WEB OF DATA

FATIHA SAÏS

UNIVERSITÉ PARIS SACLAY
MASTER 2 OF COMPUTER SCIENCE – DATA SCIENCE

COURSE OUTLINE

- **Introduction to linked data principles**
- Linked data publication
- Ontology alignment
- Data Linking
- Dataset interoperability and validation
- Link invalidation
- …

SOME OF HISTORY …

1960 — First idea of computer network

J.C.R Licklider
@MIT

1966 — Creation of ARPANET, first computer network

1986 — Creation of NSFNET: new protocols for transparent networks linking

1990 — First skeleton of Internet: a browser and a Web serveur

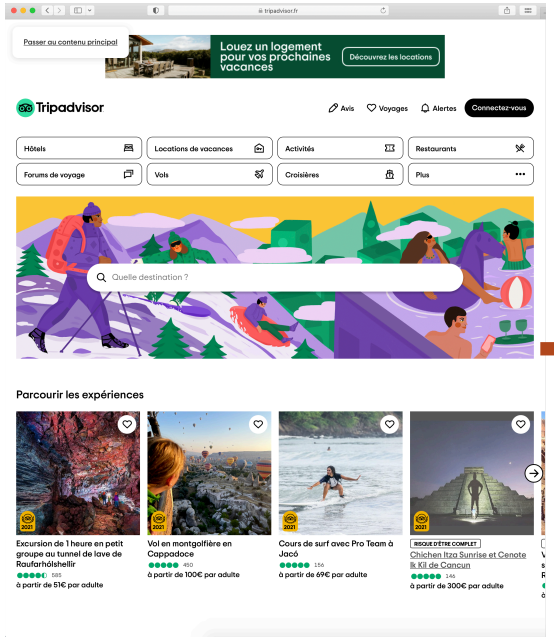
T. Berners Lee
@CERN

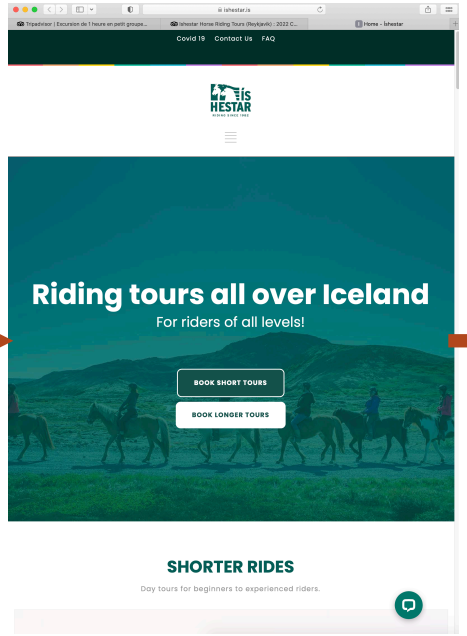
1993 — First search engines, e.g. WebCrawler

1994 — Creation of **World Wide Web** Consortium

Source: https://www.w3.org/2012/08/history-of-the-web/origins.htm

THE WORLD WIDE WEB





Unqualified links

THE WORLD WIDE WEB

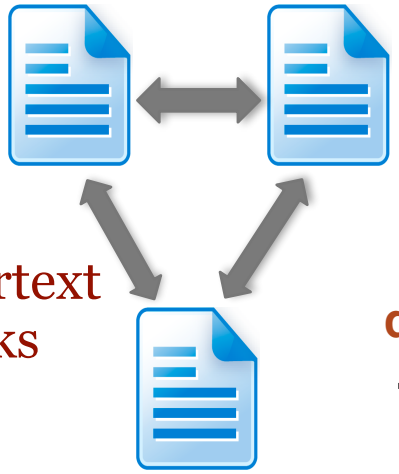
- A global network of linked documents
- A place where anyone can say anything about anything
- A vast collection of **human-readable knowledge**
- Documents are linked, but links are not **qualified**

FROM THE WEB OF DOCUMENTS TO THE WEB OF LINKED DATA

Tim Berners-Lee
The next web
Posted Mar 2009

TED

Web of documents

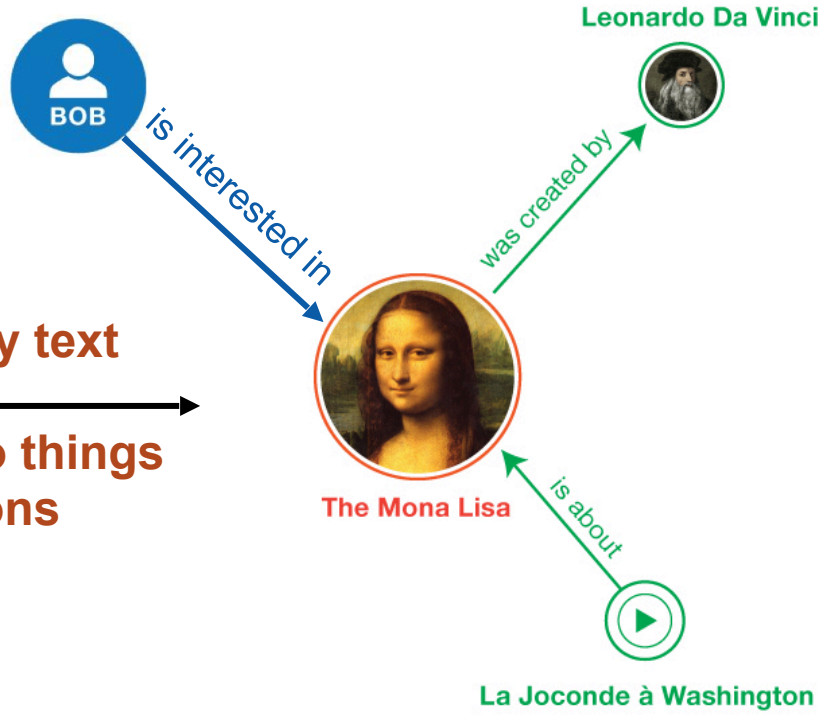


Hypertext links

data in not only text

… but also things and relations

Linked Data

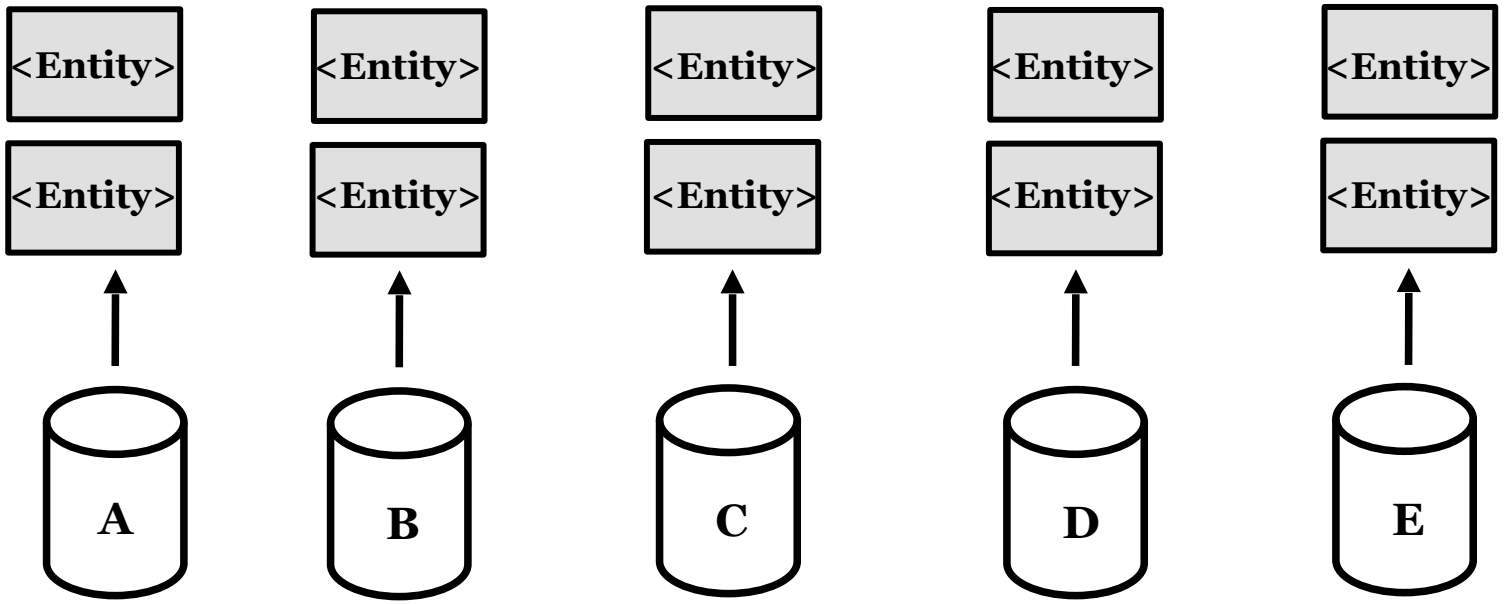


WEB OF DATA CREATION

Use semantic web technologies to:

- Publish structured (**semantic**) data to the web

➡ processable by machines

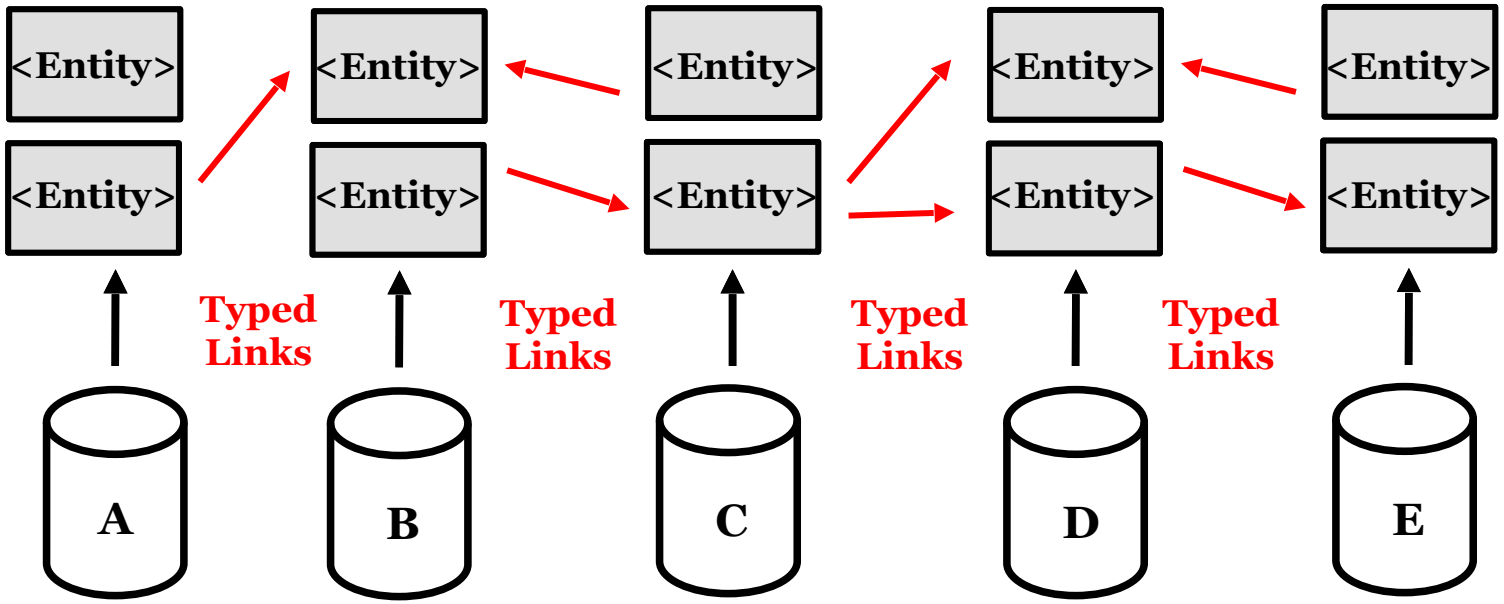


WEB OF DATA CREATION

Use semantic web technologies to:

- Publish structured (**semantic**) data to the web
- Make links between data from one source to data from other sources already published

➡ processable by machines



WEB OF DATA CREATION

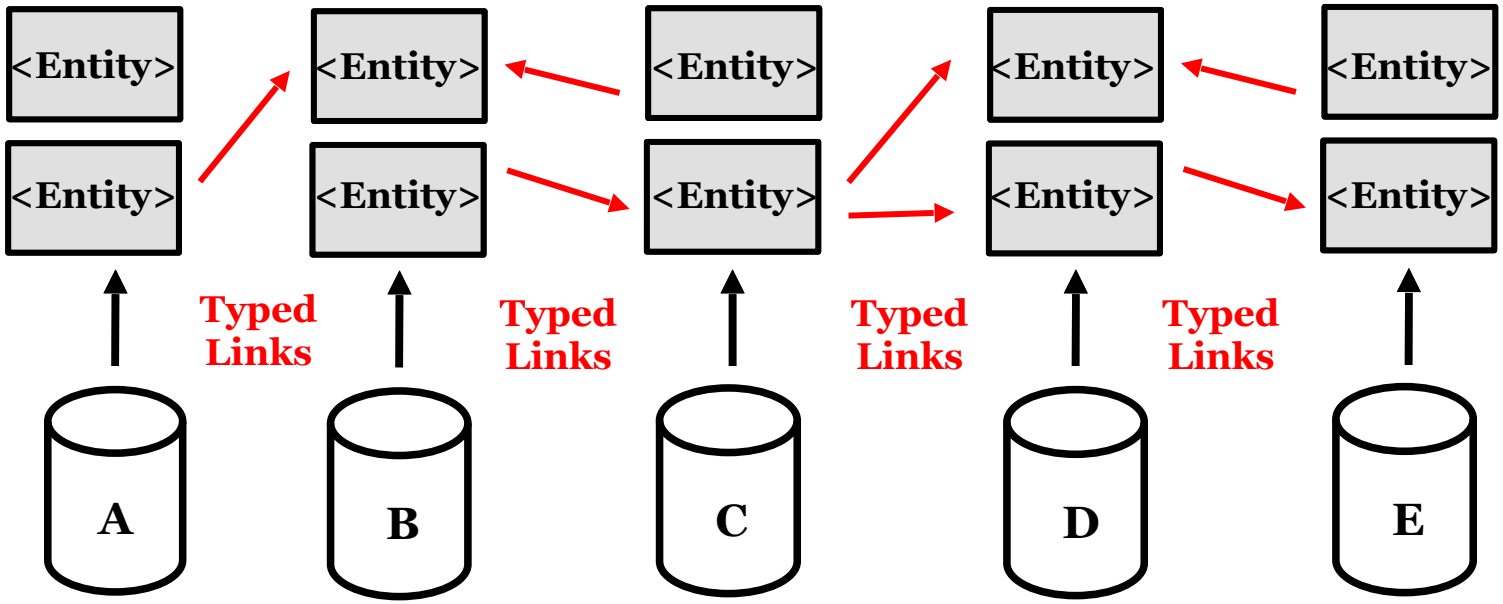
Use semantic web technologies to:

- Publish structured (**semantic**) data to the web
- Make links between data from one source to data from other sources already published

➡ processable by machines

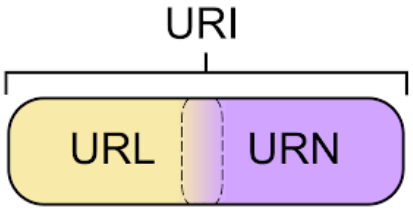
➡ discover more information

➡ unlock the potential of isolated repositories



LINKED DATA PRINCIPLES

"Linked data is a set of design principles for sharing machine-readable data on the Web for use by public administrations, business and citizens."



① **Use HTTP URIs as identifiers for resources**

→ so people can look up the data

② **Provide data at the location of URIs**

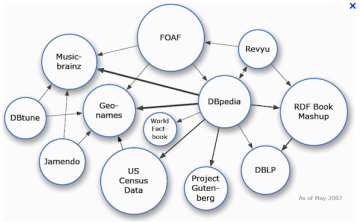
→ to provide data for interested parties

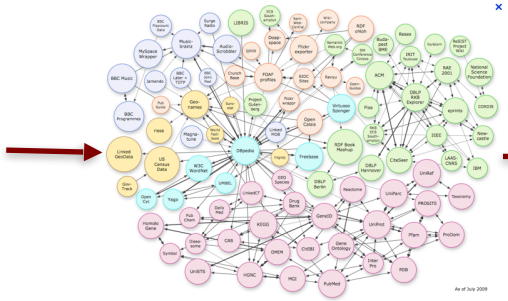
③ **Include links to other resources**

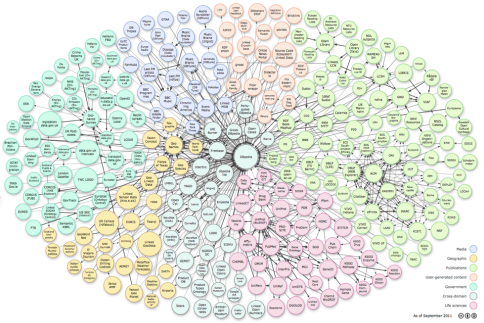
→ so people can discover more information

→ unlock the potential of isolated repositories (islands)

WEB OF DATA GROWTH ...







