CITY
UNIVERSITY OF LONDON
EST 1894

Ontology (Knowledge Graph) Alignment

Ernesto Jiménez-Ruiz

Lecturer in Artificial Intelligence

Before we start...

Additional projects

- Data Science team at **Perspectum** (`https://perspectum.com/`)
  - A **knowledge graph to orchestrate** the multi-organ quantitative assessment in **long-COVID**.
  - The data is currently exchanged via CSV and JSON files.
  - More information in moodle (‘Potential MSc Projects’).
  - Contact: `valentina.carapella@perspectum.com` and myself.

Additional projects

- Data Science team at **Perspectum** (`https://perspectum.com/`)
  - A **knowledge graph to orchestrate** the multi-organ quantitative assessment in **long-COVID**.
  - The data is currently exchanged via CSV and JSON files.
  - More information in moodle ('Potential MSc Projects').
  - Contact: `valentina.carapella@perspectum.com` and myself.

- **Kew Gardens** - Medicinal Plant Names Services
  - Automatic construction of a Knowledge Graph.
  - (Semantic) entity disambiguation.
  - Contact: Tillman Weyde and myself.

Where are we?

- ✓ Introduction.
- ✓ RDF-based knowledge graphs.
- ✓ SPARQL 1.0
- ✓ RDFS Semantics and RDF(S)-based knowledge graphs.
- ✓ OWL (2) ontology language. Focus on modelling.
- ✓ Application to Data Science.
- ✓ OWL 2 Profiles, SPARQL 1.1 and Entailment Regimes

Where are we?

✓ Introduction.
✓ RDF-based knowledge graphs.
✓ SPARQL 1.0
✓ RDFS Semantics and RDF(S)-based knowledge graphs.
✓ OWL (2) ontology language. Focus on modelling.
✓ Application to Data Science.
✓ OWL 2 Profiles, SPARQL 1.1 and Entailment Regimes
8. **Ontology (Knowledge Graph) Alignment** (today)

Where are we?

✓ Introduction.
✓ RDF-based knowledge graphs.
✓ SPARQL 1.0
✓ RDFS Semantics and RDF(S)-based knowledge graphs.
✓ OWL (2) ontology language. Focus on modelling.
✓ Application to Data Science.
✓ OWL 2 Profiles, SPARQL 1.1 and Entailment Regimes

8. **Ontology (Knowledge Graph) Alignment** (today)
9. Machine Learning and Knowledge Graphs (March 24).
10. Graph Database Solutions and Invited Talks (March 31).

Ontologies

Why do we need ontologies?

- Ontologies **standardize**, define and structure concepts in a domain
- They are essential for **FAIR** (Findable, Accessible, Interoperable, Reusable) data:
  - Using standard vocabularies to describe data is key for **Findability**, **Interoperability** and **Reusability**
  - A hierarchical structure improves **Findability** and facilitates **Interoperability**
  - Having a public knowledge model is key for **Accessibility**

FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence (2020)

What ontologies are good for?

- **Independence** of logical/physical schema
- Formulation of queries closer to **domain experts**
- **Incomplete** and **semi-structured** data
- Help identify and resolve **disagreements** in the domain
- Integration of **heterogeneous** sources

What ontologies are good for?

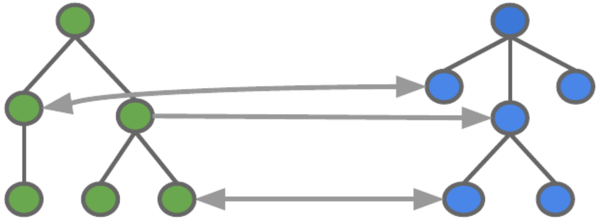
– **Independence** of logical/physical schema
– Formulation of queries closer to **domain experts**
– **Incomplete** and **semi-structured** data
– Help identify and resolve **disagreements** in the domain (‡)
– Integration of **heterogeneous** sources (‡)

‡ **Ontology alignment** will play a key role

Ontology Alignment

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.



