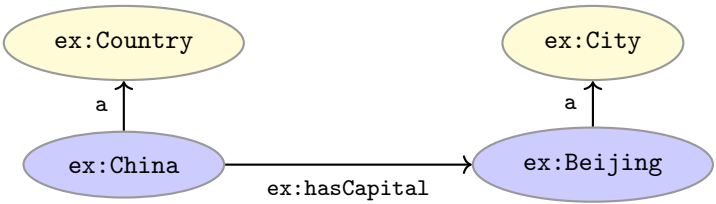
```
ex:China rdf:type ex:Country
ex:Beijing rdf:type ex:City
ex:China ex:hasCapital ex:Beijing
...
```



Exposing data as RDF: Enhanced Mapping/Transformation (ii)

Return capital of China (for the KG $\mathcal{G}$ below):

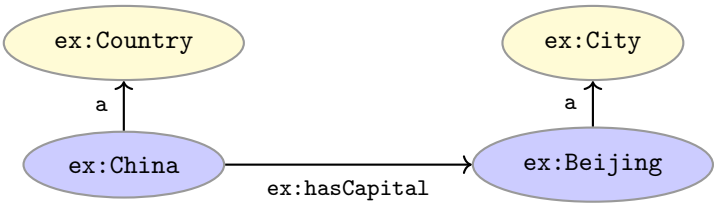
```
PREFIX ex:  <http://example.org/>
SELECT DISTINCT ?capital WHERE {
    ex:China ex:hasCapital ?capital .
}
```



Exposing data as RDF: Enhanced Mapping/Transformation (ii)

Return capital of China (for the KG $\mathcal{G}$ below): Query Result= {ex:Beijing}

```
PREFIX ex:  <http://example.org/>
SELECT DISTINCT ?capital WHERE {
    ex:China ex:hasCapital ?capital .
}
```



Semantic Understanding of Tabular Data

Adding Semantics to Tabular Data

- **Semi-automatic** process.
- Key for an **enhanced transformation** to RDF triples.
- But also for other tasks with independence of a final KG creation.
  - Tabular data in the form of CSV files is the common input format in a **data analytics pipeline**.
  - The **lack of semantics and context in datasets** hinders their usability.
  - Gaining **semantic understanding** will be very valuable for data integration, data cleaning, data mining, machine learning and knowledge discovery tasks.

Adding Semantics to Tabular Data: Basic Tasks

- Matching a cell to a KG entity (**CEA task** - Cell-Entity Annotation)
- Assigning a semantic type (*e.g.*, a KG class) to an (entity) column (**CTA task** - Column-Type Annotation)
- Assigning a KG property to the relationship between two columns (**CPA task** - Columns-Property Annotation)

Ernesto Jiménez-Ruiz and others. **SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems**. ESWC 2020

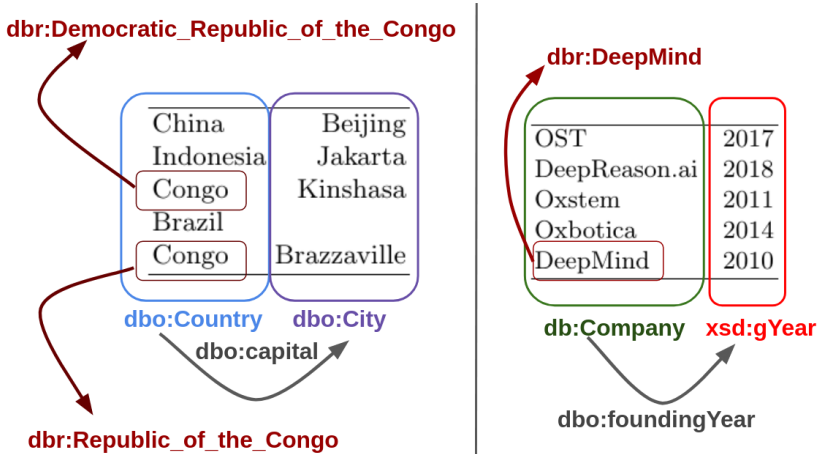
Adding Semantics to Tabular Data: Basic Tasks

- Matching a cell to a KG entity (**CEA task** - Cell-Entity Annotation)
- Assigning a semantic type (*e.g.*, a KG class) to an (entity) column (**CTA task** - Column-Type Annotation)
- Assigning a KG property to the relationship between two columns (**CPA task** - Columns-Property Annotation)

† *For a semi-automatic process, we assume the existence of a (possibly incomplete)* ***Knowledge Graph (KG)*** *relevant to the domain.*
‡ *When transforming to RDF, if no KG matching then create a fresh entity URI.*

Ernesto Jiménez-Ruiz and others. **SemTab 2019: Resources to Benchmark Tabular Data to Knowledge Graph Matching Systems**. ESWC 2020

Adding Semantics to Tabular Data: Basic Tasks (with DBpedia)



SemTab Challenge

- **Systematic evaluation** of Tabular Data to KG matching systems.
- Evaluates the **three basic tasks**: CTA, CEA and CPA.
- Relies on:
  - an **automatic** dataset generator, and
  - **manually** curated datasets.
- **Target KGs**: DBpedia, Wikidata and Schema.org.
- Co-organised and sponsored by **IBM Research**.

SemTab: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching. Collocated with the International Semantic Web Conference (ISWC): `http://www.cs.ox.ac.uk/isg/challenges/sem-tab/`

Semantic Understanding of Tabular Data: Techniques

Common Techniques

- **Pre-processing**: spelling error, stopwords, unicode fixing, etc.
- **Regular expressions** to identify data formats (*e.g.*, numbers, phones, dates, names).
- **NLP techniques** to process textually-rich fields.

Common Techniques

- **Pre-processing**: spelling error, stopwords, unicode fixing, etc.
- **Regular expressions** to identify data formats (*e.g.*, numbers, phones, dates, names).
- **NLP techniques** to process textually-rich fields.
- **Fuzzy search** over a KG
  - Via online look-up services
  - Or local indexes
- Access to the **KG's SPARQL Endpoint** (local or online)

Common Techniques

- **Pre-processing**: spelling error, stopwords, unicode fixing, etc.
- **Regular expressions** to identify data formats (*e.g.*, numbers, phones, dates, names).
- **NLP techniques** to process textually-rich fields.
- **Fuzzy search** over a KG
  - Via online look-up services
  - Or local indexes
- Access to the **KG's SPARQL Endpoint** (local or online)
- **Lexical similarity** (*e.g.*, Levenshtein)
- Word and KG **embeddings**. And **LLMs!** (More next session).

Common Knowledge Graphs

Wikidata: `https://www.wikidata.org/`

- $>$100 million entities
- Free and public (anyone can edit)

DBpedia: `https://dbpedia.org/` **(Extracted from Wikipedia)**

- $>$100 million entities
- $>$900 million triples

Google KG: `https://developers.google.com/knowledge-graph`

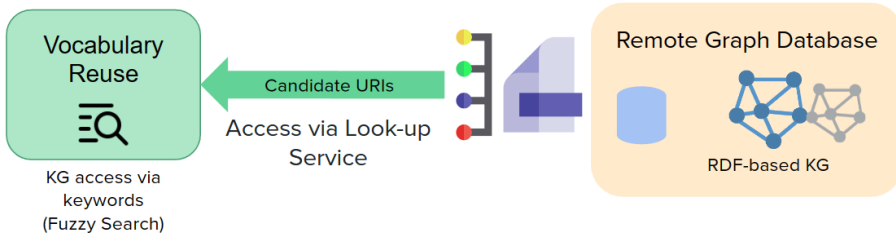
- Private, only accessible via look-up
- $>$5 billion entities

Fuzzy Search: KG look-up Services

- Given a string (*e.g.*, “Congo”)
- Return a set of candidate KG entities, *e.g.*,
  `http://dbpedia.org/resource/Republic_of_the_Congo`
  `http://dbpedia.org/resource/Congo_River`
- Typical starting point for CEA and CTA tasks
- DBPedia, Wikidata and Google KG provide look-up services via a REST API.
- Some systems have built their own local index for fuzzy search.