2009 2011 2014

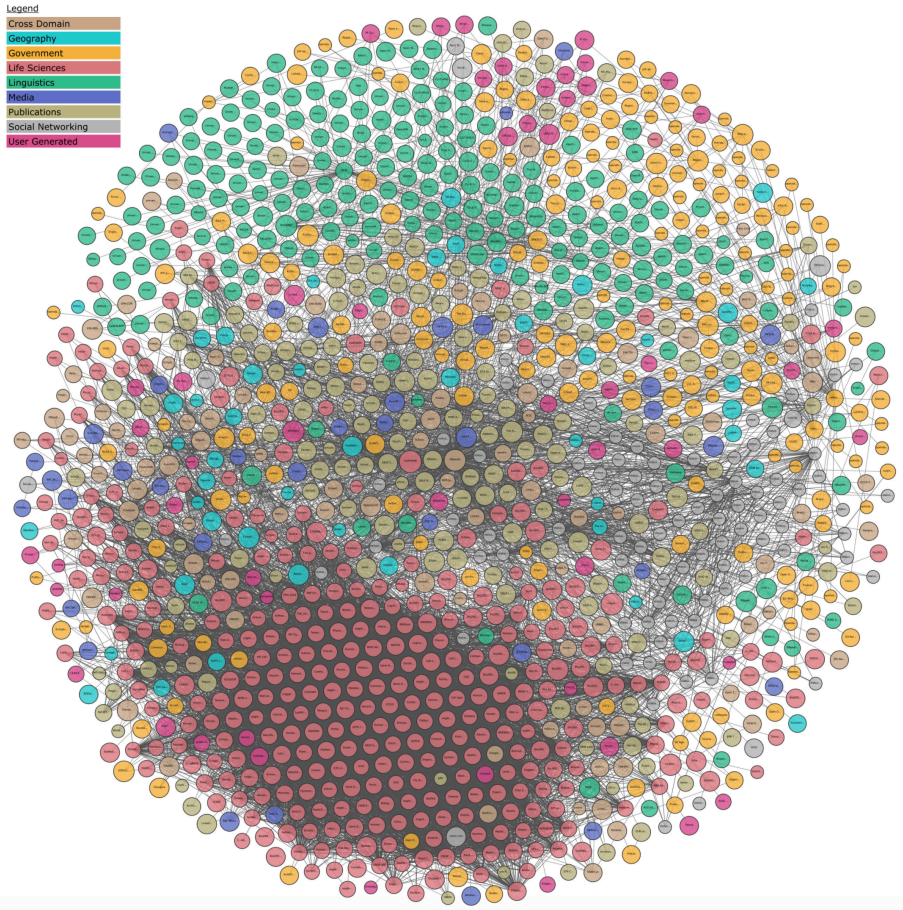
LINKED OPEN DATA

Linked Data - Datasets under an open access

- 650K datasets
- over 28B triples
- over 500M links
- several domains

Ex. DBPedia : 1.5 B triples, 230M entities

Linked Open Data cloud



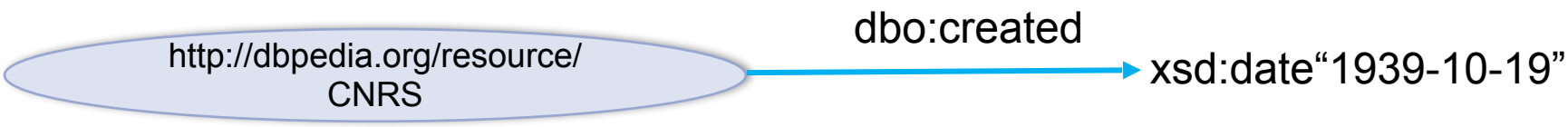
"Linking Open Data cloud diagram 2017, by Andrejs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/"

RDF – RESOURCE DESCRIPTION FRAMEWORK

- **RDF**: a data model for declaring metadata that describe resources on the Web
- **Resources**: Web pages, video or music files, PDF files, Web services, … identified by **URIs** (**Uniform Resource Identifiers**).

RDF – RESOURCE DESCRIPTION FRAMEWORK

- **RDF**: a data model for declaring metadata that describe resources on the Web
- **Resources**: Web pages, video or music files, PDF files, Web services, … identified by **URIs** (**Uniform Resource Identifiers**).
- **Statements of < subject predicate object >**



Subject Predicate Object

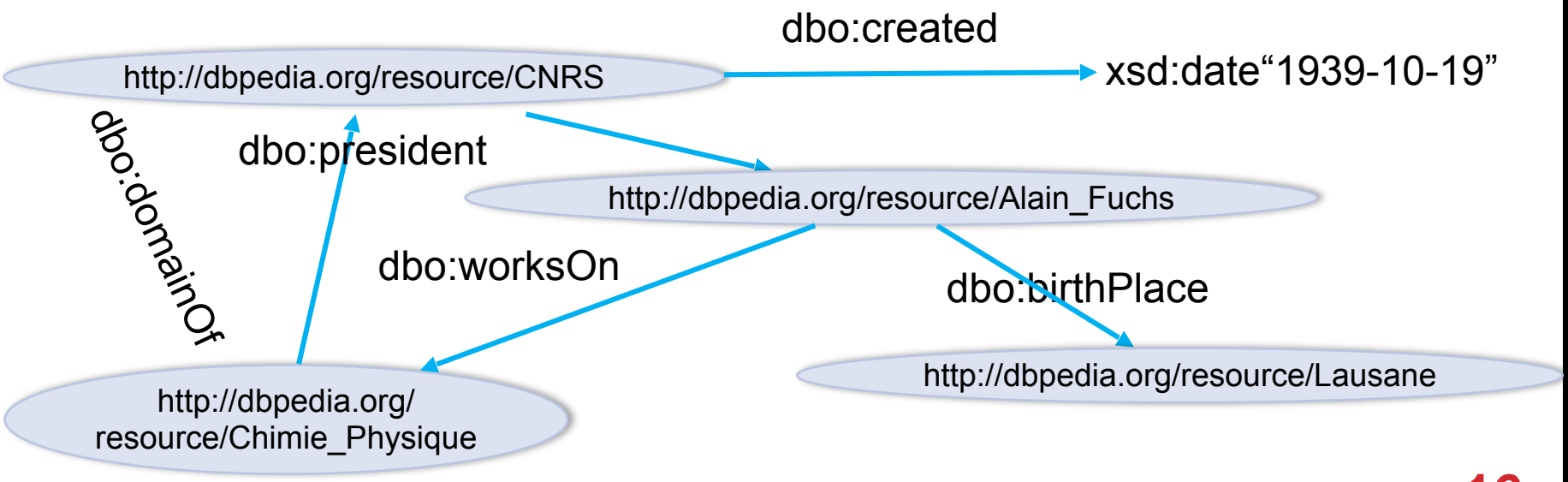
… is called a triple

RDF – RESOURCE DESCRIPTION FRAMEWORK

- **An RDF Graph** is a set of triples.
  - Its **nodes** are (labelled by) the subjects and objects appearing in the triples.
  - Its **edges** are labelled by the properties

RDF – RESOURCE DESCRIPTION FRAMEWORK

- **An RDF Graph** is a set of triples.
  - Its **nodes** are (labelled by) the subjects and objects appearing in the triples.
  - Its **edges** are labelled by the properties



WHO IS DEVELOPING KNOWLEDGE GRAPHS?

2007

2008

yago
select knowledge

2012

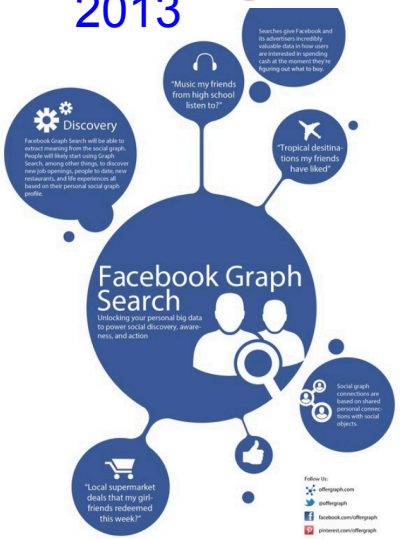
WIKIDATA

2007

Freebase™

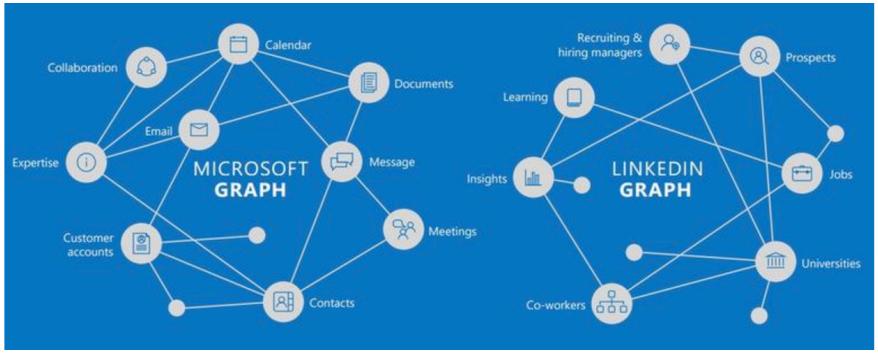
2012

2013



2015

2016



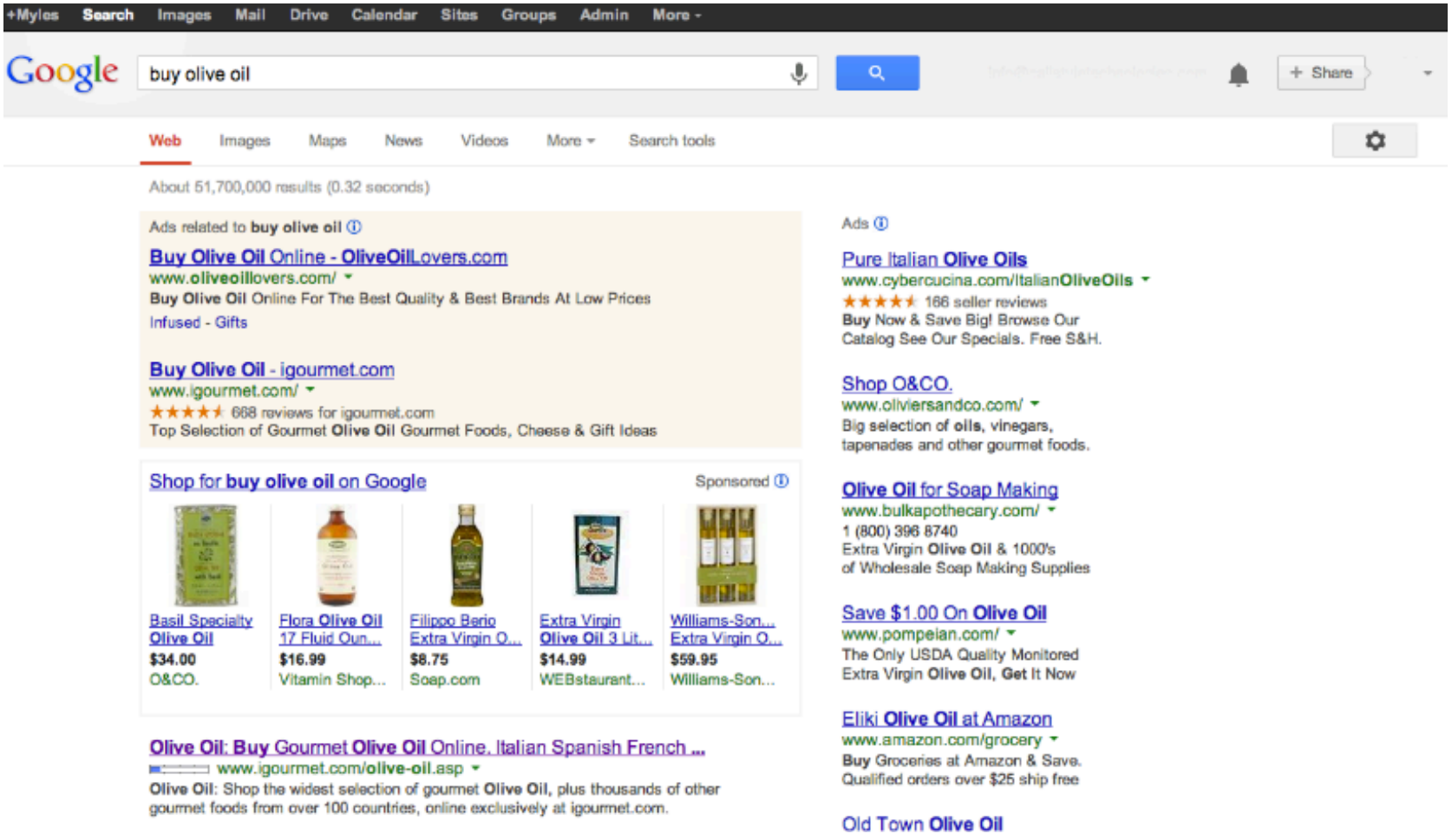
2013

Yahoo's new SERP designs mobile and knowledge graph

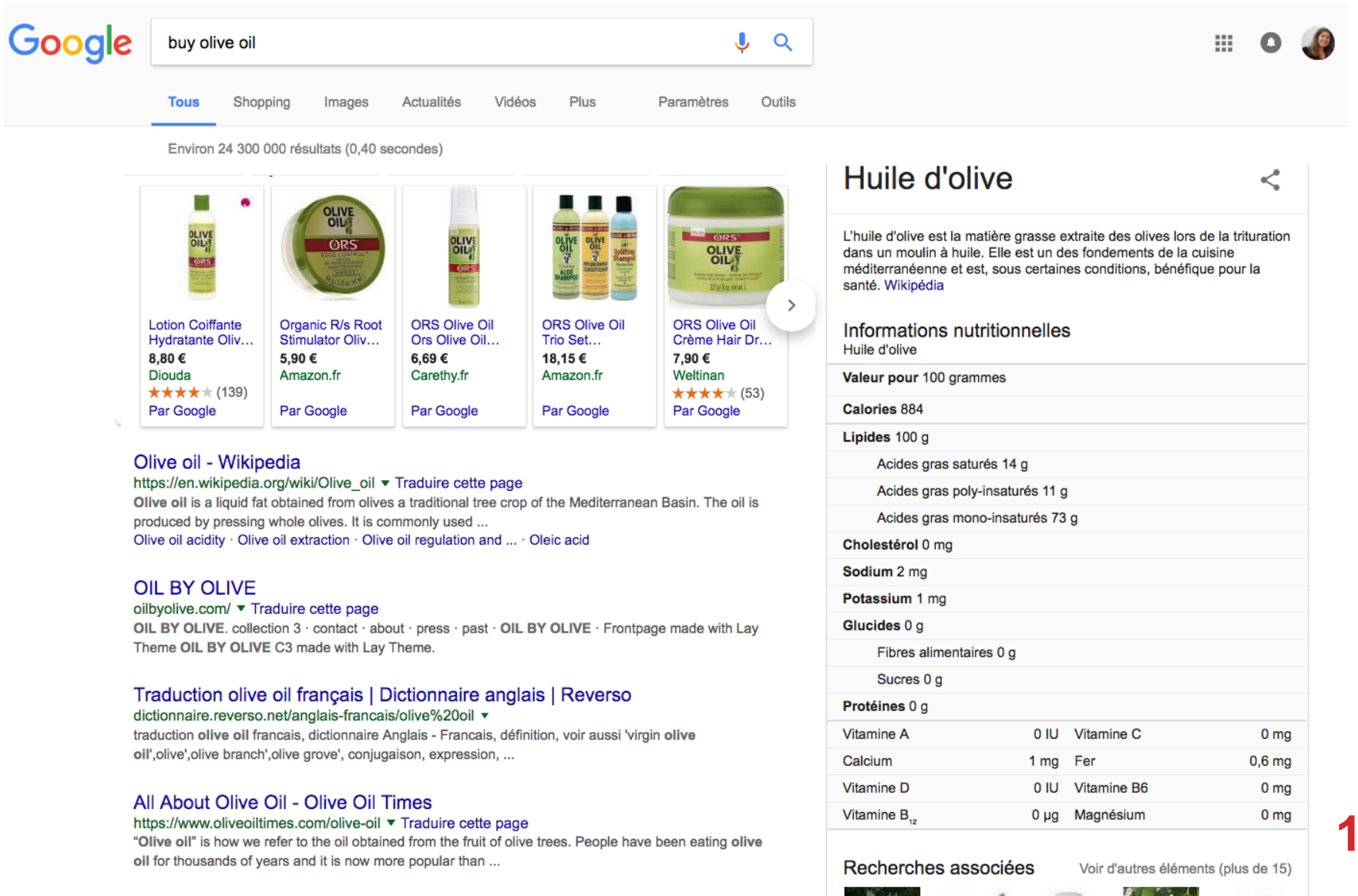
Academic side

Commercial side

WEB SEARCH WITHOUT KNOWLEDGE GRAPHS



WEB SEARCH WITH KNOWLEDGE GRAPHS



QUESTION ANSWERING WITH KNOWLEDGE GRAPHS

barack obama mother

All | Images | Videos | Maps | News | My saves

15 900 000 Results | Date | Language | Region

Barack Obama · Mother

Ann Dunham

Ann Dunham - Wikipedia

https://en.wikipedia.org/wiki/**Ann_Dunham**

Stanley Ann Dunham (November 29, 1942 – November 7, 1995) was an American anthropologist who specialized in the economic anthropology and rural development of ...