Ontology alignment motivation: Interoperability

– 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

– 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

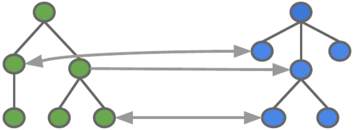
Ontology alignment: Nomenclature

– Knowledge graph alignment as a type of **ontology alignment** or **ontology matching**.

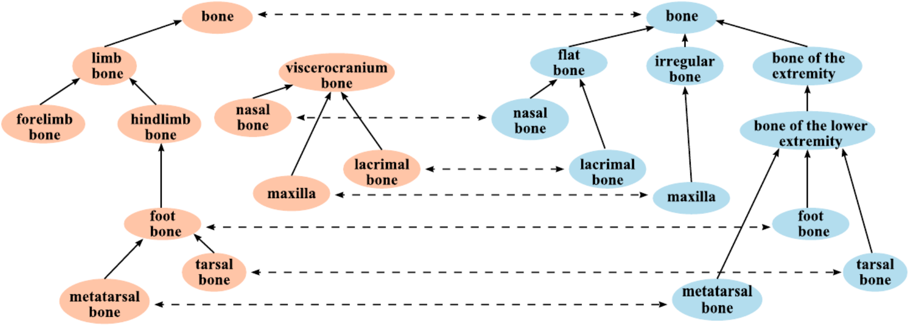
– **To match or align or map**: the process that produces an alignment or mapping.

– **An alignment or mapping set**: the final output of matching or aligning.

– **A mapping or match**: a single link between related entities; also called a cross reference.

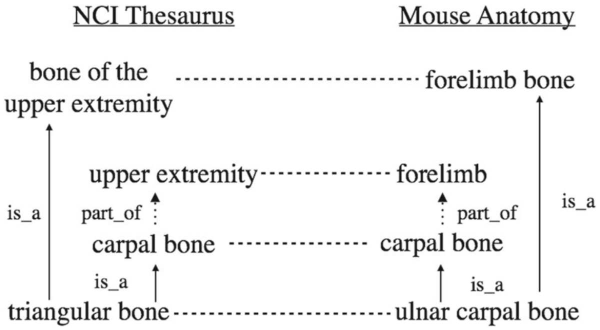


Ontology alignment: Example (i)



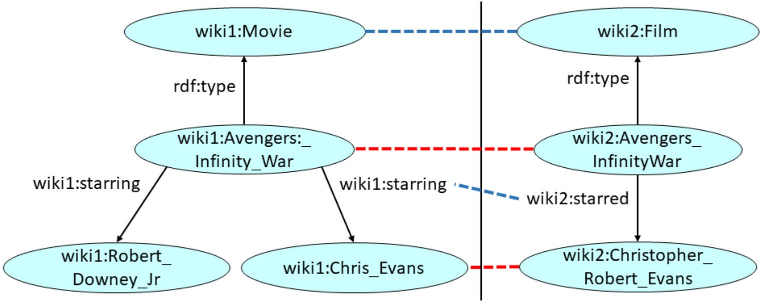
P. Lambrix and V. Ivanova. A unified approach for debugging is-a structure and mappings in networked taxonomies. Journal of Biomedical Semantics 2013

Ontology alignment: Example (ii)



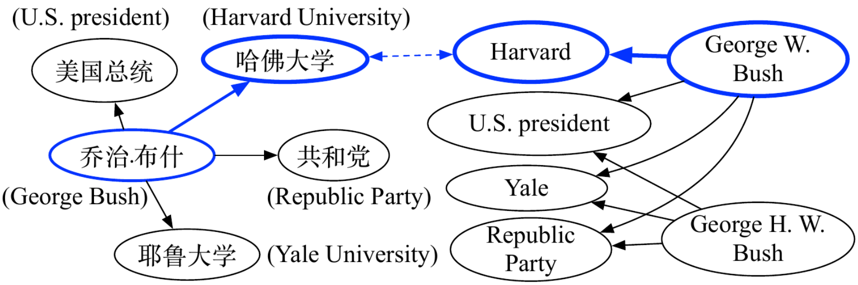
P. Kolyvakis, A. Kalousis, B. Smith, D. Kiritsis. Biomedical ontology alignment: an approach based on representation learning. Journal of Biomedical Semantics 2018