GitHub repositories: `https://github.com/city-knowledge-graphs`

KG look-up Services

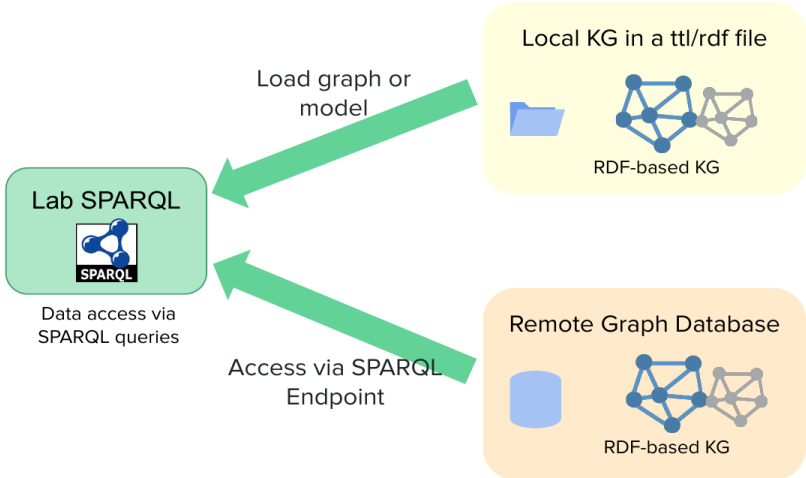


Access to KG SPARQL Endpoint

- Get additional **contextual information**:
  - Additional type information (*e.g.*, `dbr:London rdf:type dbo:City`)
  - Relationships (*e.g.*, `dbr:London dbo:country dbr:United_Kingdom`)
  - Labels (*e.g.*, `dbr:London rdfs:label "London"`)
  - Members of a given class.
- Access via **SPARQL queries** (no fuzzy search)
- Typically required for:
  - the **CPA task**
  - **disambiguation** in CTA and CEA tasks

GitHub repositories: `https://github.com/city-knowledge-graphs`

SPARQL: local and remote KG access



Lexical Processing and Similarity

- **Datatype prediction**, *e.g.*, ptype:
  `https://github.com/alan-turing-institute/ptype`
- **Spelling corrector:** `https://norvig.com/spell-correct.html`

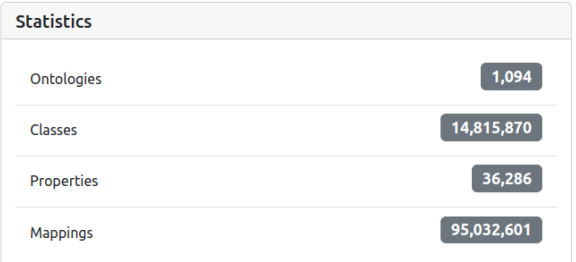
Lexical Processing and Similarity

- **Datatype prediction**, *e.g.*, ptype:
  `https://github.com/alan-turing-institute/ptype`
- **Spelling corrector:** `https://norvig.com/spell-correct.html`
- **Lexical similarity (as in today's lab):**
  - Levenshtein distance:
    `levenshtein('Congo', 'Republic of Congo')=12`
  - Jaro Winkler:
    `jaro_winkle('Congo', 'Republic of Congo')=0.0`
    `jaro_winkle('Congo', 'Congo Republic')=0.893`
  - **I-Sub**:
    `isub('Congo', 'Republic of Congo')=0.727`

PART II: Linking to other Knowledge Graphs

Ontology alignment motivation: Interoperability (i)

BioPortal, comprehensive repository of biomedical ontologies:
`https://bioportal.bioontology.org/` (*) Stats: March 2024



Whetzel PL et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011

Ontology alignment motivation: Interoperability (ii)

- An application domain can be modelled with **different points of view and purposes**
- Ontologies with **different naming and modelling conventions** exist for the same domain

Ontology alignment motivation: Interoperability (ii)

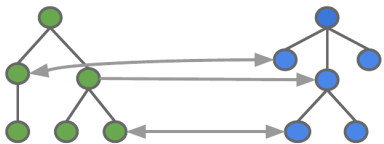
- An application domain can be modelled with **different points of view and purposes**
- Ontologies with **different naming and modelling conventions** exist for the same domain
- Aligning these ontologies will **enable interoperability** between ontology-based information systems and **data migration**
- **Reusing** vocabulary from domain ontologies is a good practice in ontology engineering

What is ontology alignment?

Ontology matching (or alignment) is the process of **finding relationships** or correspondences **between** two or more **entities** in two or more **independent ontologies**.

For example:

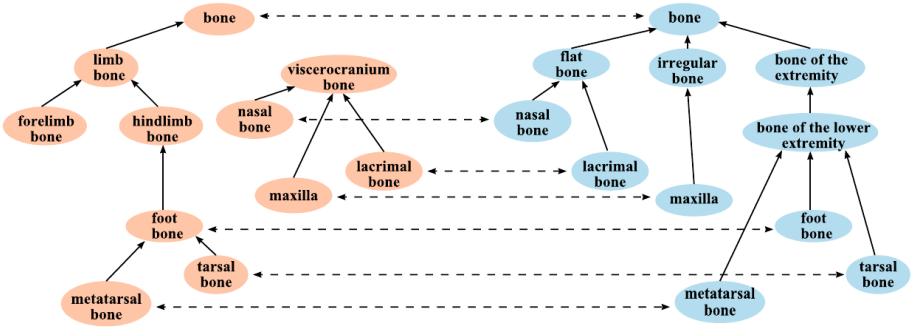
- `Person` in one ontology (`o`) is equivalent to the concept `Human` in another ontology (`o'`).
- `o:Person owl:equivalentClass o':Human`



Ontology alignment: Nomenclature

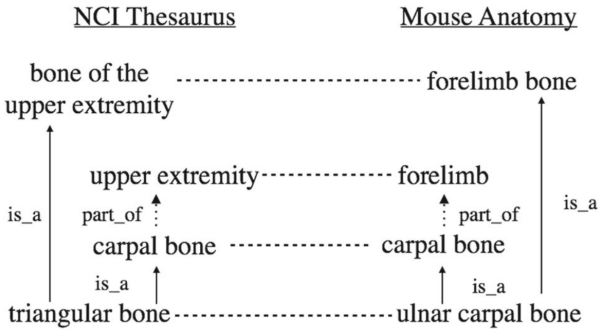
- Knowledge graph alignment as a type of **ontology alignment** or **ontology matching**.
- **To match or align or map**: the process that produces an alignment or mapping.
- **An alignment ($\mathcal{A}$)** or **mapping set ($\mathcal{M}$)**: the output of matching or aligning.
- **A mapping or match**: a single link between related entities; also called a cross reference.