Barack Obama Sr · Zarai Taraqiati Bank Limited · Lolo Soetoro · Wikipedia:Good Articles

Family of Barack Obama - Wikipedia

https://en.wikipedia.org/wiki/**Family_of_Barack_Obama**

The family of **Barack Obama**, the 44th President of the United States, and his wife Michelle **Obama** is made up of people of Kenyan (Luo), African-American, and Old Stock ...

United States Citizen · Craig Robinson · Barack Obama Sr · Jonathan Singletary Dunham

Ann Dunham

Anthropologue

Stanley Ann Dunham, née le 29 novembre 1942 à Wichita et morte le 7 novembre 1995 à Honolulu, est une anthropologue américaine spécialisée dans l'anthropologie économique et le développement rural. Elle est la mère de Barack Obama, le 44e … +

Wikipedia

Parents: Madelyn Dunham (Mother) · Stanley Armour Dunham (Father)

Spouse: Lolo Soetoro (m. 1965 - 1980) · Barack Obama, Sr. (m. 1961 - 1964)

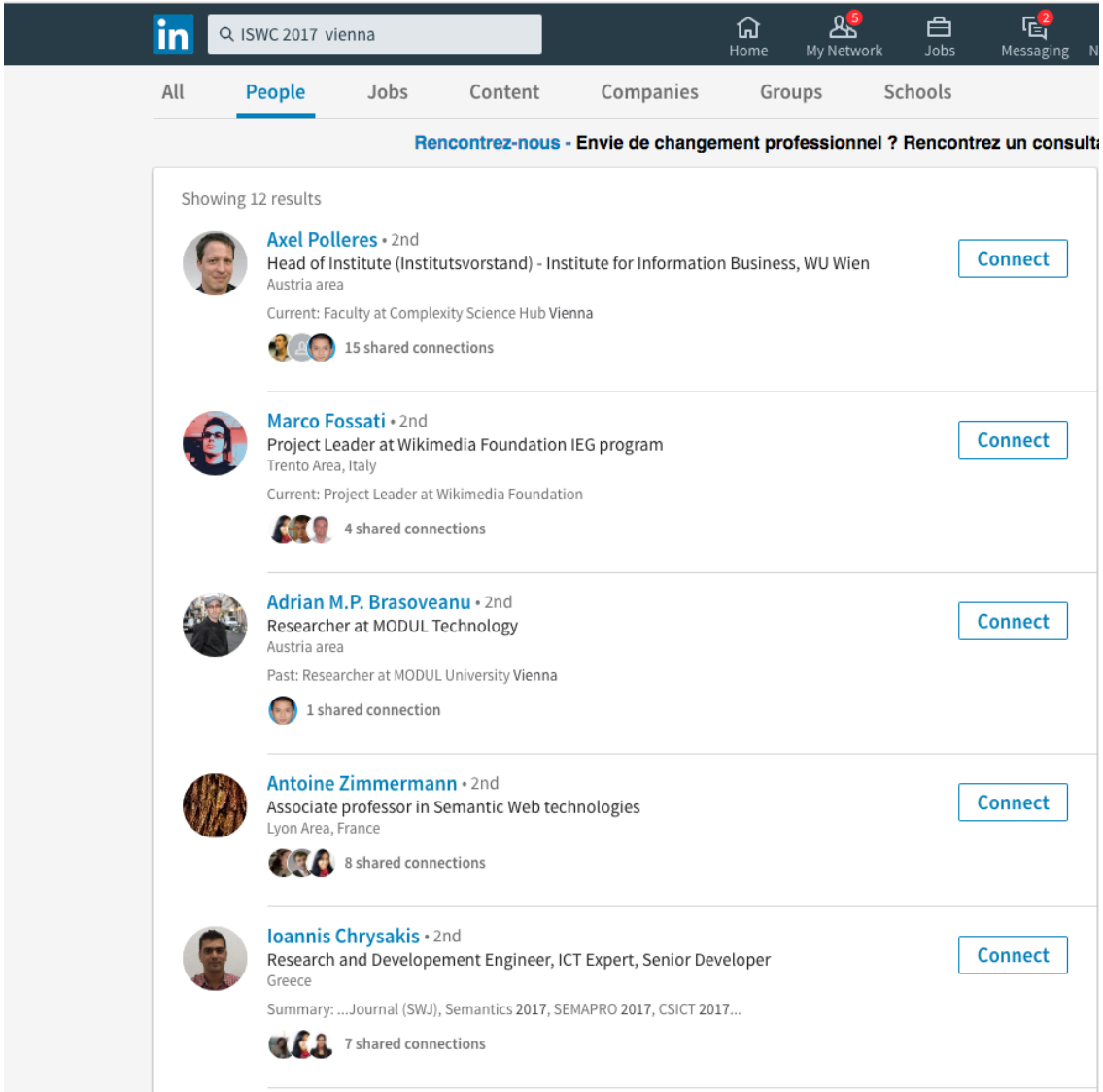
Children: Barack Obama (Son) · Maya Soetoro-Ng (Daughter)

Lived: 29 nov. 1942 - 7 nov. 1995 (age 52)

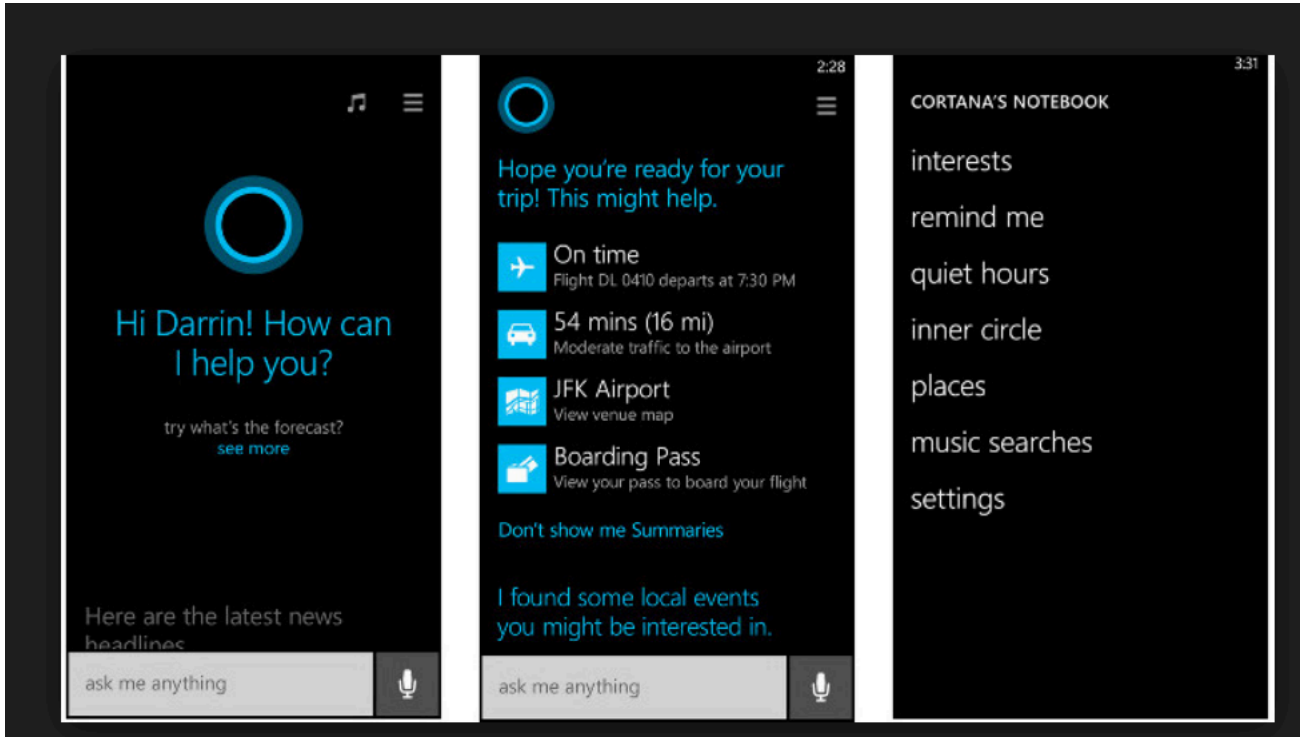
Education: Mercer Island High School · Université d'Hawaï à Mānoa · Université de Washington

Buried: Océan Pacifique

CONNECTING EVENTS AND PEOPLE WITH KNOWLEDGE GRAPHS



TOWARDS A KNOWLEDGE-POWERED DIGITAL ASSISTANT



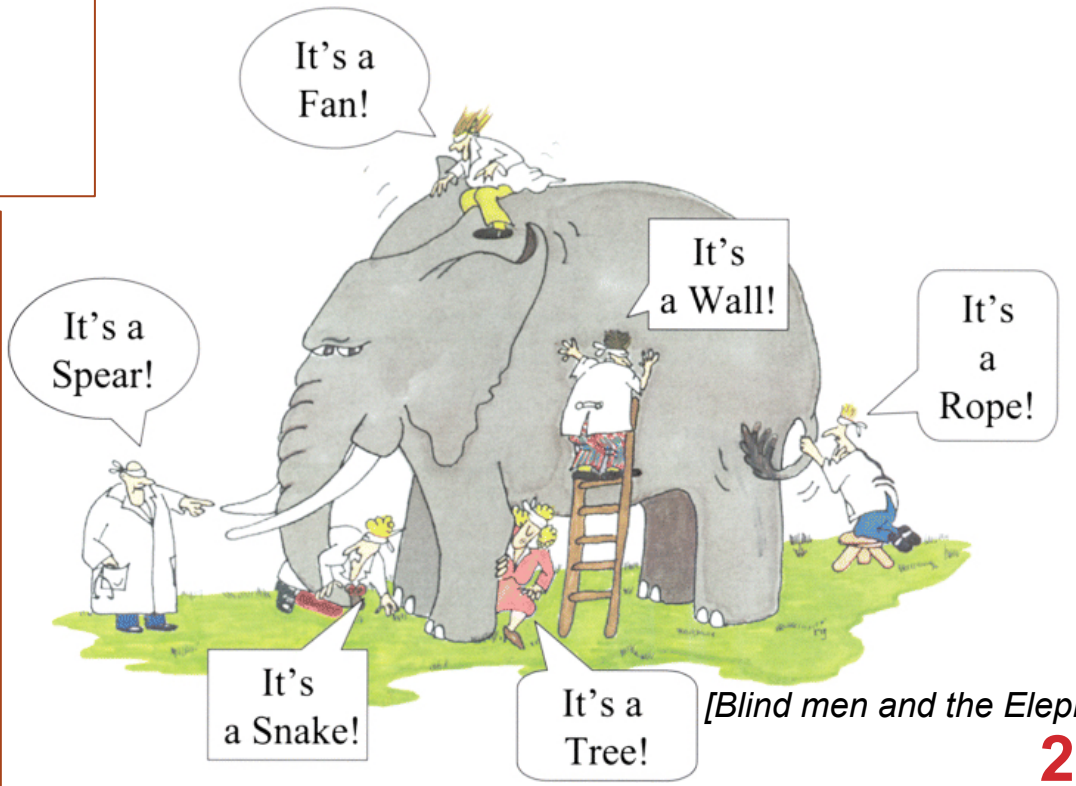
Cortana (Microsoft)

- Natural access and storage of knowledge
- Chat bots
- Personalization
- Emotion

KNOWLEDGE GRAPH ADOPTION [2019]

source: https://fr.slideshare.net/Frank.van.Harmelen/adoption-of-knowledge-graphs-mid-2019

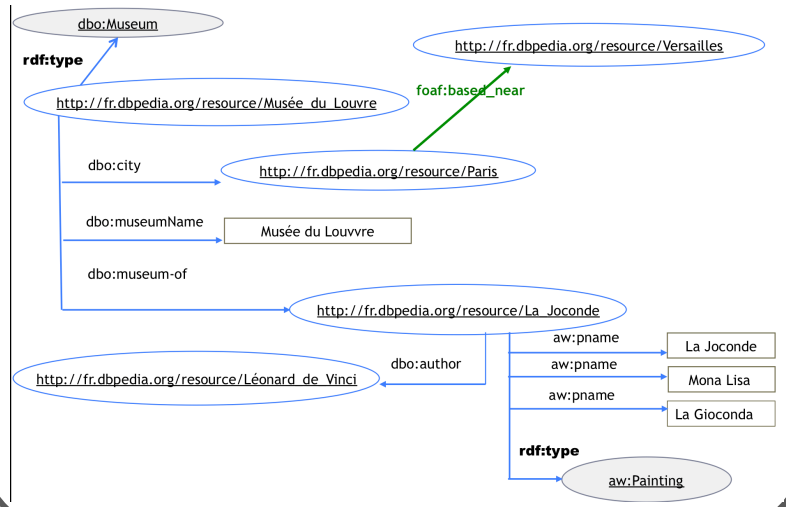
KNOWLEDGE GRAPH: A DEFINITION …



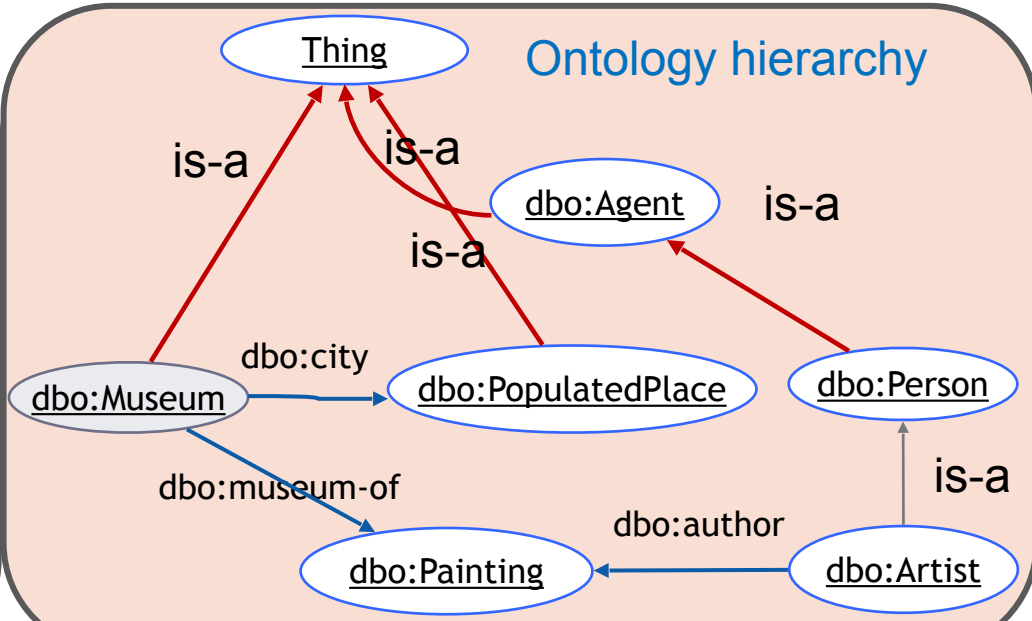
[Blind men and the Elephant]

KNOWLEDGE GRAPH (KG)

RDF Graphs



Ontology hierarchy



Querying (SPARQL)

```
PREFIX dbo: <http://dbpedia.org/ontology#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
SELECT ?m ?p
WHERE { ?m rdf:type dbo:Museum . ?m dbo:musuem-of ?p .}
```

Reasoners: (Pellet, Fact++, Hermit, etc.)