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

Ontology alignment: definition (atomic alignment)

- 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 (*e.g.*, subsumption, equivalence or disjointness)

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

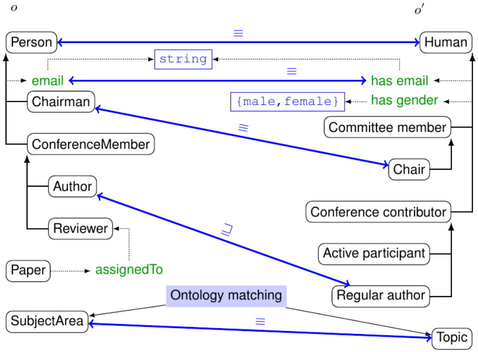
Ontology alignment: definition (atomic alignment)

- 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 (*e.g.*, subsumption, equivalence or disjointness)
- Can be **formalized as OWL 2 axioms**
  - Where the semantic relationship ρ is one of $\{\equiv, \sqsubseteq, \sqsupseteq, \bot\}$
  - Confidence values n are represented as axiom annotations

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

Ontology alignment: formalization example

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

o':Regular_author rdfs:subClassOf o:Author

o1:Person rdfs:subClassOf o2:Human

o:email owl:equivalentProperty o':has_email

o:OM owl:sameAs o':OntologyMatching

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

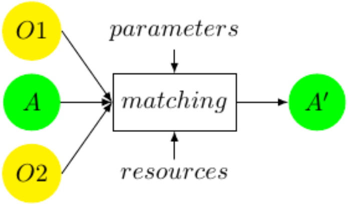


J. Euzenat, C. Meilicke, P. Shvaiko, H. Stuckenschmidt, C. Trojahn dos San-tos. Ontology Alignment Evaluation Initiative: six years of experience. Journal on Data Semantics 2011

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**.



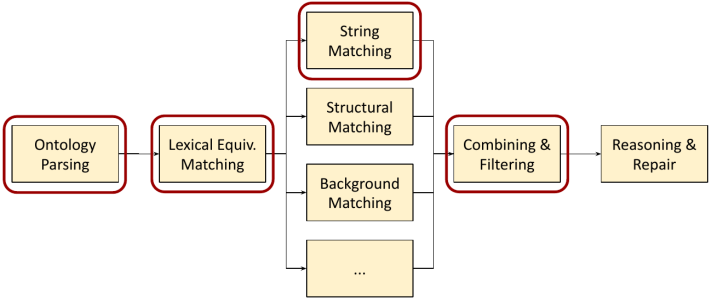
Alignment techniques (i)

- **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



† **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. Possibly 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.

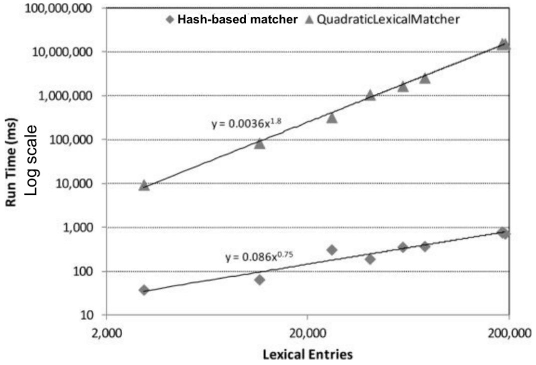
Solutions for large ontologies (i)

- 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 (ii)

Hash-based searching (aka inverted index):

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

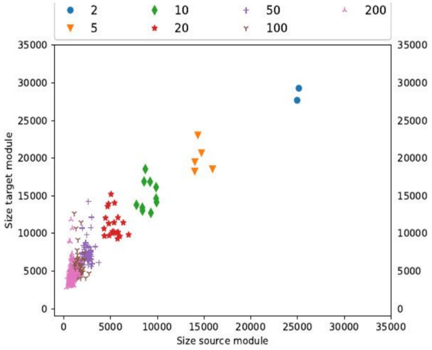


D. Faria, et al. Tackling the challenges of matching biomedical ontologies. J. Biomed. Semant 2018
E. Jiménez-Ruiz, B. Cuenca Grau. LogMap: Logic-based and Scalable Ontology Matching. ISWC 2011

Solutions for large ontologies (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.