Ontology alignment: Example (i)



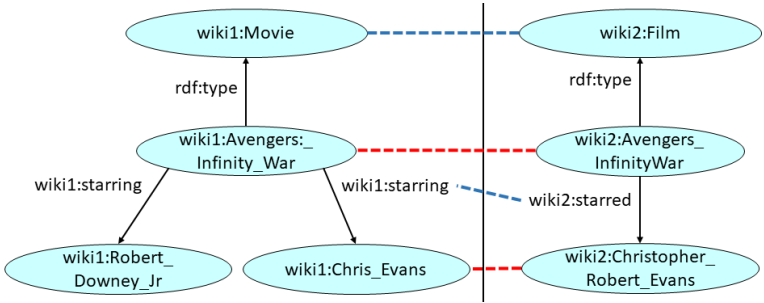
P. Lambrix and V. Ivanova. A unified approach for debugging is-a structure and mappings in networked taxonomies. Journal of Biomedical Semantics 2013

Ontology alignment: Example (ii)



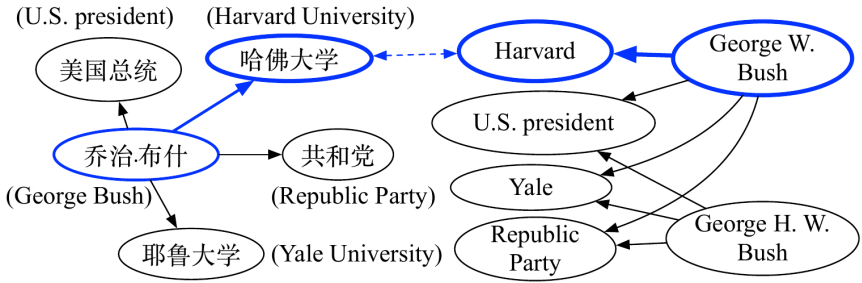
P. Kolyvakis, A. Kalousis, B. Smith, D. Kiritsis. Biomedical ontology alignment: an approach based on representation learning. Journal of Biomedical Semantics 2018

Ontology alignment: Example (iii)



S. Hertling and H Paulheim. The Knowledge Graph Track at OAEI: Gold Standards, Baselines, and the Golden Hammer Bias. ESWC 2020. `http://oaei.ontologymatching.org/2020/knowledgegraph/`

Ontology alignment: Example (iv)

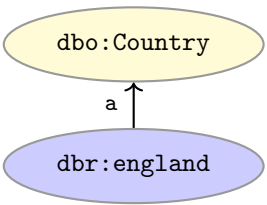


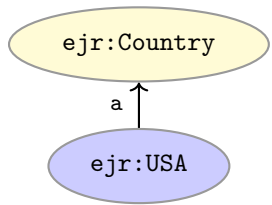
K. Xu, L. Song, Y. Feng, Y. Song, D. Yu. Coordinated Reasoning for Cross-Lingual Knowledge Graph Alignment. AAAI 2020

SPARQL Example: with alignment (and reasoning)

Return all Countries:

```
SELECT DISTINCT ?country WHERE {
    ?country rdf:type dbo:Country .
}
```

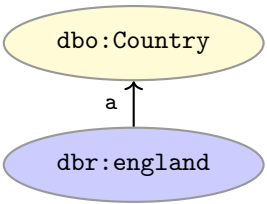


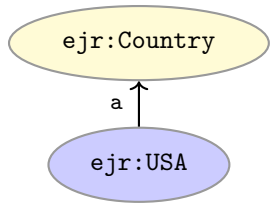


SPARQL Example: with alignment (and reasoning)

Return all Countries: Query Result= {dbr:england}

```
SELECT DISTINCT ?country WHERE {
    ?country rdf:type dbo:Country .
}
```

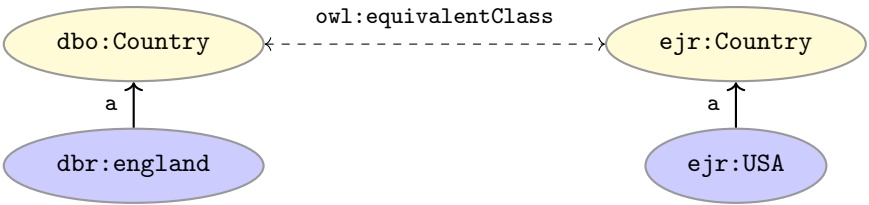




SPARQL Example: with alignment (and reasoning)

Return all Countries: Query Result= {dbr:england}

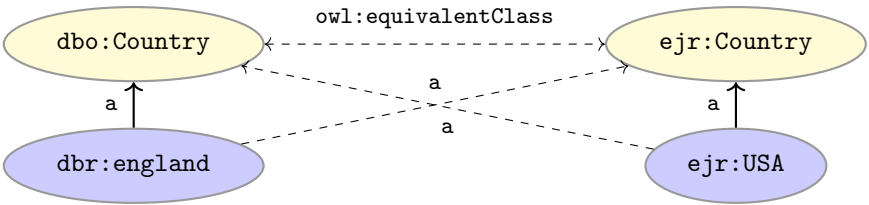
```
SELECT DISTINCT ?country WHERE {
    ?country rdf:type dbo:Country .
}
```



SPARQL Example: with alignment (and reasoning)

Return all Countries: **Query Result= {dbr:england, ejr:USA}**

```
SELECT DISTINCT ?country WHERE {
    ?country rdf:type dbo:Country .
}
```



Ontology Alignment: definitions

Ontology alignment: definition (atomic mappings)

- Basic definition in the OM community.
- An **ontology alignment** $\mathcal{M}$ (or $\mathcal{A}$) is a set of tuples $\langle e_1, e_2, n, \rho \rangle$
  - e_1, e_2 are **entities** in the input ontologies ($e_1 \in \mathcal{O}_1$ and $e_2 \in \mathcal{O}_2$)
  - n a **confidence** value between 0 and 1
  - ρ is the **semantic relationship** between e_1 and e_2
    - OM: subsumption, equivalence, disjointness
    - Life Sciences (SKOS vocabulary): broadMatch, narrowMatch, closeMatch, relatedMatch, exactMatch.

P. Shvaiko, J. Euzenat. Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering 2013

Ontology alignment: (exchange) formats

- **RDF Alignment** format (OM Community)
- A Simple Standard for Sharing Ontology Mappings (**SSSOM**)
  - `https://github.com/mapping-commons/SSSOM`
- **OWL 2 axioms**
  - Where the semantic relationship ρ is one of $\{\equiv, \sqsubseteq, \sqsupseteq, \perp\}$
    - $pizza{:}Margherita\_Pizza \equiv ejr{:}Pizza\_Margarita$
  - Confidence values n are represented as axiom annotations
  - Enables OWL 2 reasoning.
  - Strong interpretation/assumption.

Ontology alignment: OWL predicates

Predicates for OWL relationships ($\equiv$ and $\sqsubseteq$).

Mappings between classes:

- `owl:equivalentClass`
- `rdfs:subClassOf`

Mappings between properties:

- `owl:equivalentProperty`
- `rdfs:subPropertyOf`

Mappings between instances:

- `owl:sameAs`

Ontology alignment: as triples

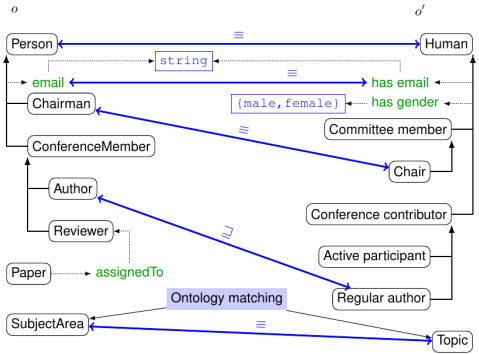
```
o:Person owl:equivalentClass o':Human

o':Regular_author rdfs:subClassOf o:Author

o:email owl:equivalentProperty o':has_email

o:OM owl:sameAs o':OntologyMatching

o:ernesto owl:sameAs o':ejimenez
```

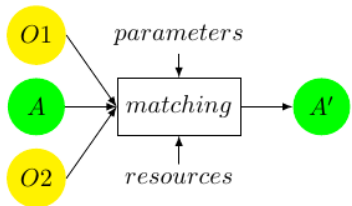


J. Euzenat, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, C. Trojahn dos San-tos. Ontology Alignment Evaluation Initiative: six years of experience. Journal on Data Semantics 2011

Ontology Alignment System

Alignment systems

- Given two input ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$ **generate an alignment** $\mathcal{A}'$ as output.
- In addition a system can get as input a **partial alignment** $\mathcal{A}$, **matching parameters** and **external resources**.