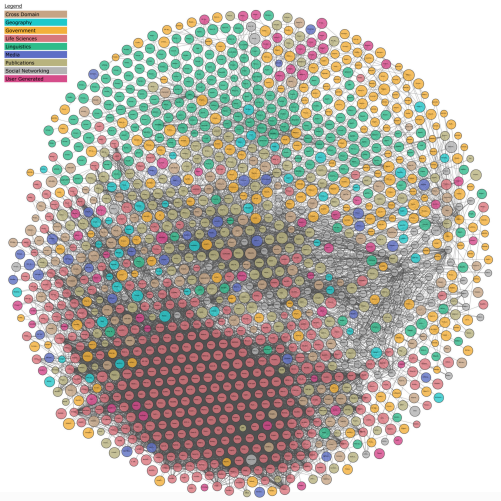
- KG saturation: infer whatever can be inferred from the KG.
- KG consistency checking: no contradictions
- KG repairing
- …

Ontology axioms and rules

owl:equivalentClass(dbo:Municipality, dbo:Place)
owl:equivalentClass(dbo:Place, dbo:Wikidata:Q532)
owl:equivalentClass(dbo:Village, dbo:PopulatedPlace)
owl:equivalentClass(dbo:PopulatedPlace,dbo:Municipality)
owl:disjointClass(dbo:PopulatedPlace, dbo:Artist)
owl:disjointClass(dbo:PopulatedPlace, dbo:Painting)
owl:FunctionalProperty(dbo:city)
owl:InverseFunctionalProperty(dbo:museum-of)

dbo:birthPlace(X, Y) => dbo:citizsenOf(X, Y)
dbo:parentOf(X, Y) => dbo:child(Y, X)

HOW TO PUBLISH LINKED DATA



5 STAR-SCHEMA OF LINKED (OPEN) DATA

T. Berners Lee 2006

★ Make your stuff available on the Web (whatever format) under an open license. Optional

★★ Make it available as structured data (e.g., Excel instead of image scan of a table)

★★★ Use non-proprietary formats (e.g., CSV instead of Excel)

★★★★ Use URIs to denote things, so that people can point at your stuff

★★★★★ Link your data to other data to provide context

5 STAR-SCHEMA OF LINKED (OPEN) DATA

data.gouv.fr

[data.gouv.fr](), more than 40370 dataset publicly available

Visualisation
4 jeux de données
0

Rapport financier 2020

Application
4 jeux de données
Open Mobility Indicators

Open Mobility Indicators

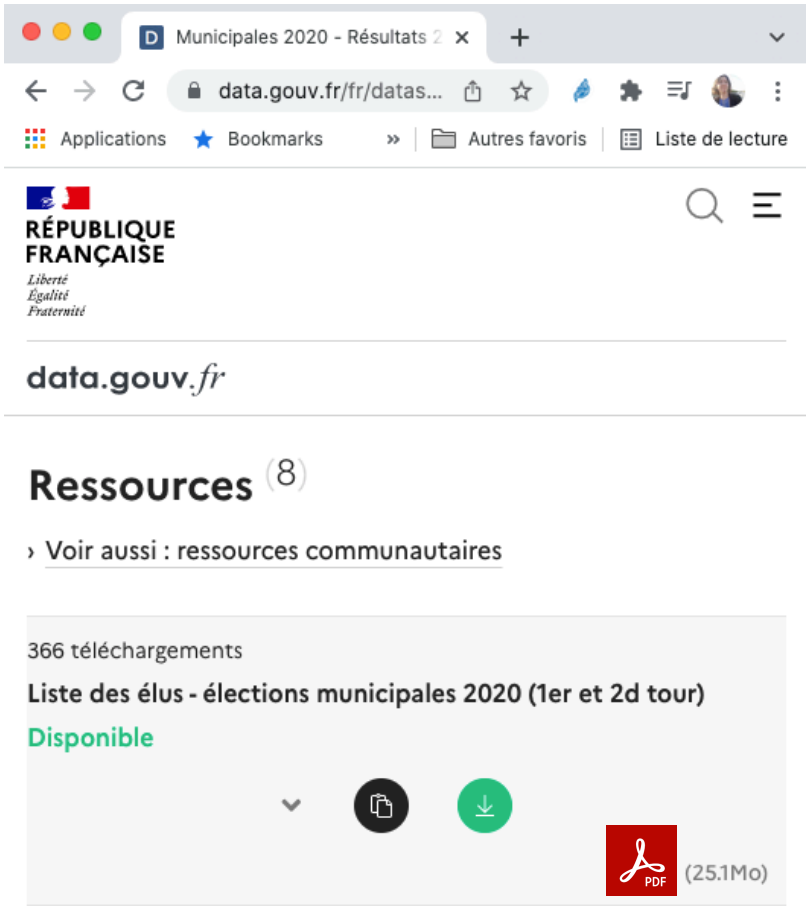
Visualisation
4 jeux de données
0

Aménagements cyclables

...

5 STAR-SCHEMA OF LINKED (OPEN) DATA

★ Make your stuff available on the Web (whatever format) under an open license.



5 STAR-SCHEMA OF LINKED (OPEN) DATA

★ Make your stuff available on the Web (whatever format) under an open license.

As a consumer …	As a publisher
You can look at it	It is simple to publish
You can store it locally	You do not have to explain to others how they can use your data
You can enter the data into any other system	
You can change the data	
You can share the data with anyone	

5 STAR-SCHEMA OF LINKED (OPEN) DATA

data.gouv.fr

★★ Make it available as structured data (e.g., Excel instead of image scan of a table)

	Année	Organisme	Code ancienne région	Ancienne région
1	2020	Urssaf Corse	94	Corse
2	2020	Urssaf Aquitaine	72	Aquitaine
3	2020	Urssaf Bourgogne	26	Bourgogne
4	2020	Urssaf Centre	24	Centre
5	2020	CGSS Guadeloupe	01	Guadeloupe
6	2020	Urssaf Haute-Normandie	23	Haute-Normandie
7	2020	Urssaf Languedoc-Roussillon	91	Languedoc-Roussillon
8	2020	Urssaf Limousin	74	Limousin
9	2020	Urssaf Poitou-Charentes	54	Poitou-Charentes
10	2020	Urssaf Rhône-Alpes	82	Rhône-Alpes
11	2020	Urssaf Auvergne	83	Auvergne
12	2020	Urssaf Aquitaine	72	Aquitaine
13	2020	CGSS Guyane	03	Guyane

5 STAR-SCHEMA OF LINKED (OPEN) DATA

★★ Make it available as structured data (e.g., Excel instead of image scan of a table)

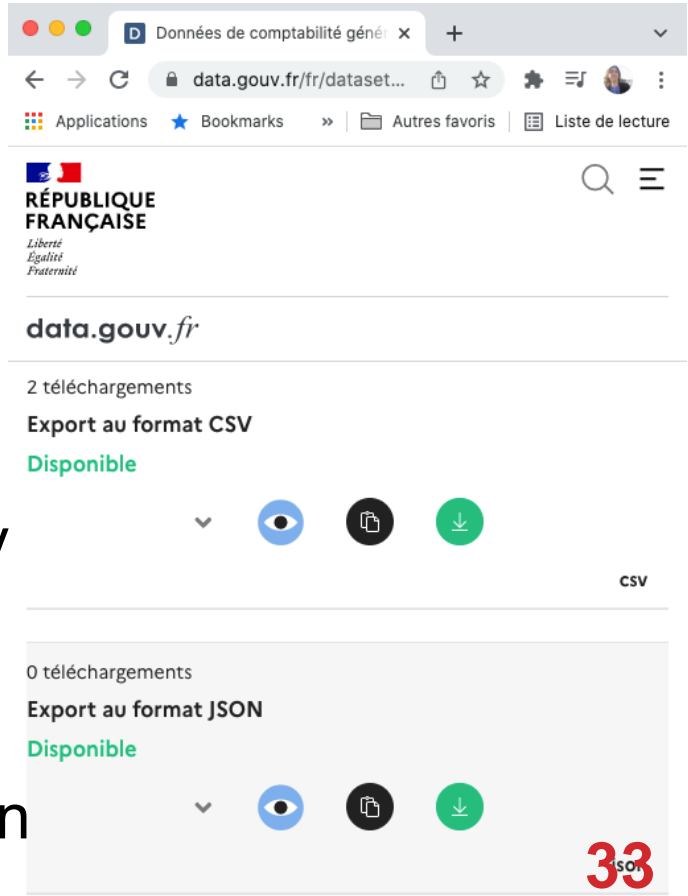
All the benefits of ★ open data; **plus**

As a consumer …	As a publisher
You can directly process it with proprietary software to aggregate it, perform calculations, visualise it, etc.	
You can export it into another (structured) format.	

5 STAR-SCHEMA OF LINKED (OPEN) DATA

★★★ Use non-proprietary formats (e.g., CSV instead of Excel)

Proprietary: Excel, Word, PDF…
Non-proprietary: XML, CSV, RDF, JSON, …



5 STAR-SCHEMA OF LINKED (OPEN) DATA

★★★ Use non-proprietary formats (e.g., CSV instead of Excel)

All the benefits of ★ ★ open data; **plus**

As a consumer …	As a publisher
You can manipulate the data in any way you like, without being confined by the capabilities of any particular software.	It is still simple to publish.
	But, you do need converters or plug-ins to export the data from the proprietary format.

5 STAR-SCHEMA OF LINKED (OPEN) DATA

★★★★ Use URIs to denote things, so that people can point at your stuff

For example, creating a URI for French government

https://dbpedia.org/page/Government_of_France

5 STAR-SCHEMA OF LINKED (OPEN) DATA

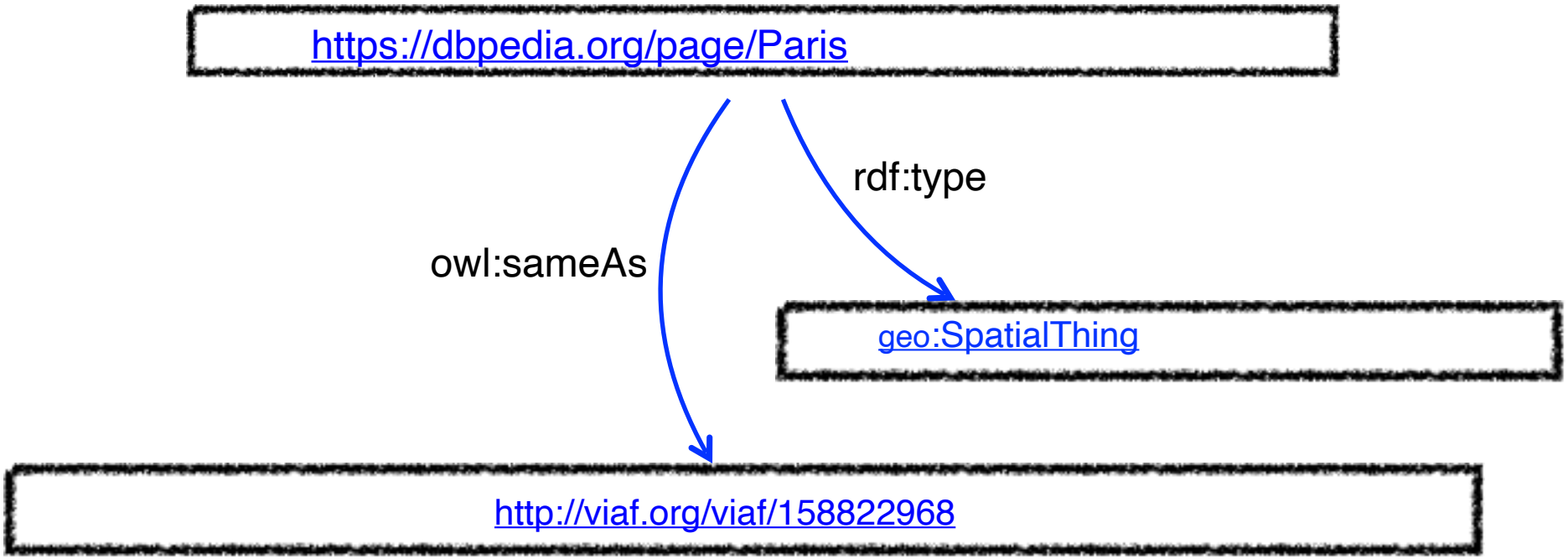
★★★★ Use URIs to denote things, so that people can point at your stuff

All the benefits of ★ ★ ★ open data; **plus**

As a consumer …	As a publisher
You can link to it from any other place	You have fine-granular control over the data items and can optimise their access
You can bookmark it	Other data publishers can now link into your data, promoting it to 5 star.
You can reuse parts of the data	You will be able to reuse vocabularies, data and metadata, and URI design
You may be able to reuse existing tools and libraries.	
You can combine the data safely with other data.	- But you typically need to invest some time in slicing and dicing your data.

5 STAR-SCHEMA OF LINKED (OPEN) DATA

★★★★★ Link your data to other data to provide context



5 STAR-SCHEMA OF LINKED (OPEN) DATA

★★★★★ Link your data to other data to provide context

All the benefits of ★★★★ open data; **plus**

As a consumer …	As a publisher
You can discover more (related) data while consuming the data	You make your data discoverable
You can directly learn about the data schema.	You increase the context, expressivity, quality, value and visibility of your data
You can combine data from different sources, be innovative, gain new knowledge,...	- This requires an investment in time, money, technology and skills.

BENEFITS OF USING LINKED DATA

- Allows assigning **identifiers** in the form of HTTP URIs to things (e.g. people, products, business, locations…)
- The use of structured formats and standard Web interfaces (such as HTTP and SPARQL) **can simplify the use of data for machines.**

- Provides **context** to data – richer and more expressive data

- Allows for flexible **integration of datasets** from different sources, without needing the data to be moved.
- Fosters the **reuse** of information (**data and knowledge**) from reference/authoritative sources.

NEED OF KNOWLEDGE

THE ROLE OF KNOWLEDGE IN AI

[Artificial Intelligence 47 (1991)]

ON THE THRESHOLDS OF KNOWLEDGE

Douglas B. Lenat

MCC
3500 W. Balcones Center
Austin, TX 78759

Edward A. Feigenbaum

Computer Science Department
Stanford University
Stanford, CA 94305

Abstract

We articulate the three major fmdings of AI to date: (1) The Knowledge Principle: if a program is to perform a complex task well, it must know a great deal about the world in which it operates. (2) A plausible extension of that principle, called the Breadth Hypothesis: there are two additional abilities necessary for intelligent behavior in unexpected situations: falling back on increasingly general knowledge, and analogizing to specific but far-flung knowledge. (3) AI as Empirical Inquiry: we must test our ideas experimentally, on large problems. Each of these three hypotheses proposes a particular threshold to cross, which leads to a qualitative change in emergent intelligence. Together, they determine a direction for future AI research.

opponent is Castling.) Even in the case of having to search

*The knowledge principle: "if a program is to perform a complex task well, **it must know a great deal about the world** in which it operates."*

The Well-Formedness Threshold: For each task, there is some minimum knowledge needed for one to even formulate it.

ONTOLOGY, A DEFINITION

“An ontology is an **explicit**, **formal specification** of a **shared conceptualization.”**

[Thomas R. Gruber, 1993]

Conceptualization: abstract model of domain related expressions

Specification: domain related

Explicit: semantics of all expressions is clear

Formal: machine-readable

Shared: consensus (different people have different perceptions)