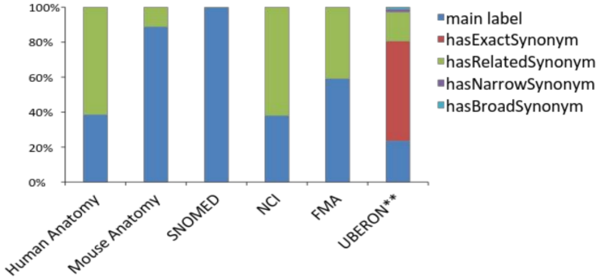
E. Jiménez-Ruiz et al. Dividing the Ontology Alignment Task with Semantic Embeddings and Logic-based Modules. ECAI 2020.

Exploiting rich and complex vocabularies (i)

How can we handle different types of labels?

UBERON_0000948

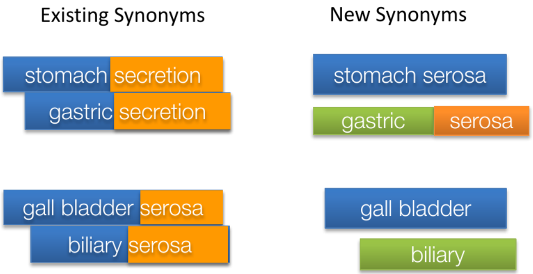
- rdfs:label: “heart”
- exact synonyms: “vertebrate heart”, “chambered heart”
- narrow synonym: “branchial heart”
- related synonym: “cardium”



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”

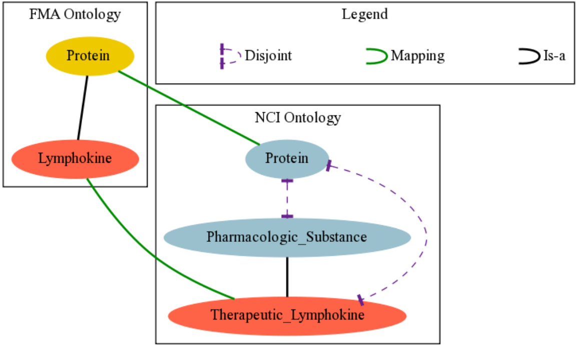


C. Pesquita et al. What’s in a ‘nym’? Synonyms in Biomedical Ontology Matching 2013

Different modelling views (i)

The integration of different models can cause **unsatisfiabilities**.

– **Possible solution:** repair/remove mappings.

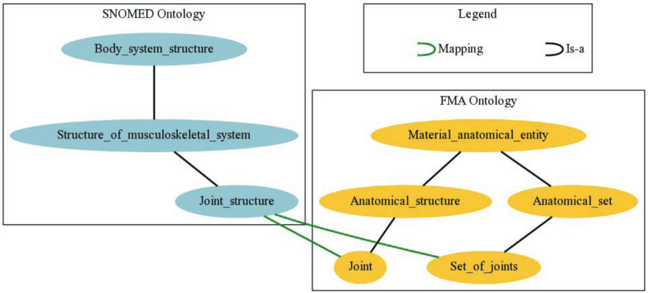


E. Santos et al. Ontology alignment repair through modularization and confidence-based heuristics. PloS one 2015
E. Jimenez-Ruiz and B. Cuenca-Grau. LogMap: Logic-based and Scalable Ontology Matching. ISWC 2011

Different modelling views (ii)

The integration of different models can lead to **unintended logical consequences** (others than unsatisfiabilities).