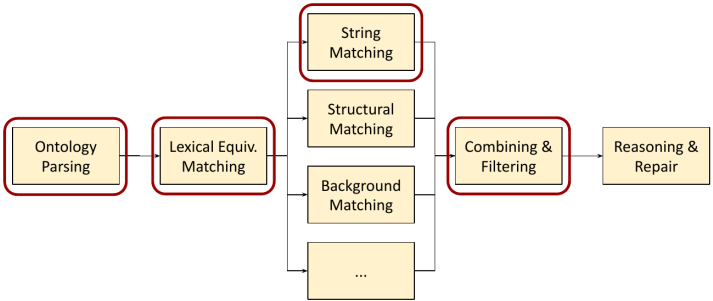
Alignment techniques (i)

- **element-level vs structure-level:** analyse entities in isolation, or how they appear together in the ontology structure.
- **syntactic vs semantic:** analyse lexical and/or structural characteristics of the entities and/or employ formal semantics
- **internal vs external:** rely solely on the information contained in the ontologies to match, or use external (background) knowledge sources to assist in the matching.

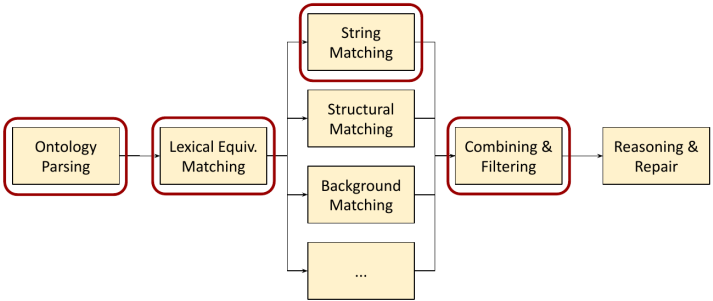
Alignment techniques (ii)

- **similarity vs logic relationship**: assert similarity between ontology entities and/or formally assert a logic relation (*e.g.*, OWL axiom).
- **atomic vs complex**: relate individual entities and/or combinations of entities (possibly in complex expressions).
- **schema vs instance**: relate schema-level entities and/or instance-level entities.
- **homogeneous vs heterogeneous**: relate only entities of the same kind or allow relations between an individual with a class, for example.

Typical alignment pipeline



Typical alignment pipeline



† **In the lab today:** we are creating a (basic) syntactic element-level (lexical) matcher, using internal information only, and producing (atomic and homogeneous) logical relationships. Applying some filtering based on lexical similarity.

Challenges (and Solutions) in Ontology Alignment

Challenges

- ✓ Large ontology size
- ✓ Rich and complex vocabularies
- ✓ Different modelling views
- ✓ Use of background knowledge

Challenges

- ✓ Large ontology size
- ✓ Rich and complex vocabularies
- ✓ Different modelling views
- ✓ Use of background knowledge
- ✓ Combination with ML techniques
- ✓ User involvement
- ✗ Need for complex mappings beyond atomic equivalence/subsumption

Solutions for large ontologies (i)

- Ontologies may be large (*i.e.*, tens of thousands of classes or **even hundreds of thousands**) like SNOMED Clinical Terms.
- The matching problem has quadratic **complexity**: Size($\mathcal{O}_1$)$\times$ Size($\mathcal{O}_2$) potential candidates.

Solutions for large ontologies (i)

- Ontologies may be large (*i.e.*, tens of thousands of classes or **even hundreds of thousands**) like SNOMED Clinical Terms.
- The matching problem has quadratic **complexity**: Size($\mathcal{O}_1$)$\times$ Size($\mathcal{O}_2$) potential candidates.
- Strategies:
  - **Pruning**: avoid comparing all entities - e.g. hash-based searching
  - **Dividing** the matching tasks into independent subtasks - parallelize
    - **Partitioning**: split into vertical blocks.
    - **Modularization**: identify overlapping self-contained sub-ontologies.

Solutions for large ontologies: hash-based search (ii)

Hash-based searching (aka inverted index):

Table 1: Inverted lexical index Lexl. For readability, index values have been split into elements of $\mathcal{O}_1$ and $\mathcal{O}_2$. '-' indicates that the ontology does not contain entities for that entry.

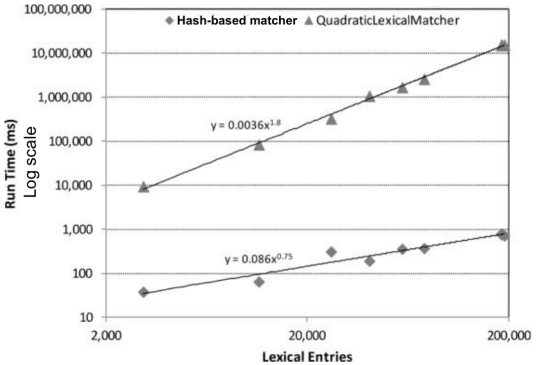
#	Index key	Index value	
		Entities $\mathcal{O}_1$	**Entities $\mathcal{O}_2$**
1	{ disorder }	$\mathcal{O}_1$:Disorder_of_pregnancy, $\mathcal{O}_1$:Disorder_of_stomach	$\mathcal{O}_2$:Pregnancy_Disorder
2	{ disorder, pregnancy }	$\mathcal{O}_1$:Disorder_of_pregnancy	$\mathcal{O}_2$:Pregnancy_Disorder
3	{ carcinoma, basaloid }	$\mathcal{O}_1$:Basaloid_carcinoma	$\mathcal{O}_2$:Basaloid_Carcinoma, $\mathcal{O}_2$:Basaloid_Lung_Carcinoma
4	{ follicul, thyroid, carcinom }	$\mathcal{O}_1$:Follicular_thyroid_carcinoma	$\mathcal{O}_2$:Follicular_Thyroid_carcinoma
5	{ hamate, lunate }	$\mathcal{O}_1$:Lunate_facet_of_hamate	-

D. Faria, et al. Tackling the challenges of matching biomedical ontologies. J. Biomed. Semant 2018
E. Jiménez-Ruiz, B. Cuenca Grau. LogMap: Logic-based and Scalable Ontology Matching. ISWC 2011

Solutions for large ontologies: hash-based search (ii)

Hash-based searching (aka inverted index):

- reduces the time complexity of the matching problem **from quadratic to linear.**

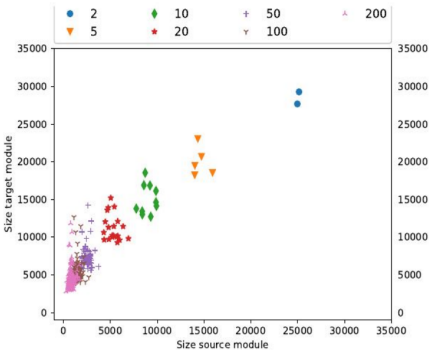


D. Faria, et al. Tackling the challenges of matching biomedical ontologies. J. Biomed. Semant 2018
E. Jiménez-Ruiz, B. Cuenca Grau. LogMap: Logic-based and Scalable Ontology Matching. ISWC 2011

Solutions for large ontologies division (iii)

Division (facilitate parallelization):

- **Partitioning**: divides ontologies into (vertical) partitions.
- **Modularization**: extracts self-contained sub-ontologies preserving logical properties.



Division of FMA-NCI into matching subtasks: from 2 modules (blue) to 200 (pink) per ontology.

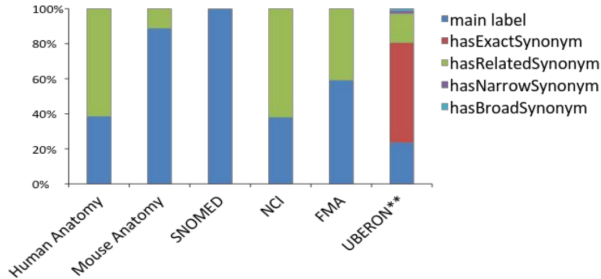
E. Jiménez-Ruiz et al. Dividing the Ontology Alignment Task with Semantic Embeddings and Logic-based Modules. ECAI 2020.

Exploiting rich and complex vocabularies (i)

How can we handle different types of labels?