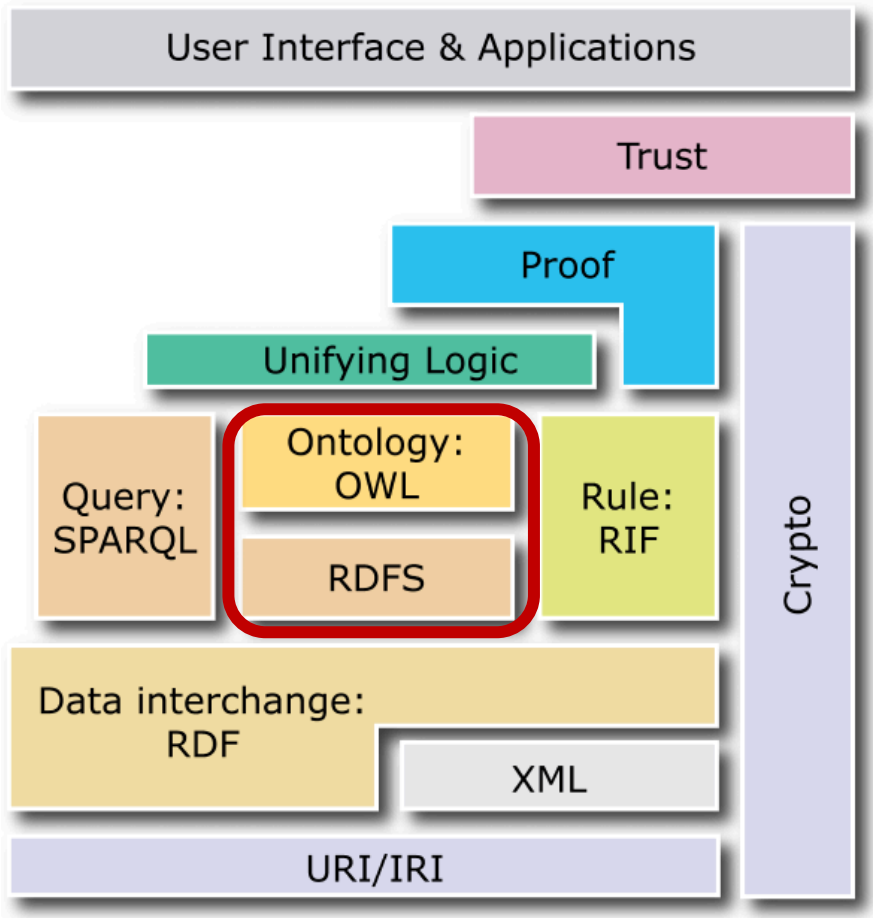
SEMANTIC WEB: ONTOLOGIES

RDFS – Resource Description Framework Schema

- Lightweight ontologies

OWL – Web Ontology Language

- Expressive ontologies

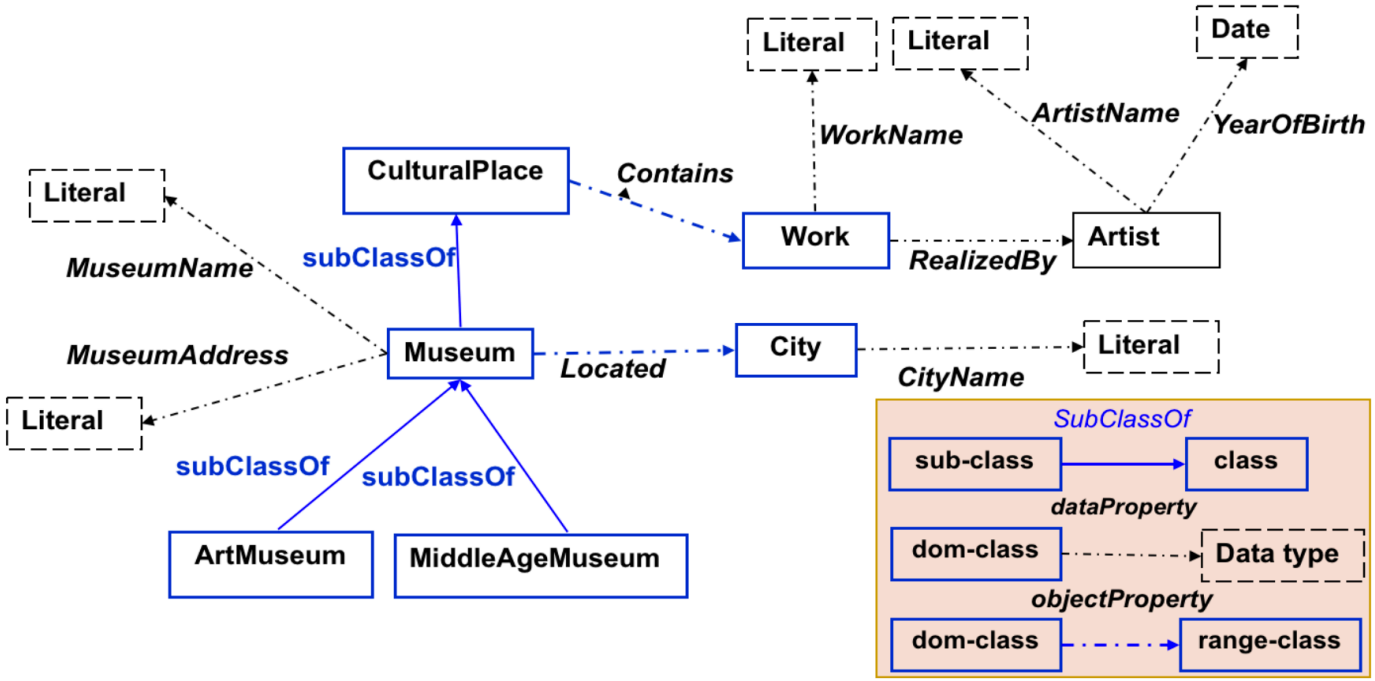


Source: https://it.wikipedia.org/wiki/File:W3C-Semantic_Web_layerCake.png

OWL – WEB ONTOLOGY LANGUAGE

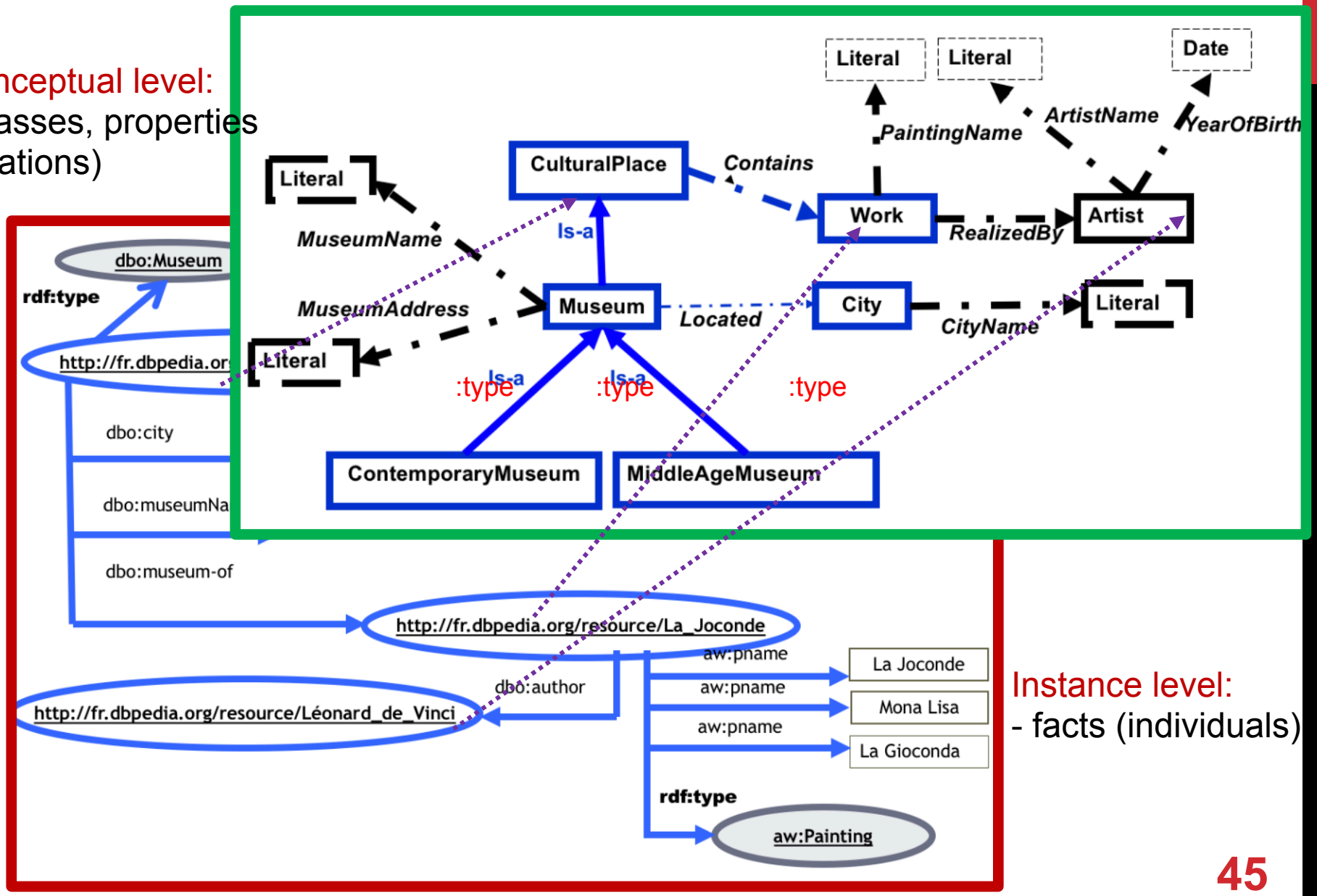
- **Classes:** concepts or collections of objects (individuals)
- **Properties**:
  - owl:DataTypeProperty (attribute)
  - owl:ObjectProperty (relation)
- **Individuals**: ground-level of the ontology (instances)

- **Axioms**
  - owl:subClassOf
  - owl:subPropertyOf
  - owl:inverseProperty
  - owl:FunctionalProperty
  - owl:minCardinality
  - …



ONTOLOGY LEVELS: KNOWLEDGE ENGINEERING VIEW

Conceptual level:
- classes, properties (relations)



Instance level:
- facts (individuals)

OWL ONTOLOGY - REASONING

- **Axioms:** knowledge definitions in the ontology that were **explicitly defined** and have **not been proven true**.
  - **Reasoning over an ontology**

    → Implicit knowledge can be made explicit by logical reasoning

- **Example:**

  Pompidou museum is an **Art Museum**

  < Pompidou_museum rdf:type ArtMuseum**> .**

  Pompidou museum contains ***Hallucination partielle***

  < Pompidou_museum ao:contains ***Hallucination_partielle*****> .**

OWL ONTOLOGY - REASONING

- **Axioms:** knowledge definitions in the ontology that were **explicitly defined** and have **not been proven true**.
  - **Reasoning over an ontology**
    → Implicit knowledge can be made explicit by logical reasoning

- **Example:**

  Pompidou museum is an **Art Museum**

  < Pompidou_museum rdf:type ArtMuseum**> .**

  Pompidou museum contains ***Hallucination partielle***

  < Pompidou_museum ao:contains ***Hallucination_partielle*****> .**

- **Infer that:**

➔ Pompidou museum is a CulturalPlace

< Pompidou_museum rdf:type CulturalPlace> .

Because: **Museum** subsumes **ArtMuseum** and **CulturalPlace** subsumes **Museum**

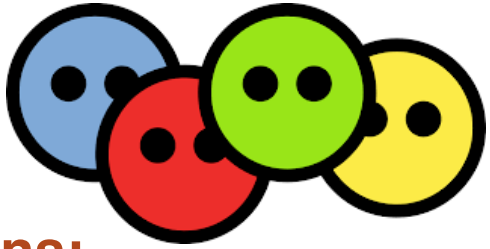
➔ ***Hallucination partielle*** is a Work

<*Hallucination_partielle* rdf:type ao:Work> .

Because: the **range** of the object property **contains** is the class **Work**.

LINKED DATA ONTOLOGIES

FOAF Ontology



Description of persons:

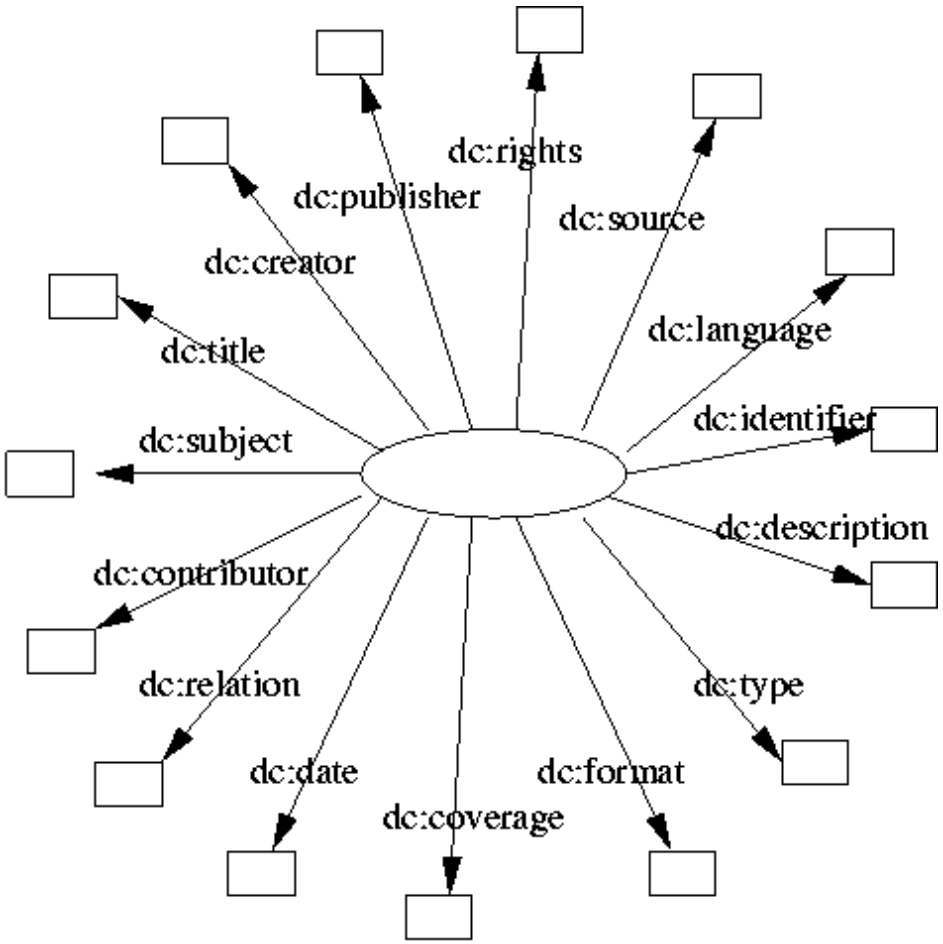
<mariejean> <foaf:knows> <ineskhan>
<ineskhan> <foaf:interest> “classic music ‘’
<ineskhan> <foaf:knows> <billybob>

A machine could infer that Marie Jean might like to know Billy Bob.

LINKED DATA ONTOLOGIES

Dublin core ontology

Description of document metadata



LINKED DATA ONTOLOGIES

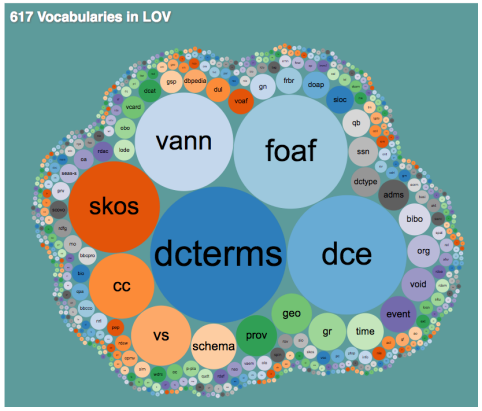
schema.org

- Common vocabularies that search engines can understand
- Micro-data data on the web in their HTML

Goals

- **Create a web for both humans and machines**
- Help webmasters to make metadata available through web standards and structured HTML
- Gain access to the meaning of web sites
- Establish relationships between data that allow for exploration and discovery

AVAILABLE ONTOLOGIES/ VOCABULARIES



Linked Open Vocabularies

- Keeps track of available open ontologies and provides them as a graph
- Search for available ontologies, open for reuse
- Example:

http://lov.okfn.org/dataset/lov/vocabs/foaf

BioPortal

https://bioportal.bioontology.org/

AgroPortal

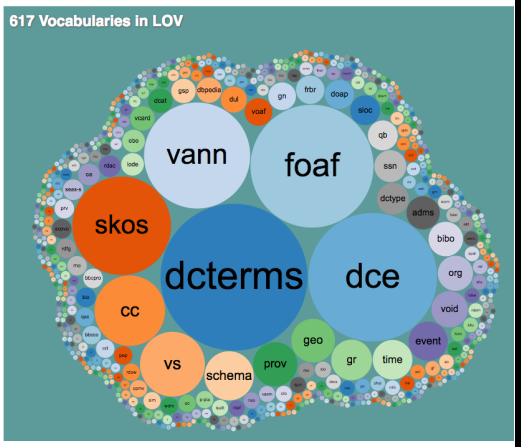
http://agroportal.lirmm.fr/

https://lov4iot.appspot.com/

Repositories of domain specific ontologies

LINKED DATA ONTOLOGIES

- Only a small portion of use-cases are covered by the well-known vocabularies (**foaf, dc, time, schema, …**)
- But, for the sake of **interoperability**, it is not judicious to create new vocabularies each time you create a new application
- Exploit resources that allow to s**earch and browse** public vocabularies



https://lov.linkeddata.es/

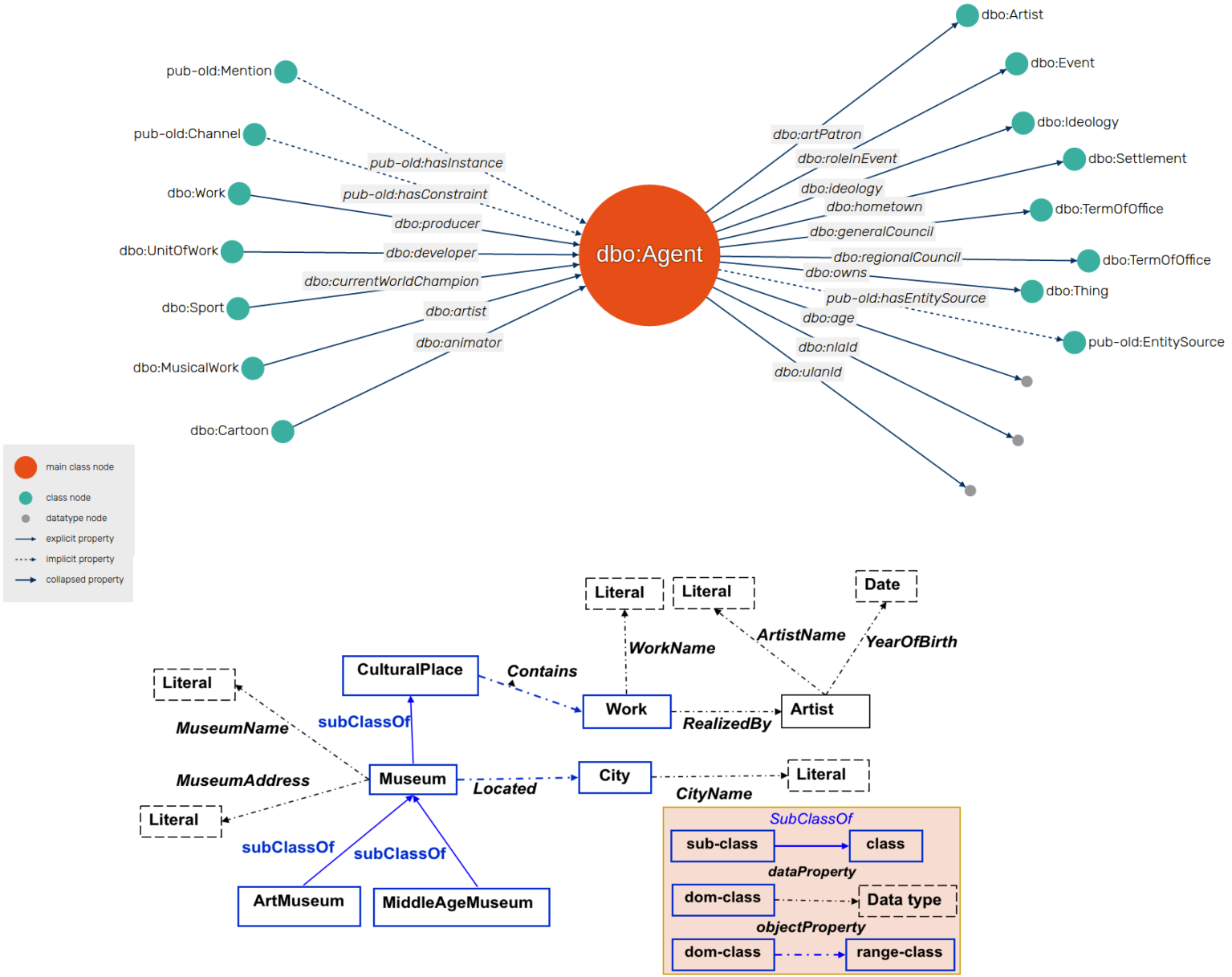
Still, coverage problem may happen
==> Import + mapping to existing ontologies.