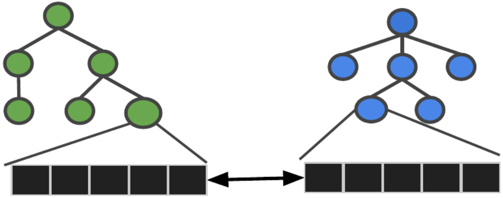
- **Possible solution:** repair/remove mappings.



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.

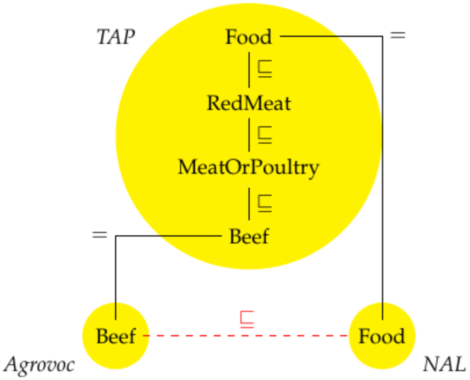


‡Embeddings: vector representation capturing the context/semantics of a word or entity.

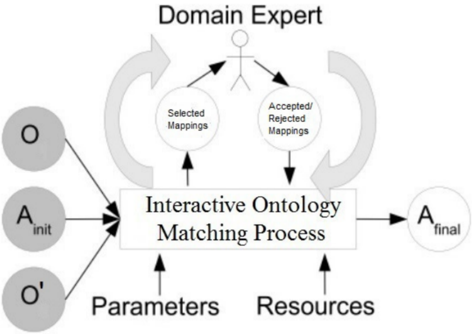
P. Kolyvakis et al. Biomedical ontology alignment: an approach based on representation learning. J. of Biomed. Semantics 2018
J. Chen et al. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. ESWC 2021

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)

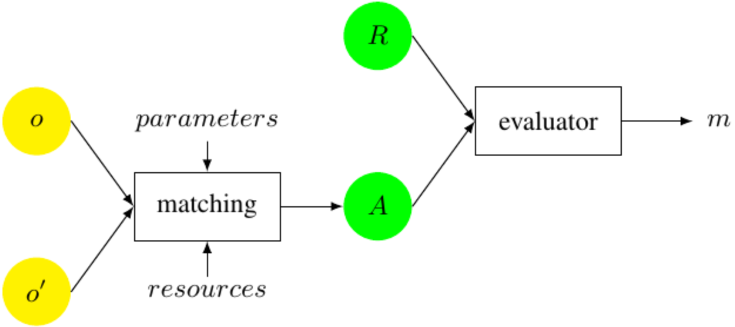
E. Thiéblin et al. Survey on complex ontology matching. Semantic Web 2020

Ontology Alignment Evaluation

Assessment of alignment systems (i)

- **Precision** and **recall** wrt 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 2021: summary of tasks and participants

System	ALIN	ALOD2Vec	AML	AMLC	AROA	ATBox	DESKMatcher	CANARD	FTRLIM	Lily	LogMap	LogMap-Bio	LogMapLt	OntoConnect	RADON	RE-miner	Silk	VeeAlign	WktMtchr	Total=19
Confidence		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓		✓	✓	✓	16
anatomy	●	●	●	○	○	●	●	○	○	●	●	●	●	●	○	○	○	○	●	11
conference	●	●	●	○	○	●	●	○	○	●	●	○	●	○	○	○	○	●	●	10
multifarm	○	○	●	○	○	○	○	○	○	◐	●	○	●	○	○	○	○	●	●	6
complex	○	○	○	●	●	○	○	●	○	○	○	○	○	○	○	○	○	○	○	3
interactive	●	○	●	○	○	○	○	○	○	○	●	○	○	○	○	○	○	○	○	3
largebio	○	◐	●	○	○	●	◐	○	○	○	●	●	●	○	○	○	○	○	●	8
phenotype	○	●	●	○	○	●	○	○	○	○	●	●	●	○	○	○	○	○	●	7
biodiv	○	◐	●	○	○	●	○	○	○	○	●	●	●	○	○	○	○	○	◐	7
spimbench	○	○	●	○	○	○	○	○	●	●	●	○	○	○	○	●	○	○	○	5
link discovery	○	○	●	○	○	○	○	○	○	○	○	○	○	○	●	○	●	○	○	3
geolink cruise	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	○	0
knowledge graph	○	●	●	○	○	●	●	○	○	○	●	●	◐	○	○	○	○	○	●	8
total	3	6	10	1	1	6	4	1	1	4	9	5	7	1	1	1	1	2	7	71

Results of the Ontology Alignment Evaluation Initiative 2020. Ontology Matching workshop.