UBERON_0000948

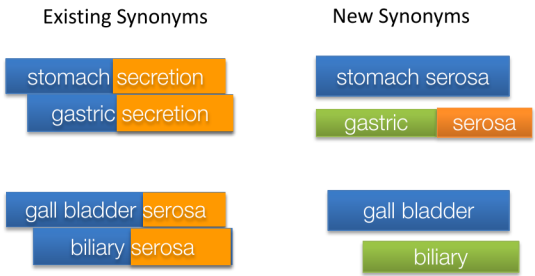
- rdfs:label: “heart”
- exact synonyms: “vertebrate heart”, “chambered heart”
- narrow synonym: “branchial heart”
- related synonym: “cardium”



C. Pesquita et al. What’s in a ‘nym’? Synonyms in Biomedical Ontology Matching 2013

Exploiting rich and complex vocabularies (ii)

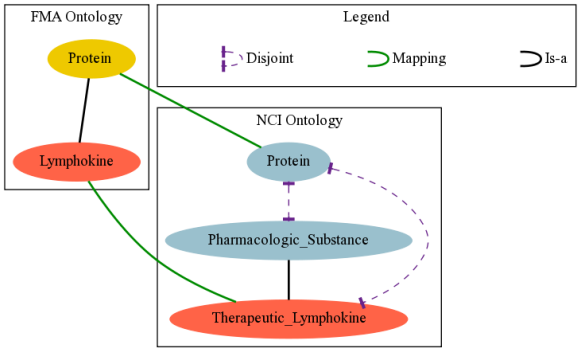
- Existing synonymous can **derive new synonyms** (see example).
- *e.g.*, “stomach” ∼ “gastric”
- *e.g.*, “gall bladder” ∼ “billiari”



C. Pesquita et al. What’s in a ‘nym’? Synonyms in Biomedical Ontology Matching 2013

Different modelling views (i)

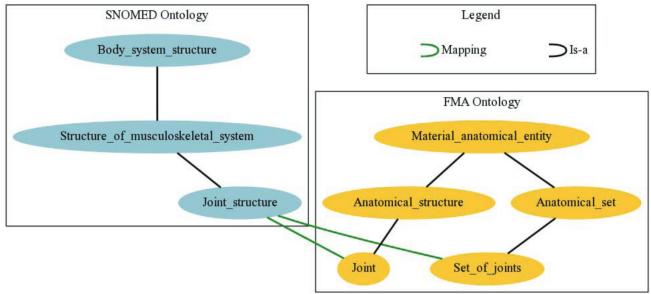
- When reasoning with $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \mathcal{M}$
- The integration of different models can cause **unsatisfiabilities** (logical errors).
- $\mathcal{M}$ is also called **incoherent**.
- Possible **sources of the errors**:
  - Different modelling views
  - Wrong mappings.



E. Santos et al. Ontology alignment repair through modularization and confidence-based heuristics. PloS one 2015
E. Jimenez-Ruiz and B. Cuenca-Grau. LogMap: Logic-based and Scalable Ontology Matching. ISWC 2011

Different modelling views (ii)

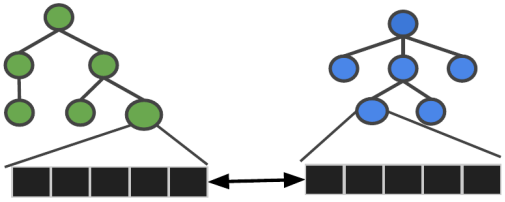
- When reasoning with $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \mathcal{M}$
- The integration of different models can lead to **unintended logical consequences** (others than unsatisfiabilities).
- Possible **sources of the errors**:
  - Different modelling views
  - Wrong mappings.



A. Solimando, E. Jiménez-Ruiz, G. Guerrini: Minimizing conservativity violations in ontology alignments: algorithms and evaluation. Knowl. Inf. Syst. 2017

Machine learning for ontology alignment

- ML models to **learn mappings**.
  - Supervised.
  - Distant-supervision.
- Source of embeddings (‡):
  - Use of **pre-trained language models** to obtain word embeddings for the entity labels.
  - **Ontology embedding** techniques.

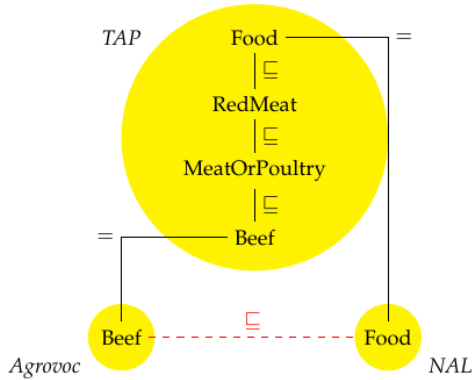


‡Embeddings: vector representation capturing the context/semantics of a word or entity.

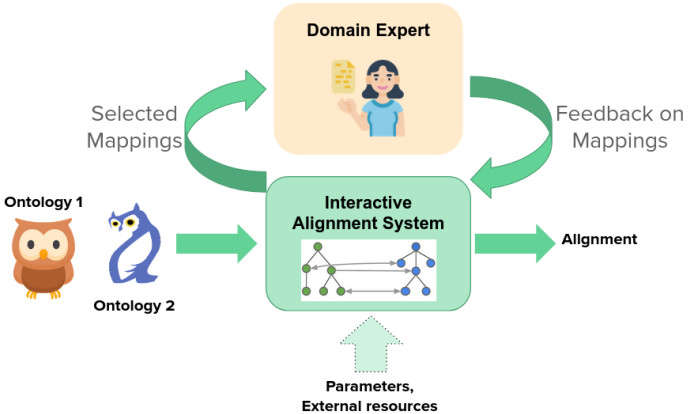
P. Kolyvakis et al. Biomedical ontology alignment: an approach based on representation learning. J. of Biomed. Semantics 2018
J. Chen et al. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. ESWC 2021
V. Iyer et al. VeeAlign: Multifaceted Context Representation Using Dual Attention for Ontology Alignment. EMNLP 2021

External resources and background knowledge

- Third ontology as **mediator**
- **WordNet** thesaurus
- **UMLS** metathesaurus (life sciences)
- **Repository** of ontologies (*e.g.*, BioPortal)
- **Pre-trained embeddings**.
- Online **multilingual translators**
- **BabelNet** multilingual semantic network.



User involvement in ontology alignment



H Li et al. User validation in ontology alignment: functional assessment and impact. KER 2019
J. da Silva et al. Alin: improving interactive ontology matching by interactively revising mapping suggestions. KER 2020

Complex ontology alignment

Links across ontologies involving complex constructors, potentially complex transformations (extends the mapping definition).

Source entity	rel.	Target construction	type
$cmt{:}ExternalReviewer$	$\equiv$	$\exists\ conference{:}invited\_by.\top$	CAE
$conference{:}Submitted\_contribution$	$\equiv$	$\exists cmt{:}submitPaper^{-}.\top$	CIAE
$cmt{:}ProgramCommitteeMember$	$\equiv$	$\exists conference{:}was\_a\_member\_of.$ $conference{:}Program\_committee$	CAT
$conference{:}Conference\_part$	$\equiv$	$\exists ekaw{:}hasPart^{-}.\ ekaw{:}Conference$	CIAT
$ekaw{:}ScientificEvent$	$\equiv$	$conference{:}Conference\_part \sqcup$ $conference{:}Conference$	union(c)
$ekaw{:}SubmittedPaper$	$\sqsupseteq$	$conference{:}Submitted\_contribution \sqcap$ $conference{:}Paper$	inters(c)
$cmt{:}hasProgramCommitteeMember$	$\equiv$	$conference{:}has\_members.$ $conference{:}Program\_committee.\top$	dom(rel)
$ekaw{:}reviewerOfPaper$	$\equiv$	$conference{:}contributes \circ conference{:}reviews$	chain(rel)
$cmt{:}writeReview$	$\equiv$	$ekaw{:}reviewWrittenBy^{-}$	inv(rel)