Ontology Alignment

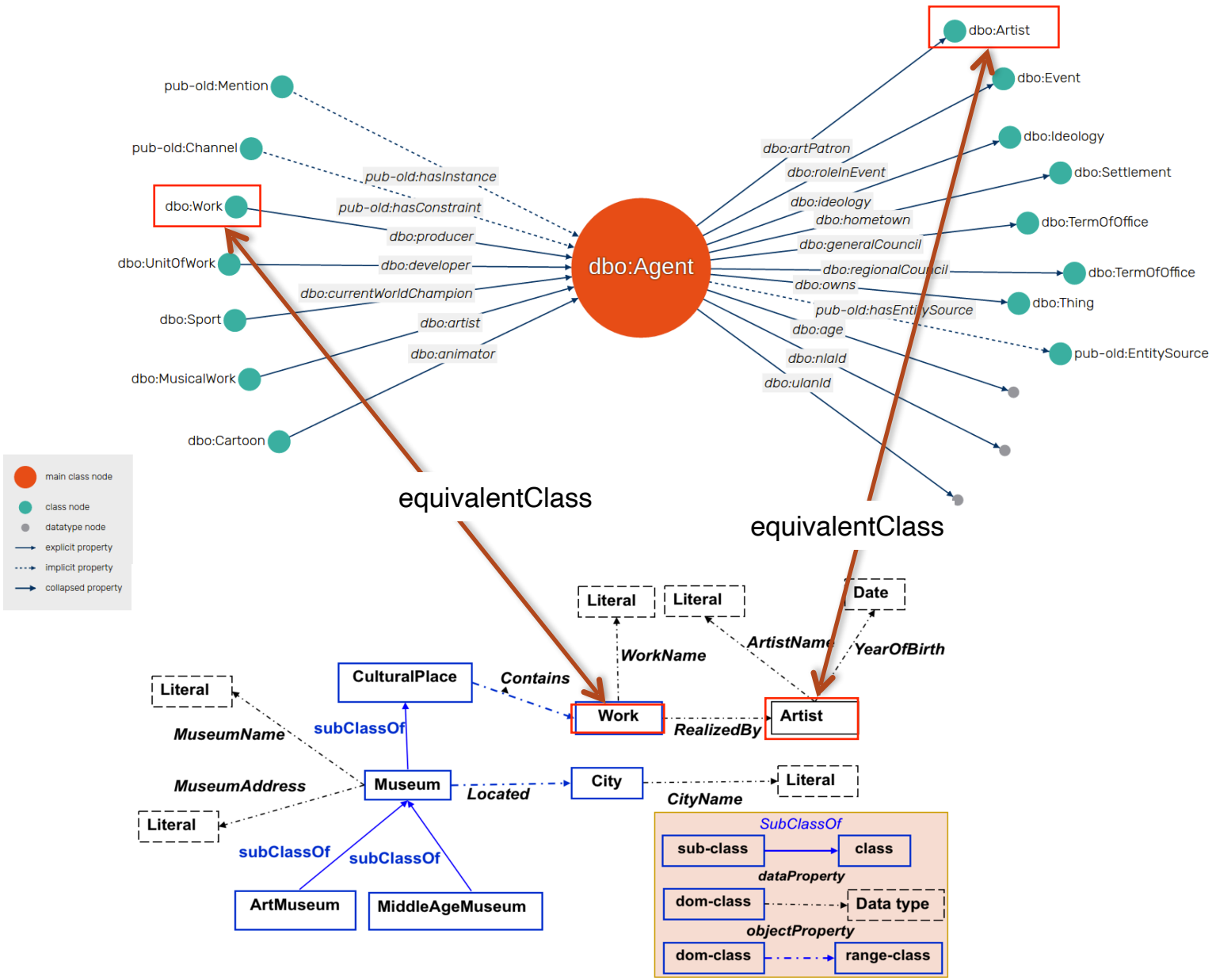
COURSE OUTLINE

- Introduction to linked data principles
- Linked data publication
- **Ontology alignment**
- Data Linking
- Dataset interoperability and validation
- Link invalidation

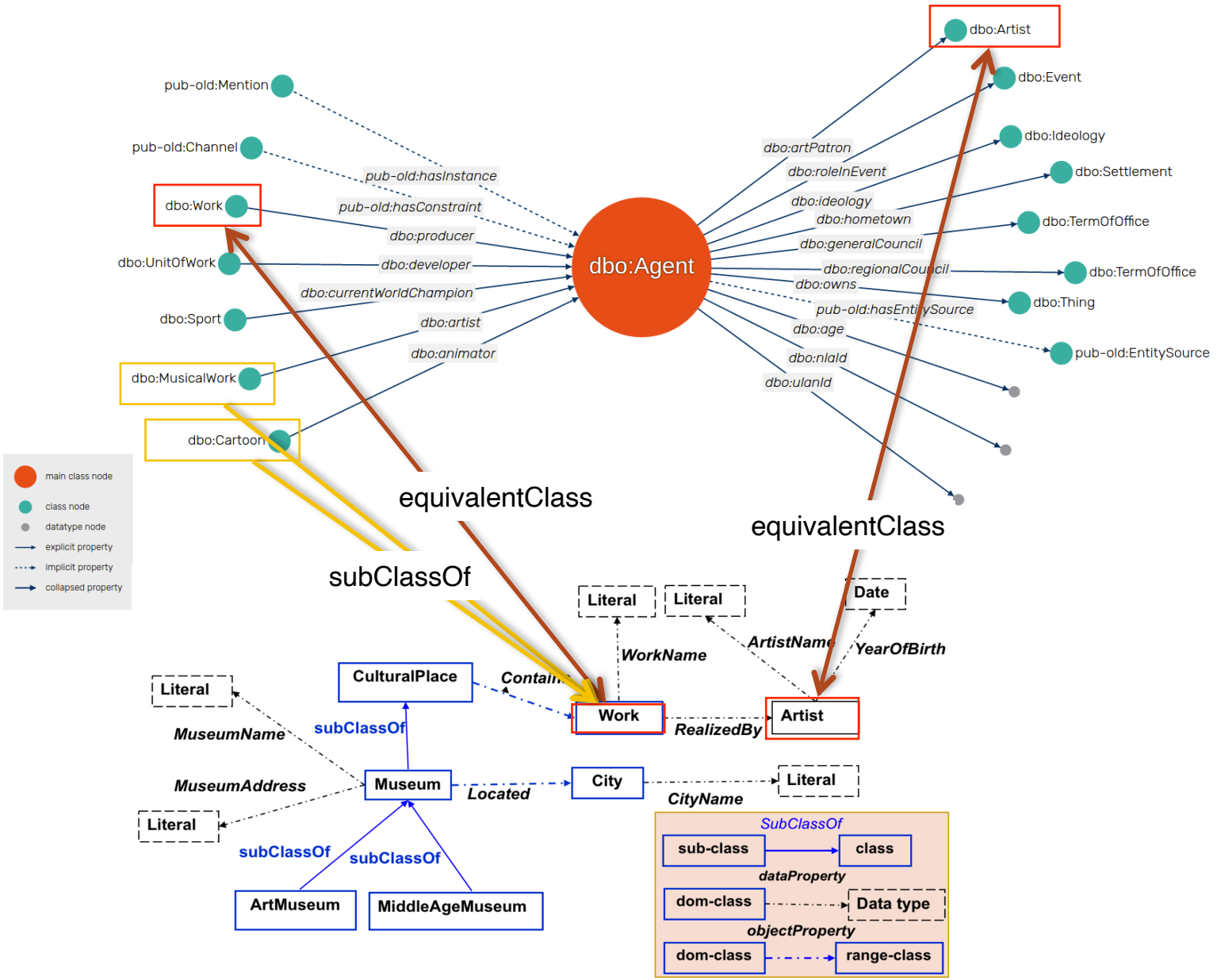
A MUSEUM ONTOLOGY



A MUSEUM ONTOLOGY



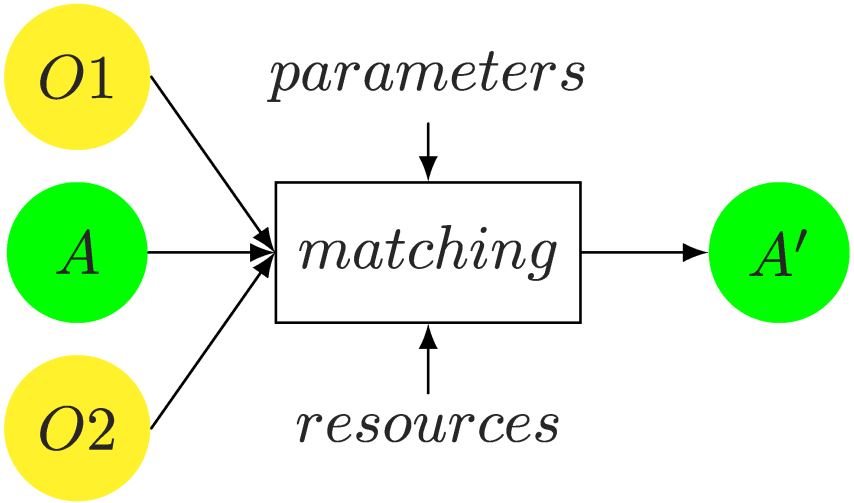
A MUSEUM ONTOLOGY



ONTOLOGY ALIGNMENT

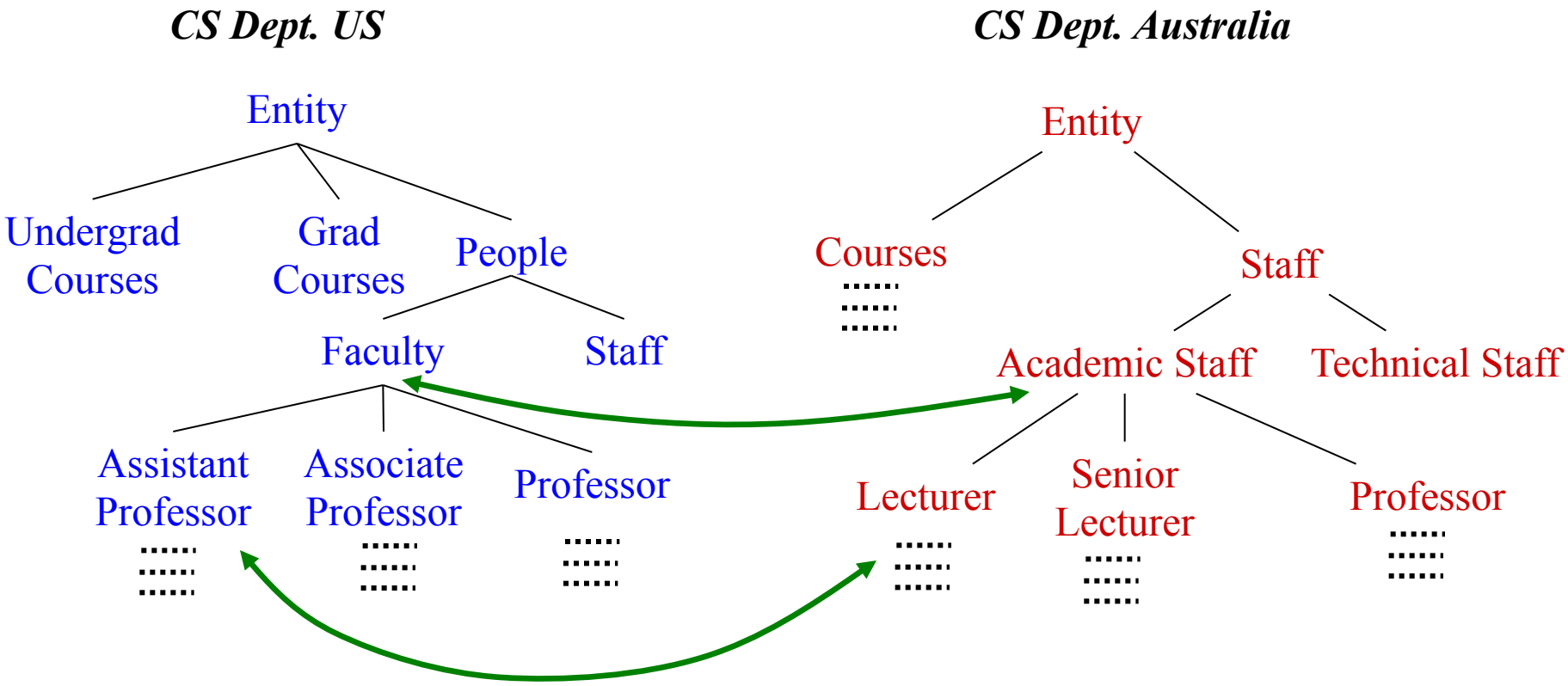
Ontology alignement is a matching process that:

Establishes an **alignment (A')** from two ontologies (O1 and O2) and optionally an input **alignment (A),** parameters and external resources, like dictionaries, … [EUZENAT and SHVAIKO 2013]).



A' may belong to as set **Θ** of possible **alignment relations** (e.g. owl:disjointWith, owl:equivalentClass, rdfs:subclassOf, closeTo, …)

ALIGNMENT OF A SIMPLE TAXONOMY/HIERARCHY (Θ=EQUIVALENTCLASS)



PROBLEM OF HETEROGENEITY

WHAT KIND OF HETEROGENEITY?

Syntaxic: two ontologies are not described using the same language (OWL/RDFS)

- ➔ Equivalence between language constructs, transformations, abstractions are needed.

WHAT KIND OF HETEROGENEITY?

Syntaxic: two ontologies are not described using the same language (OWL/RDFS)

➔ Equivalence between language constructs, transformations, abstractions are needed.

Terminological: variations in equivalent concept labels (**Homonymies, Synonymies, various languages)**

example : KG/Knowledge Graph, paper/publication, book/livre

CONCEPTUAL HETEROGENEITY

Conceptual: differences between two models of the same domain (**coverage**, **granularity**, **different points of view**)

- **Coverage**: the two ontologies describe different domains but share a more or less important set of concepts and properties.
- **Granularity**: the two ontologies describe the same domain, but they are more or less accurate (ex: detailed bio-medical ontology for experts, versus simple ontology used by ordinary patients).
- **Distinct Point of views** (ex : geographical vs geopolitical)

WHAT IS AN ONTOLOGY ALIGNMENT?

Alignment A: set of mappings declared between ontology entities (properties or classes) of two ontologies O1 and O2.

$$f(O1,O2)=A$$

is the set of relations that can be used to express a mapping (chosed by the alignment approach).

WHAT IS AN ONTOLOGY ALIGNMENT?

Alignment A: set of mappings declared between ontology entities (properties or classes) of two ontologies O1 and O2.

$$f(O1,O2)=A$$

is the set of relations that can be used to express a mapping (chosen by the alignment approach).

Example

A = {

owl:equivalentClass(http://amaz.com/dvd, http://fnac.com/etol/filmdvd),
skos:closeMatch(http://amaz.com/road, http://ign.com/tronçon-route)

}

WHAT IS AN ONTOLOGY ALIGNMENT ?

- Each mapping can be associated with a **confidence** degree

skos:closeMatch(http://amaz.com/road, http://ign.com/tronçon-route, 0.8)

- A mapping can be **complex**

film(y) Λ realised(x,y) $\xrightarrow{0.9}$ (realisation(y,concat(x.firstName,x.Name)))

WHAT IS AN ONTOLOGY ALIGNMENT ?