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)

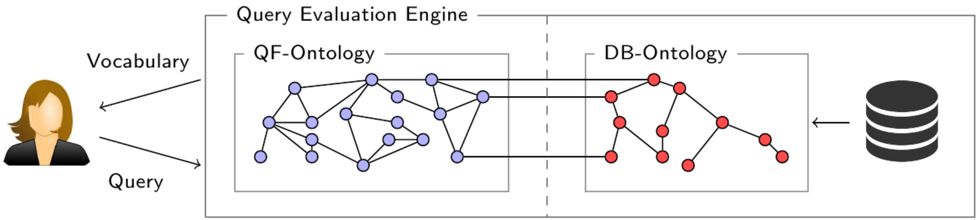
M. Berkan Sesen et al. Lung Cancer Assistant: a hybrid clinical decision support application for lung cancer care. Journal of the Royal Society Interface. 2014.

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)



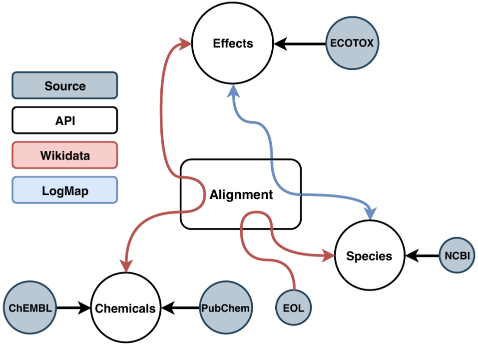
† **Coursework**: Similar setting. Ontology describing tabular data, alignment with the `pizza.owl` ontology to enable the formulation of queries with the vocabulary of `pizza.owl`.

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.
`https://github.com/NIVA-Knowledge-Graph/`

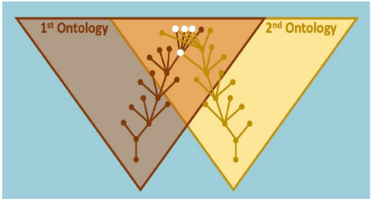
Laboratory analytics domain: Pistoia Alliance

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**.

I. Harrow et al. 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

Pistoia Alliance partners and collaborators



EMBL European Molecular Biology Laboratory, Allotrope Foundation, OSTHUS success with R&D, OAEI

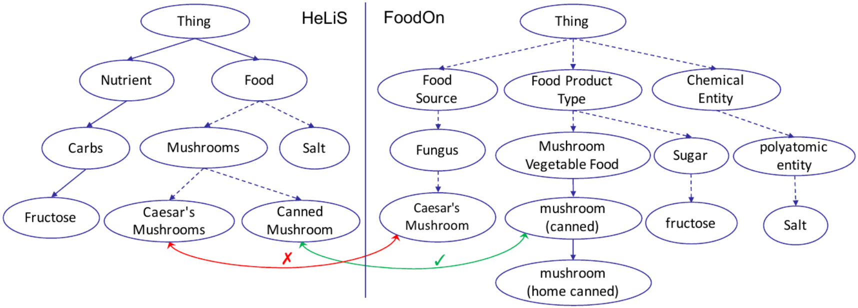
Roche, Genentech, gsk GlaxoSmithKline, Bristol-Myers Squibb

AstraZeneca, abbvie, eagle genomics, BAYER, NOVARTIS, SciBite The language of science, ELSEVIER, Linguamatics

PM: Ian Harrow

Current funders, partners & collaborators

Alignment of food ontologies (Samsung Research UK)



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

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

– Lexical Matcher

– Evaluation over OAEI tasks