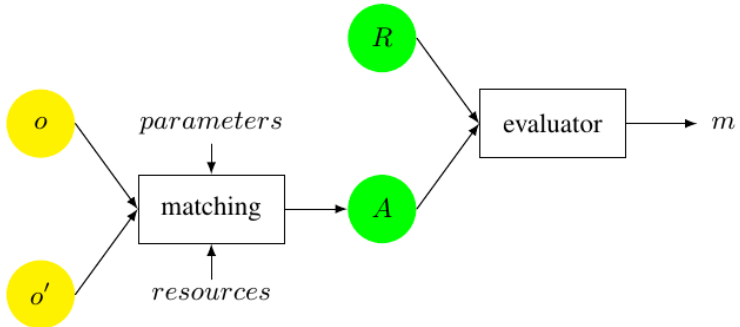
E. Thiéblin et al. Survey on complex ontology matching. Semantic Web 2020

Ontology Alignment Evaluation

Assessment of alignment systems (i)

- **Precision** and **recall** for a (system-computed) alignment $\mathcal{A}$ wrt a reference alignment or gold standard $\mathcal{R}$:
  - Precision (Pre) = $|\mathcal{A} \cap \mathcal{R}|/|\mathcal{A}|$
  - Recall (Rec) = $|\mathcal{A} \cap \mathcal{R}|/|\mathcal{R}|$
  - F-score (F) = $(2 \times \text{Pre} \times \text{Rec})/(\text{Pre} + \text{Rec})$.
- **Logical errors** of $\mathcal{A}$ wrt $\mathcal{O}_1$ and $\mathcal{O}_2$.
- Computation **times** are also considered.

Assessment of alignment systems (ii)



Ontology Alignment Evaluation Initiative (OAEI)

- **Annual Campaign** since 2004: `http://oaei.ontologymatching.org/`
- **De facto benchmark** for the OM community and driving force for tool improvement
- Collocated with the **Ontology Matching workshop** and the **International Semantic Web Conference**
- **Driven by academia**
- **Supported by industry** (*e.g.*, IBM research, Pistoia Alliance, SIRIUS)

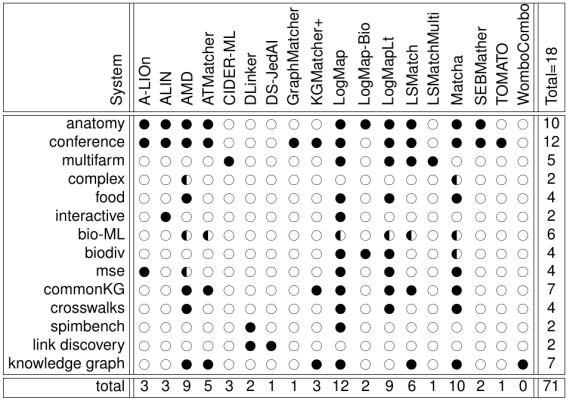
`http://om2020.ontologymatching.org/`

Ontology Alignment Evaluation Initiative (OAEI)

Common tasks and framework for the **systematic evaluation** of ontology alignment systems.

- Assessing **strengths** and **weaknesses** of alignment/matching systems
- **Comparing** performance of techniques
- Increasing **communication** among algorithm developers
- **Improving evaluation** techniques
- Helping **improve** the work on **ontology alignment**.

OAEI 2022: summary of tasks and participants



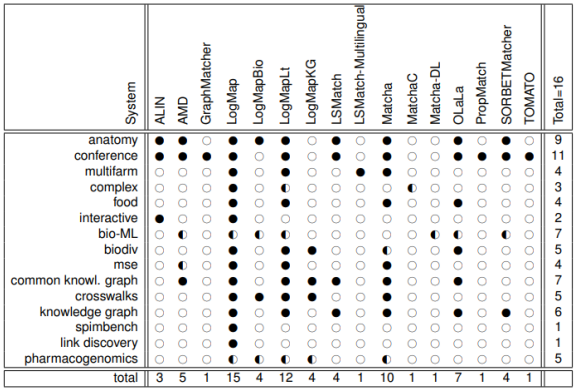
System	A-LIOn	ALIN	AMD	ATMatcher	CIDER-ML	DLinker	DS-JedAI	GraphMatcher	KGMatcher+	LogMap	LogMap-Bio	LogMapLt	LSMatch	LSMatchMulti	Matcha	SEBMather	TOMATO	WomboCombo	Total=18
anatomy	●	●	●	●	○	○	○	○	○	●	●	●	●	○	●	●	○	○	10
conference	●	●	●	●	○	○	○	●	●	●	○	●	●	○	●	●	●	○	12
multifarm	○	○	○	○	●	○	○	○	○	●	○	●	●	●	○	○	○	○	5
complex	○	○	◐	○	○	○	○	○	○	○	○	○	○	○	◐	○	○	○	2
food	○	○	●	○	○	○	○	○	○	●	○	●	○	○	●	○	○	○	4
interactive	○	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	2
bio-ML	○	○	◐	◐	○	○	○	○	○	◐	○	◐	◐	○	◐	○	○	○	6
biodiv	○	○	○	○	○	○	○	○	○	●	●	●	○	○	◐	○	○	○	4
mse	●	○	◐	○	○	○	○	○	○	●	○	●	○	○	●	○	○	○	4
commonKG	○	○	●	●	○	○	○	○	●	●	○	●	●	○	●	○	○	○	7
crosswalks	○	○	●	○	○	○	○	○	○	●	○	●	○	○	●	○	○	○	4
spimbench	○	○	○	○	○	●	○	○	○	●	○	○	○	○	○	○	○	○	2
link discovery	○	○	○	○	○	●	●	○	○	○	○	○	○	○	○	○	○	○	2
knowledge graph	○	○	●	●	○	○	○	○	●	●	○	○	●	○	●	○	○	●	7
total	3	3	9	5	3	2	1	1	3	12	2	9	6	1	10	2	1	0	71

Summary:

- 18 systems
- 14 evaluation tracks with different flavours
- One focusing on ML systems

Results of the Ontology Alignment Evaluation Initiative 2022. Ontology Matching workshop.

OAEI 2023: summary of tasks and participants



System	ALIN	AMD	GraphMatcher	LogMap	LogMapBio	LogMapLt	LogMapKG	LSMatch	LSMatch-Multilingual	Matcha	MatchaC	Matcha-DL	OLaLa	PropMatch	SORBETMatcher	TOMATO	Total=16
anatomy	●	●	○	●	●	●	○	●	○	●	○	○	●	○	●	○	9
conference	●	●	●	●	○	●	○	●	○	●	○	○	●	●	●	●	11
multifarm	○	○	○	●	○	●	○	○	●	●	○	○	○	○	○	○	4
complex	○	○	○	●	○	◐	○	○	○	○	◐	○	○	○	○	○	3
food	○	○	○	●	○	●	○	○	○	●	○	○	●	○	○	○	4
interactive	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	2
bio-ML	○	◐	○	◐	◐	◐	○	○	○	○	○	◐	◐	○	◐	○	7
biodiv	○	○	○	●	○	●	●	○	○	◐	○	○	●	○	○	○	5
mse	○	◐	○	●	○	●	○	○	○	●	○	○	○	○	○	○	4
common knowl. graph	○	●	○	●	○	●	●	○	○	●	○	○	●	○	○	○	7
crosswalks	○	○	○	●	●	●	●	○	○	●	○	○	○	○	○	○	5
knowledge graph	○	○	○	●	○	●	○	●	○	●	○	○	●	○	●	○	6
spimbench	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	1
link discovery	○	○	○	●	○	○	○	○	○	○	○	○	○	○	○	○	1
pharmacogenomics	○	○	○	◐	◐	◐	◐	○	○	◐	○	○	○	○	○	○	5
total	3	5	1	15	4	12	4	4	1	10	1	1	7	1	4	1	

Summary:

- 16 systems
- 15 evaluation tracks with different flavours
- One focusing on ML systems

Results of the Ontology Alignment Evaluation Initiative 2023. Ontology Matching workshop.