- Each mapping can be associated with a **confidence** degree

skos:closeMatch(http://amaz.com/road, http://ign.com/tronçon-route, 0.8)

- A mapping can be **complex**

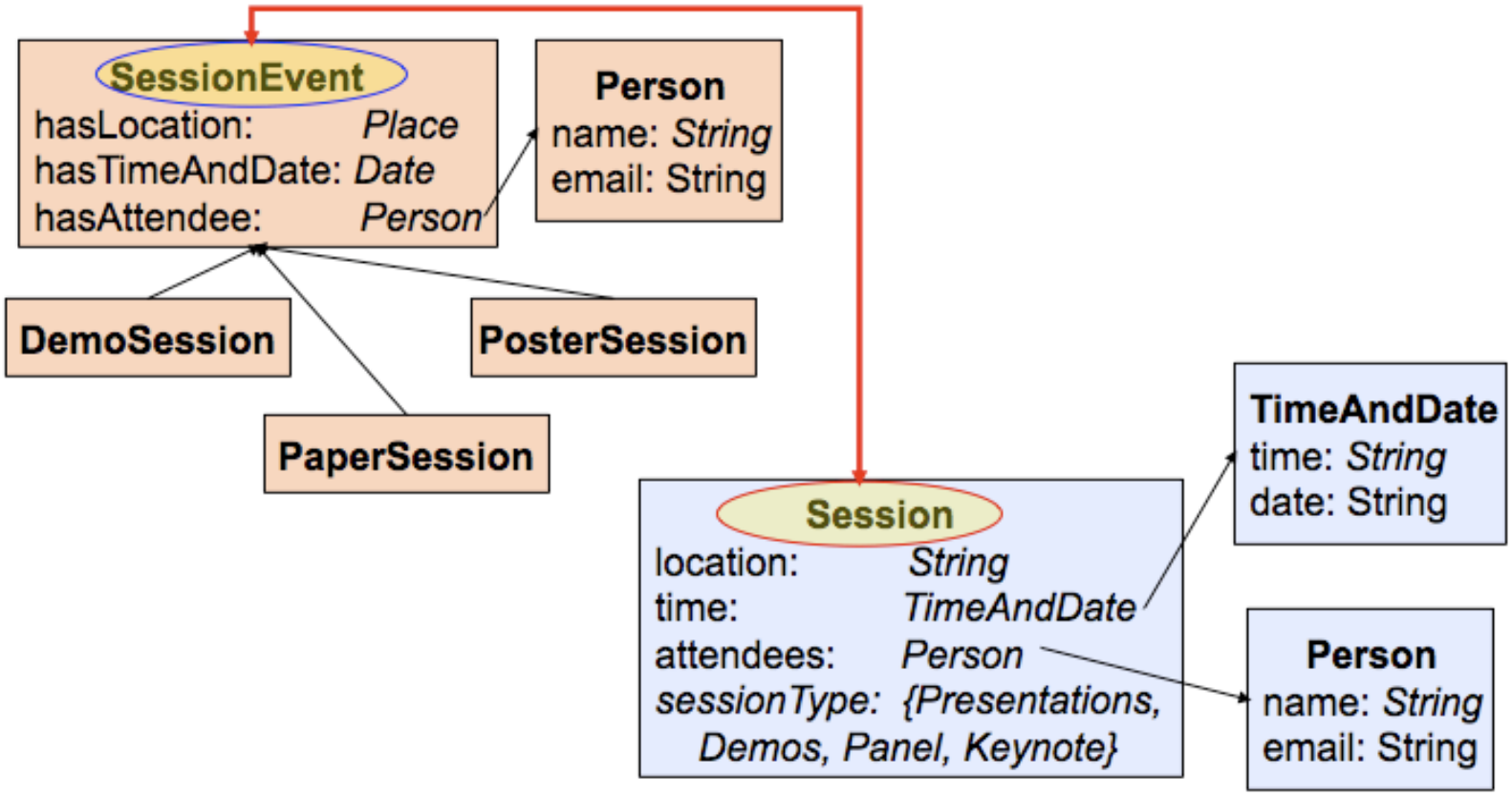
$$film(y) \wedge realised(x,y) \xrightarrow{0.9} (realisation(y,concat(x.firstName,x.Name)))$$

- An alignment can be1-1 (« **one to one** » mapping) bijective relation.

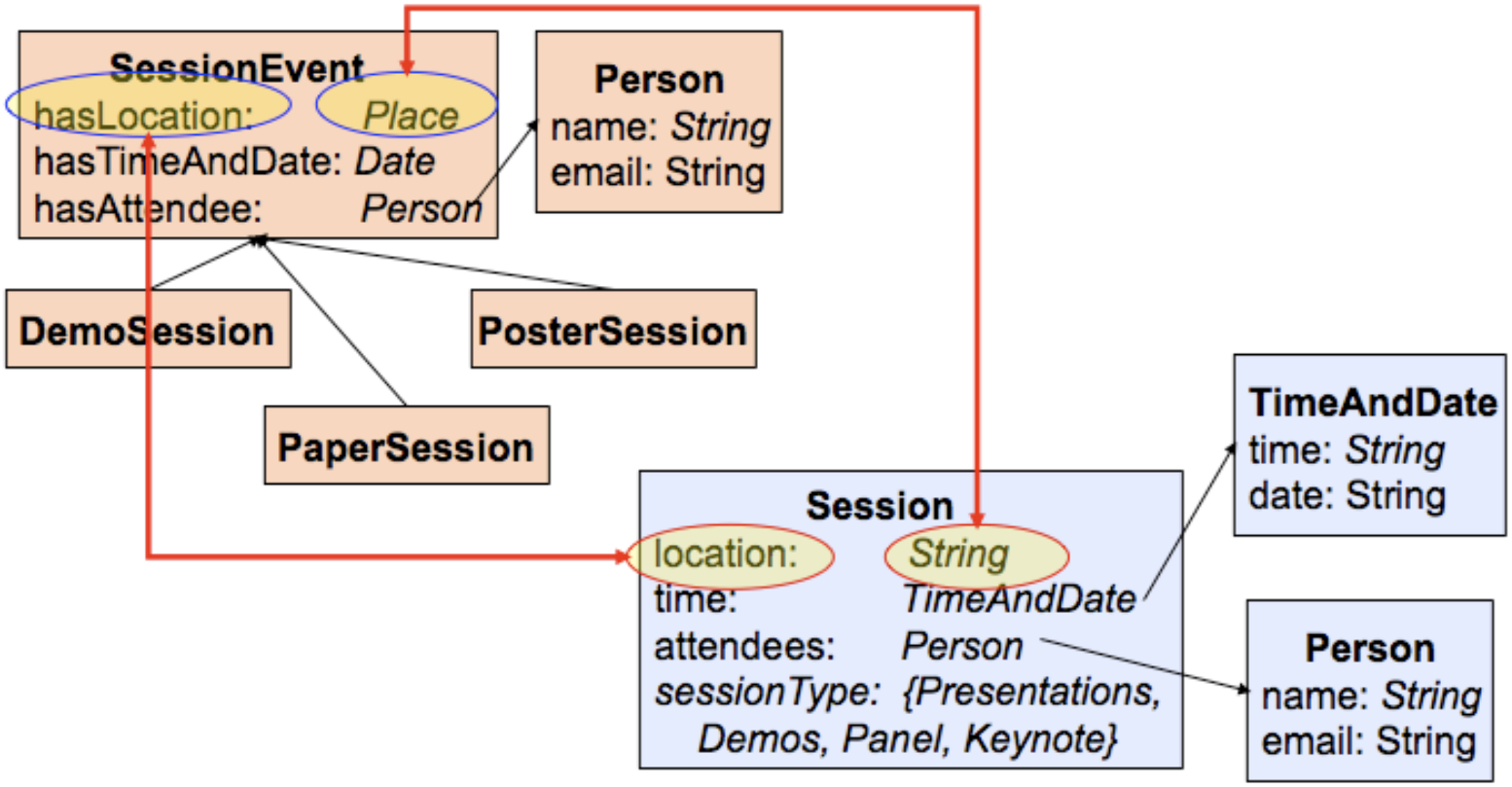
- Some alignment approaches are **not symmetric**

$$f(O1,O2) <> f(O2,O1)$$

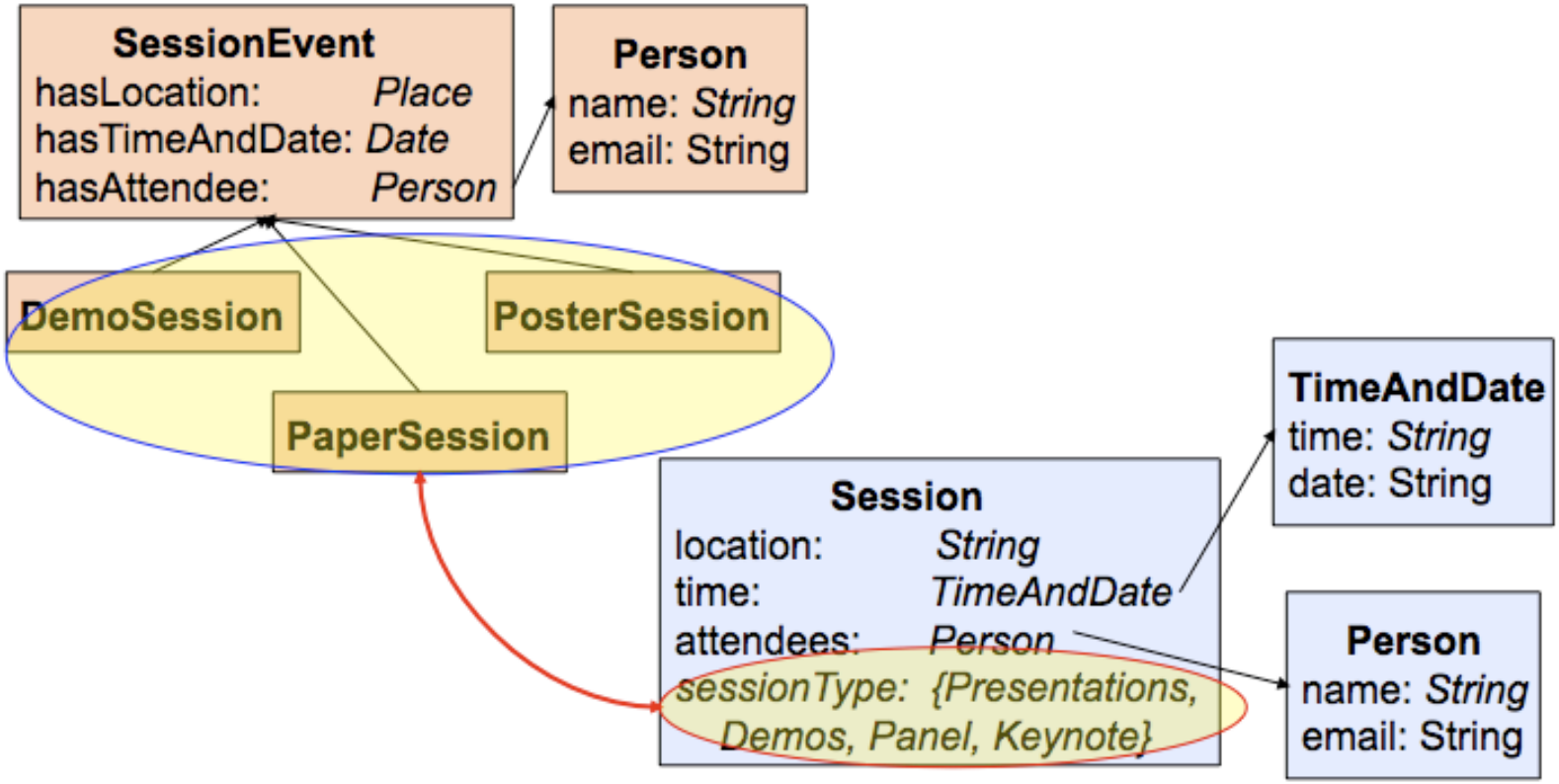
DIFFICULTIES



DIFFICULTIES



DIFFICULTIES



WHAT CAN BE TAKEN INTO ACCOUNT TO DECIDE?

- **Terminological Information** (concept labels, alt-labels, property names), comments, names of other linked entites, …)
  - Usual **similarity measures** can be used: token-based (n-grams, jaccard, or edit-based (levenstein, jaro-winkler) …

WHAT CAN BE TAKEN INTO ACCOUNT TO DECIDE?

- **Linguistic informations**
  - Stop-words (of, le, for, the,..)
  - Word weights can vary depending on its syntactic function (e.g., **an adjective is less important than a nominal** … *national road/ national organization*)

- **Terminological Information** (concept labels, alt-labels, property names), comments, names of other linked entities, …)
  - Usual **similarity measures** can be used: token-based (n-grams, jaccard, or edit-based (levenstein, jaro-winkler) …

SIMILARITY MEASURES

For more details: William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. 2003. **A comparison of string distance metrics for name-matching tasks**. In *Proceedings of the 2003 International Conference on Information Integration on the Web* (IIWEB'03), Subbarao Kambhampati and Craig A. Knoblock (Eds.). AAAI Press 73-78.

SIMILARITY MEASURES

- **Token based (e.g. Jaccard, TF/IDF cosinus) :**
  The similarity depends on the set of tokens that appear in both S and T.

- **Edit based (e.g. Levenstein, Jaro, Jaro-Winkler) :**
  The similarity depends on the smallest sequence of edit operations which transform S into T.

- **Hybrids (e.g. N-Grams, Jaro-Winkler/TF-IDF, Soundex)**

SIMILARITY MEASURES: TOKEN BASED

- **Jaccard measure:** **Jaccard(S,T) = |S∩T| / |S∪T|**

  Jaccard(« rue de la vieille pierre », « 11 rue vieille pierre ») =3/6

- **Cosinus (based on TF-IDF)**

  Widely used in traditional information retrieval (IR) approaches

  - Intuition: a term that is rare in the data is important and a term that is frequent in the string (value) is important.
  - Term frequency(TF): # of times a 'term' appears in the string compared with the size of the string.
  - Document frequency (IDF): the inverse of (# strings that contain the 'term'/ # of strings in the corpus)

SIMILARITY MEASURES: TOKEN BASED

Advantages:

- Efficient computation
- Word order is not significant

Disadvantages:

- Sensitive to spelling errors (Fathia, Sais)
- Sensitive to abbreviations (Univ. vs University)
- Sometimes order in words is meaningful (*Laurent Simon* vs *Simon Laurent*)

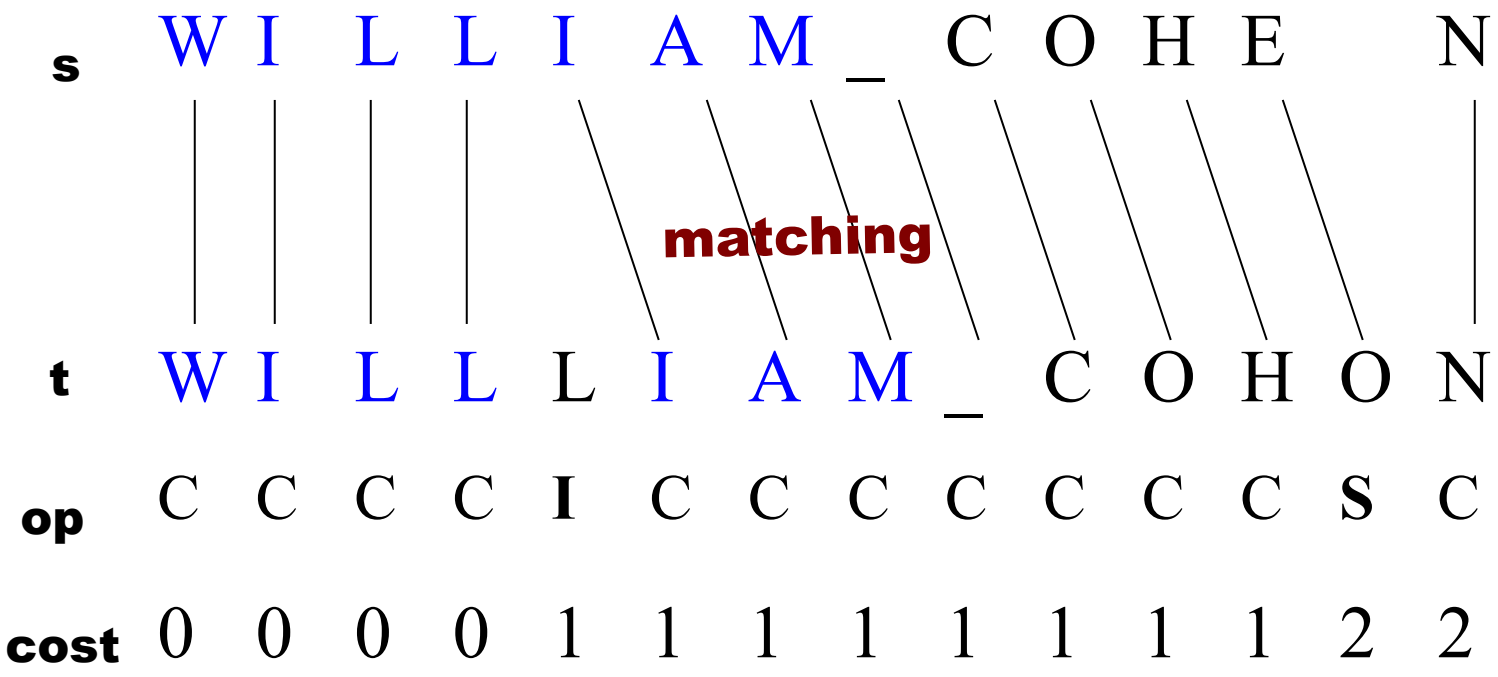
SIMILARITY MEASURES: EDIT BASED

- **Edit-based measure: "Levenstein" distance**
  - Character operations:
    - I (Insert), D(delete), R(replace), S (substitution).
  - Unit costs

  - Given two strings s,t edit(s, t):
    - Minimum cost sequence of operations to transform *s to t*.
    - Example: edit('Error', 'Eror')=1, edit('great', 'grate')=2

SIMILARITY MEASURES: EDIT BASED

- Levenstein("William Cohen", "Willliam Cohon")



SIMILARITY MEASURES: EDIT BASED

- **Jaro**
  - For (S, T), the character *c* is common for (S, T):
    if (S_i=c), (T_j=c), and |i-j| < min(|S|,|T|) / 2.
  - The character *c and d* are ***transpositions*** if *c and d* are common for S and T and appear in different orders in S and T.

$$Jaro(S,T) = \frac{1}{3}\left(\frac{m}{|S|} + \frac{m}{|T|} + \frac{m-t}{m}\right)$$

  - Example: Jaro(Texas, Texhas) = $\frac{1}{3}\left(\frac{5}{5} + \frac{5}{6} + \frac{5-2}{5}\right)$ = 0,81

SIMILARITY MEASURES: EDIT BASED

Jaro-Winkler

- An extension of Jaro by considering the size of the longest prefix between S and T.

$$Jaro-Winkler(S,T) = Jaro(S,T) + \left(\frac{\max(P,4)}{10} * (1 - Jaro(S,T)\right)$$

- Example : Jaro-Winkler(Texas, Texhas) $= 0{,}81 + \left(\frac{4}{10} * (1 - 0{,}81)\right)$

$= 0{,}88$

- Runtime efficiency
- Showed to be relevant for the comparison of person names in the literature.