OAEI datasets for the Lab today

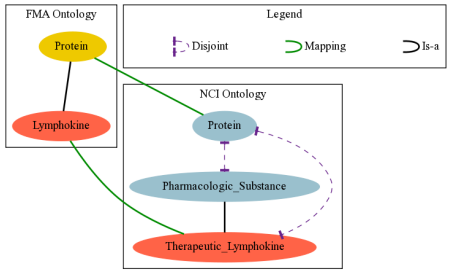
Conference. This dataset aligns 7 ontologies from the same domain (conference organization): Cmt, ConfTool, Edas, Ekaw, Iasted, Sigkdd, Sofsem.

Anatomy. This evaluation dataset consists of finding an alignment between the Adult Mouse Anatomy (mouse) and a part of the NCI Thesaurus describing the human anatomy (human).

Logical Errors in Ontology Alignment

Incoherent mappings

- A set of Mappings $\mathcal{M}$ wrt $\mathcal{O}_1$ and $\mathcal{O}_2$ is **incoherent** if
  - there exists a class A such that
  - $\mathcal{O}_1 \cup \mathcal{O}_2 \cup \mathcal{M} \models A \sqsubseteq \bot$, and
  - $\mathcal{O}_1 \cup \mathcal{O}_2 \not\models A \sqsubseteq \bot$.



Assessing ontology alignment incoherence

- System generated (OAEI)
- Mapping repositories:
  - UMLS Metathesaurus
  - BioPortal
- Gold standards

Assessment of system-computed mappings

Top OAEI, or very precise mapping sets in OAEI largebio track.

System	FMA-NCI		FMA-SNOMED		SNOMED-NCI (*)	
	Unsat.	Ratio	Unsat.	Ratio	Unsat.	Ratio
LogMap	3	0.02%	0	0%	≥1	≥0.001%
AML	2	0.01%	0	0%	≥535	≥0.6%
LogMapLt	26,429	19%	4,938	1.3%	≥305,648	≥82%
ServOMap	48,743	27%	273,242	71%	≥313,643	≥84%
GOMMA	5,574	4%	10,752	3%	≥266,051	≥71%
YAM++	50,550	29%	106,107	28%	≥269,107	≥72%

(*) Using ELK reasoner instead of HermiT.

E. Jiménez-Ruiz et al. **Evaluating Mapping Repair Systems with Large Biomedical Ontologies.** Description Logics 2013

Assessment of mapping repositories

- **UMLS** (Unified Medical Language System) Metathesaurus
  - Integrates more than 200 thesauri and ontologies
  - Contains more than 14 million names
  - Contains more than 3.8 million concepts
- **BioPortal**
  - Contains more than 1,000 ontologies
  - Represent a network of ontologies
  - More than 75 million mappings are available
  - Also includes user-submitted alignments

E. Jiménez-Ruiz, et al. **Logic-based assessment of the compatibility of UMLS ontology sources**. J. Biomed. Semantics 2011
D. Faria, E. Jiménez-Ruiz, et al. **Towards annotating potential incoherences in BioPortal mappings**. ISWC 2014.

Assessment of UMLS (mappings)

– Assessment of the integration of FMA, NCI and SNOMED CT ontologies via UMLS-based alignments

	UMLS alignments		
Ontologies	**#**	**Unsatisfiabilities**	**Ratio**
FMA-NCI	3,024	655	0.5%
FMA-SNOMED	9,072	6,179	2.5%
SNOMED-NCI (*)	19,622	$\geq$20,944	$\geq$5.6%

(*) Using ELK reasoner instead of HermiT.

E. Jiménez-Ruiz, et al. **Logic-based assessment of the compatibility of UMLS ontology sources**. J. Biomed. Semantics 2011
E. Jiménez-Ruiz, Bernardo Cuenca Grau. **LogMap: Logic-based and Scalable Ontology Matching**. ISWC 2011.

Assessment of BioPortal (mappings)

	BioPortal alignments	
Ontologies	**#**	**Logical errors**
BDO-NCIT	1,636	34,341
CCONT-NCIT	2,097	50,304
EFO-NCIT	2,507	60,347
EP-FMA	78,489	210
EP-NCIT	2,465	14,687
MA-FMA	961	850
OMIM-NCIT	5,178	70,172
SDO-EP	135	44
UBERON-FMA	1,932	4,753
ZFA-EFO	427	913
ZFA-UBERON	724	104

D. Faria, E. Jiménez-Ruiz, et al. **Towards annotating potential incoherences in BioPortal mappings**. ISWC 2014.

Assessment of gold standards

- **2004 EON** Ontology Alignment **Contest** (pre-OAEI)
- Ontologies in the domain of **bibliographic references**.
- 4 ontologies: $\mathcal{O}_{INR}$, $\mathcal{O}_{MIT}$, $\mathcal{O}_{UMBC}$ and $\mathcal{O}_{AIFB}$.
- Reasoning with $\mathcal{O}_{INR} \cup \mathcal{O}_{MIT} \cup \mathcal{M}_{GS}$ led to consequences like:
  $\mathcal{O}_{MIT} : TechnicalReport \sqsubseteq \mathcal{O}_{INR} : Date$

$$\mathcal{O}_{\text{INR}} : \text{year} \sqsubseteq \mathcal{O}_{\text{MIT}} : \text{hasYear}$$
$$\mathcal{O}_{\text{MIT}} : \text{TechnicalReport} \sqsubseteq \geqslant \mathcal{O}_{\text{MIT}} : hasYear\ 1.Literal$$
$$\mathcal{O}_{\text{INR}} : \text{TechnicalReport} \sqsubseteq \mathcal{O}_{\text{MIT}} : \text{Technicalreport}$$
$$\mathcal{O}_{\text{INR}} : \text{year}\ \ \text{hasDomain}\ \ \mathcal{O}_{\text{INR}} : \text{Date}$$

E. Jiménez-Ruiz et al.: **Ontology Integration Using Mappings: Towards Getting the Right Logical Consequences**. ESWC'09

How to fix unintended consequences

1. Detect unintended consequences (*e.g.*, classification).
2. Perform the repair:
   - Remove mappings [1,2,3].
   - Modify ontologies [4,5]
   - Be aware of the logical impact [6].

1. E. Jiménez-Ruiz et al.: **Ontology Integration Using Mappings: Towards Getting the Right Logical Consequences**. ESWC'09
2. E. Jiménez-Ruiz, Bernardo Cuenca Grau. **LogMap: Logic-based and Scalable Ontology Matching**. ISWC 2011.
3. A. Solimando, E. Jiménez-Ruiz, G. Guerrini: **Minimizing conservativity violations in ontology alignments: algorithms and evaluation**. Knowl. Inf. Syst. 2017
4. C. Pesquita et al. **To repair or not to repair: reconciling correctness and coherence in ontology reference alignments**. OM 2013
5. V. Ivanova et al. **A Unified Approach for Aligning Taxonomies and Debugging Taxonomies and Their Alignments**. ESWC 2013
6. D. Faria, E. Jiménez-Ruiz, et al. **Towards annotating potential incoherences in BioPortal mappings**. ISWC 2014.

How to fix unintended consequences

1. Detect unintended consequences (*e.g.*, classification).
2. Perform the repair:
   - **Remove mappings** [1,2,3].
   - Modify ontologies [4,5]
   - Be aware of the logical impact [6].

1. E. Jiménez-Ruiz et al.: **Ontology Integration Using Mappings: Towards Getting the Right Logical Consequences**. ESWC'09
2. E. Jiménez-Ruiz, Bernardo Cuenca Grau. **LogMap: Logic-based and Scalable Ontology Matching**. ISWC 2011.
3. A. Solimando, E. Jiménez-Ruiz, G. Guerrini: **Minimizing conservativity violations in ontology alignments: algorithms and evaluation**. Knowl. Inf. Syst. 2017
4. C. Pesquita et al. **To repair or not to repair: reconciling correctness and coherence in ontology reference alignments**. OM 2013
5. V. Ivanova et al. **A Unified Approach for Aligning Taxonomies and Debugging Taxonomies and Their Alignments**. ESWC 2013
6. D. Faria, E. Jiménez-Ruiz, et al. **Towards annotating potential incoherences in BioPortal mappings**. ISWC 2014.

Applications of Ontology Alignment

Lung Cancer Assistant (LCA)

- An **ontology-based system** which provides decision support for lung cancer treatment
- LCA exploits the English Lung Cancer Dataset (LUCADA)

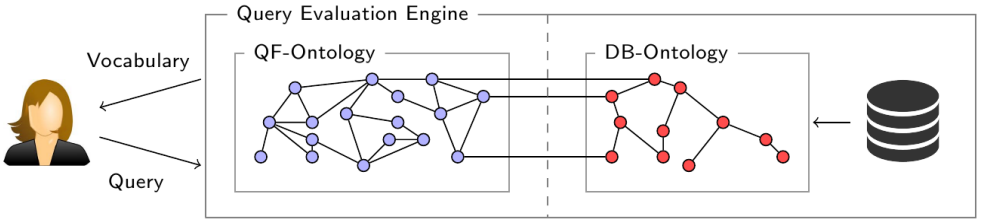
M. Berkan Sesen et al. Lung Cancer Assistant: a hybrid clinical decision support application for lung cancer care. Journal of the Royal Society Interface. 2014.

Lung Cancer Assistant (LCA)

- An **ontology-based system** which provides decision support for lung cancer treatment
- LCA exploits the English Lung Cancer Dataset (LUCADA)
- **LUCADA ontology** represents the semantic layer of the LCA,
- Required **alignment with SNOMED CT**
  - to facilitate interoperability with NHS systems

M. Berkan Sesen et al. Lung Cancer Assistant: a hybrid clinical decision support application for lung cancer care. Journal of the Royal Society Interface. 2014.

Ontology-based Data Access: Oil & Gas (EU Optique project)

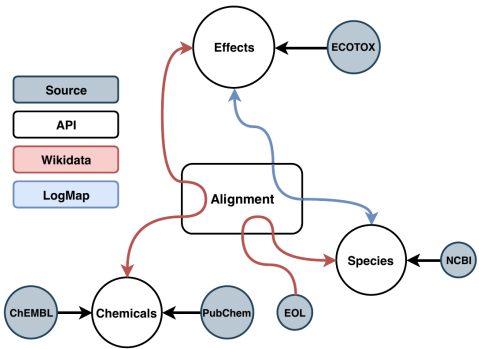


† **Coursework**: Similar setting. Ontology describing tabular data, alignment with the `pizza.owl` ontology to enable the formulation of queries with the vocabulary of `pizza.owl`.

A. Solimando, E. Jiménez-Ruiz and G, Guerrini. Minimizing conservativity violations in ontology alignments: algorithms and evaluation. Knowledge and Information Systems 2017
E. Kharlamov et al. Ontology Based Data Access in Statoil. Journal of Web Semantics 2017

Ecotoxicological Effect Prediction

- **KG Construction** for Ecotoxicological Effect Prediction
- **Integration of several resources** relevant to species and chemicals.
- ECOTOX contains data (experiments) about pairs chemical-species



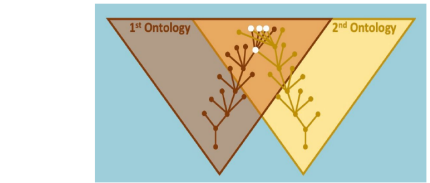
E. B. Myklebust, E. Jimenez-Ruiz et al. Knowledge Graph Embedding for Ecotoxicological Effect Prediction. ISWC In-Use 2019.
E. B. Myklebust, E. Jimenez-Ruiz et al. Prediction of Adverse Biological Effects of Chemicals Using Knowledge Graph Embeddings. Sem Web journal 2022. https://github.com/NIVA-Knowledge-Graph/

Laboratory analytics domain: Pistoia Alliance

Pistoia Alliance (Ontologies Mapping project)

- Not-for-profit alliance of **life science companies**, vendors, publishers, and academics.
- Motivation: better **integration**, **understanding** and **analysis of data**
- Interest in Semantic Web technologies and **ontology alignment**.
- OA has an important role in life sciences to achieve other tasks (*i.e.*, **drug-drug interactions**)
- OM project now integrated within a FAIR implementation project: `https://fairtoolkit.pistoiaalliance.org/`

Pistoia Alliance partners and collaborators (from OM project)



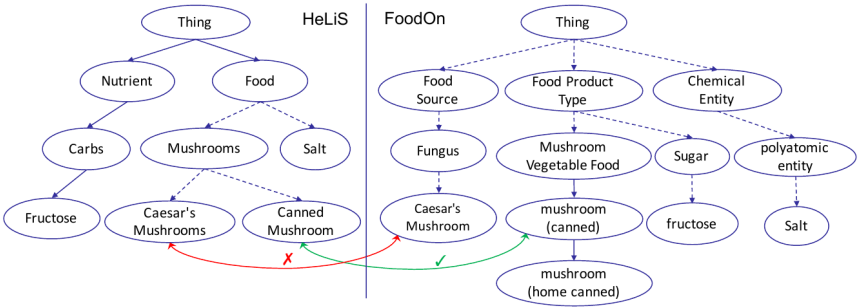
EMBL European Molecular Biology Laboratory · Allotrope Foundation · OSTHUS success with R&D · OAEI · Roche · Genentech · gsk GlaxoSmithKline · Bristol-Myers Squibb · AstraZeneca · BAYER · SciBite The language of science · abbvie · ELSEVIER · eagle genomics · NOVARTIS · Linguamatics

PM: Ian Harrow

Current funders, partners & collaborators

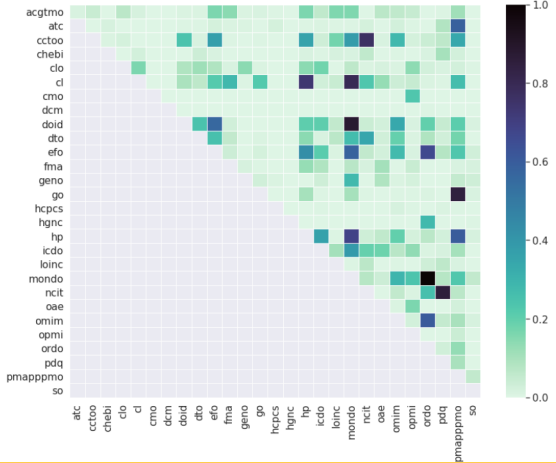
Matching Disease and Phenotype Ontologies in the Ontology Alignment Evaluation Initiative. J. Biomedical Semantics 2018
I. Harrow et al. Ontology mapping for semantically enabled applications. Drug Discovery Today, 2019

Alignment of food ontologies (Samsung Research UK)



J. Chen et al. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. ESWC 2021

Holistic Ontology Matching for Precision Medicine



Marta Contreiras Silva et al.: Matching Multiple Ontologies to Build a Knowledge Graph for Personalized Medicine. ESWC 2022

Acknowledgements

Acknowledgements

- Co-organisers of the SWAT4(HC)LS 2019 tutorial on **Ontology Matching in the Biomedical Domain**.
  - `https://tinyurl.com/tutorial-ontology-alignment`
- Ontology matching workshop organisers.
- Ontology Alignment Evaluation Initiative (OAEI) organisers.
- LogMap project contributors.
- Pistoia Alliance.
- Samsung Research UK.
- SIRIUS Lab (Norway).
- Norwegian Institute for Water Research (NIVA).

Laboratory Session

Laboratory

- Option A: a system to match tabular data to a KG.
- Option B: a system for ontology alignment.

This could also be your choice for the **final project**.

Project submission

- Students need to work on a **project** (max 2 students per group). There are two options:
  - Create a (simple) system that performs KG to KG alignment.
  - Create a (simple) system that performs CSV to KG matching.
- Submission:
  - **When:** June 16, 23:59 CEST
  - **What:** a link to the GitHub repository where the system codes are
  - **How:** via a Google form (see GitHub)