SIMILARITY MEASURES: EDIT BASED

Advantages:

- Robustness when spelling errors exist
- Word order is significant

Disadvantages:

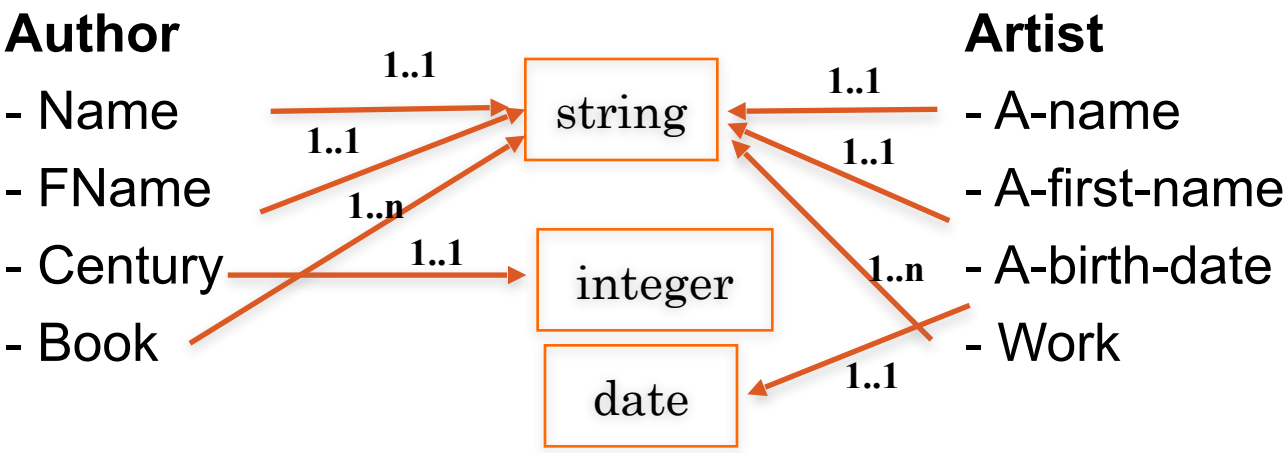
- High runtime
- Sometimes order in words is not meaningful (Univ. Paris Saclay and Paris Saclay University)

ONTOLOGY ALIGNMENT (CONT.)

WHAT CAN BE TAKEN INTO ACCOUNT TO DECIDE?

Ontology Structure :

Internal Structure (datatype properties, range, cardinalities)



SIMILARITY OF INTERNAL STRUCTURES

Compatibility measures can be defined [Valchev et Euzenat 97, 04]

→ **To compare the range (table)**

- compatibility (integer,date)=0.5
- compatibility(integer, number)=0.9

SIMILARITY OF INTERNAL STRUCTURES

Compatibility measures can be defined [Valchev et Euzenat 97, 04]

→ **To compare the range (table)**

- compatibility (integer,date)=0.5
- compatibility(integer, number)=0.9

→ **To compare cardinalities**

As it can be done for an JSON Schema : **compatibility (*, +)=0.9**

In OWL: mincardinality b /maxcardinality e

Sim([b1 e1], [b2 e2])=0

if b2>e1 or b1>e2

else = (min(e1,e2)-max(b1,b2))/(max(e1,e2)-min(b1,b2))

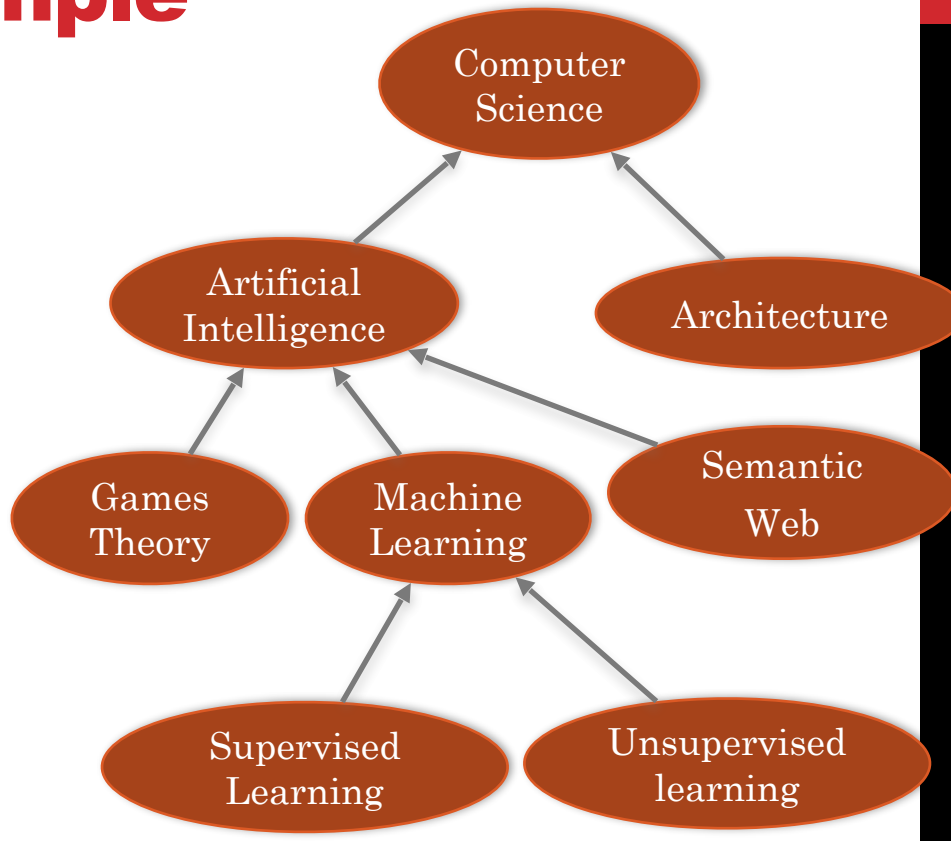
SIMILARITY OF EXTERNAL STRUCTURES

External Structure (related concepts)

Hypothesis: the more 2 concepts are similar, the more their linked concepts are similar.

- Given a property r (usually r=subclassOf)
- r : concepts that are directly linked using r
- r+: concepts that belong to the transitive closure (w.r.t the property r)
- r-1; more general concepts
- r! : leaves that belong to the transitive closure of subclassOf.

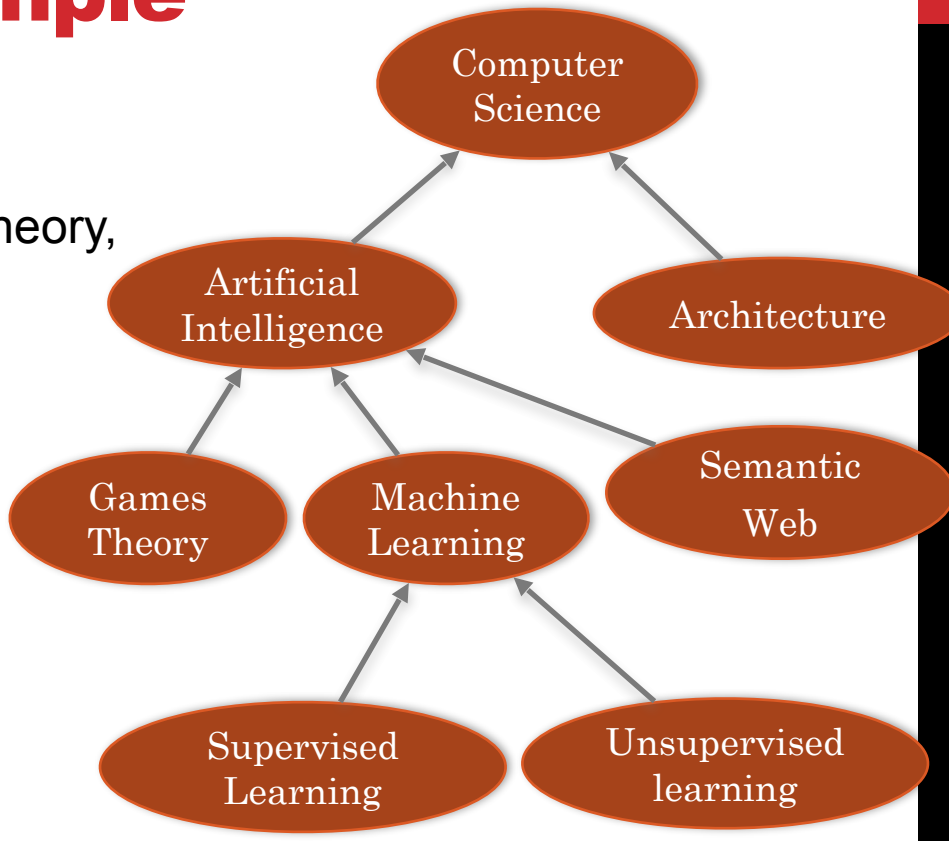
SIMILARITY OF EXTERNAL STRUCTURES: example



SIMILARITY OF EXTERNAL STRUCTURES: example

With r = subclassOf

subclassOf(Artificial Intelligence)= {Game theory, Machine Learning, Semantic web}



SIMILARITY OF EXTERNAL STRUCTURES: example

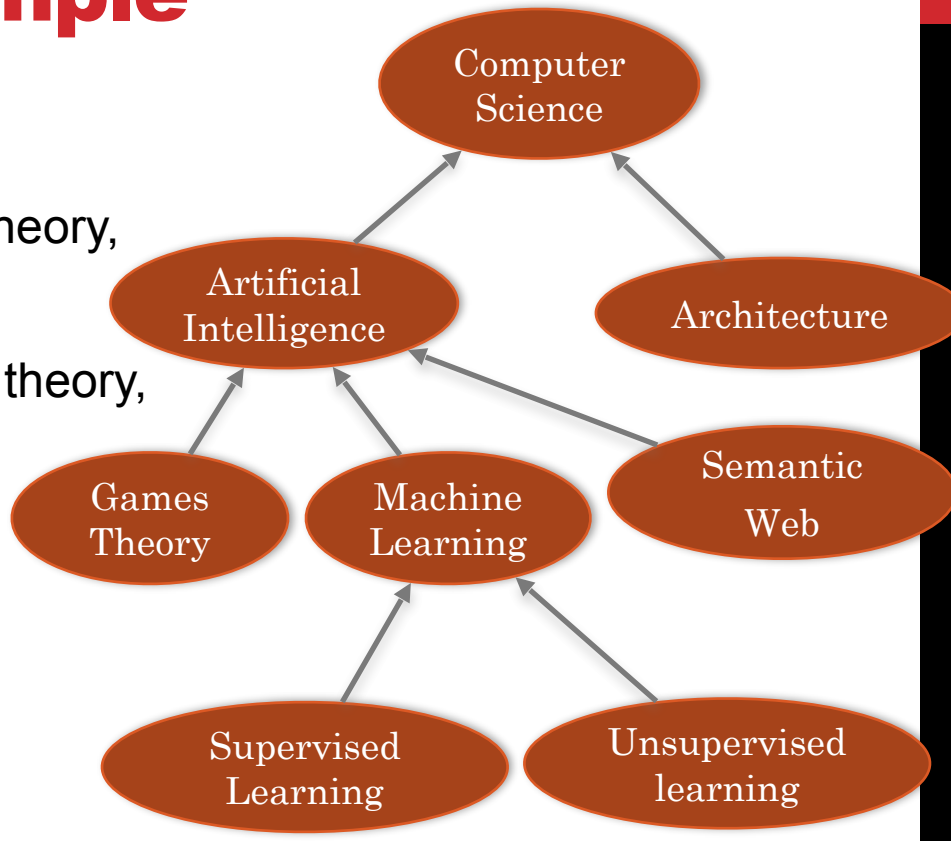
With r = subclassOf

subclassOf(Artificial Intelligence)= {Game theory, Machine Learning, Semantic web}

subclassOf+(Artificial Intelligence)= {Game theory, Machine Learning, Supervised Learning, Unsupervised learning, Semantic web}

subclassOf-1(Artificial Intelligence)=

{Computer Science}



SIMILARITY OF EXTERNAL STRUCTURES: example

With r = subclassOf

subclassOf(Artificial Intelligence)= {Game theory, Machine Learning, Semantic web}

subclassOf+(Artificial Intelligence)= {Game theory, Machine Learning, Supervised Learning, Unsupervised learning, Semantic web}

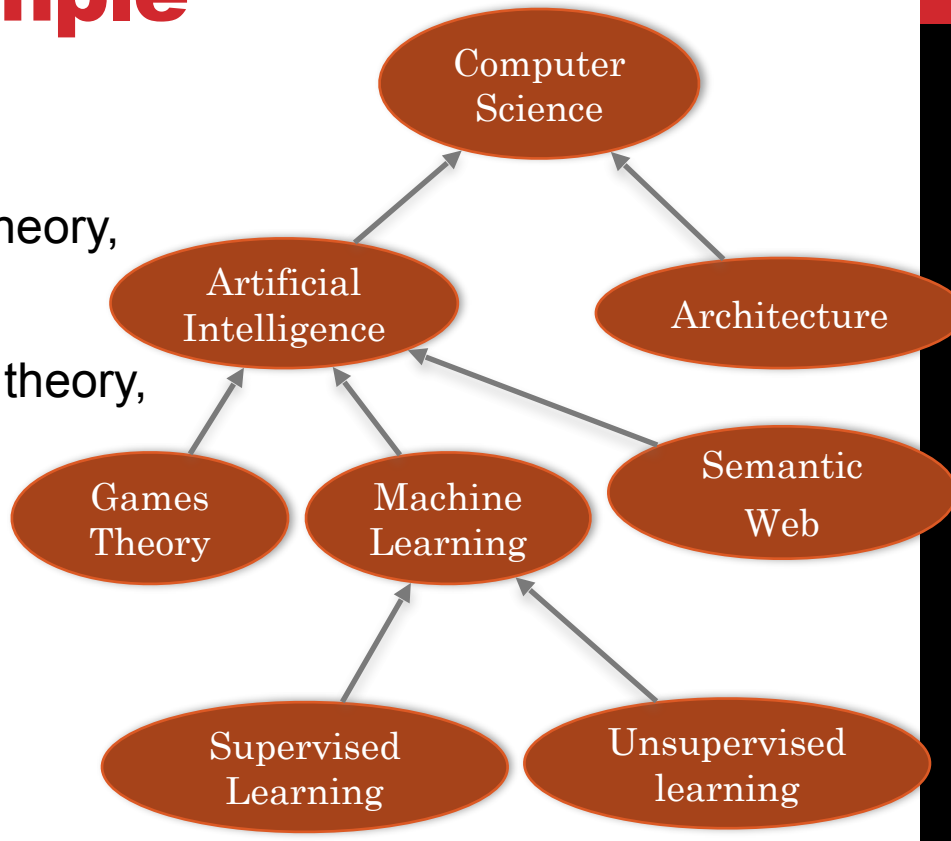
subclassOf-1(Artificial Intelligence)=

{Computer Science}

subclassOf!(Artificial Intelligence)=

{Game theory, Supervised Learning,

Unsupervised learning, Semantic web}



EXTERNAL RESOURCE

- Two concepts can be compared using an external ontology or taxonomy they also belong to (or they have been found in).

- **Similarity Measure**: Wu et Palmer [Wu &al. 94]

  *Sim(C1, C2) = 2*level(C)/ (level(C1) + level (C2))*

where C is the **least common subsumer** (LCS) of C1 and C2.

Level of the root = 1.

EXTERNAL RESOURCE

Examples

Sim(Supervised Learning, Unsupervised Learning)

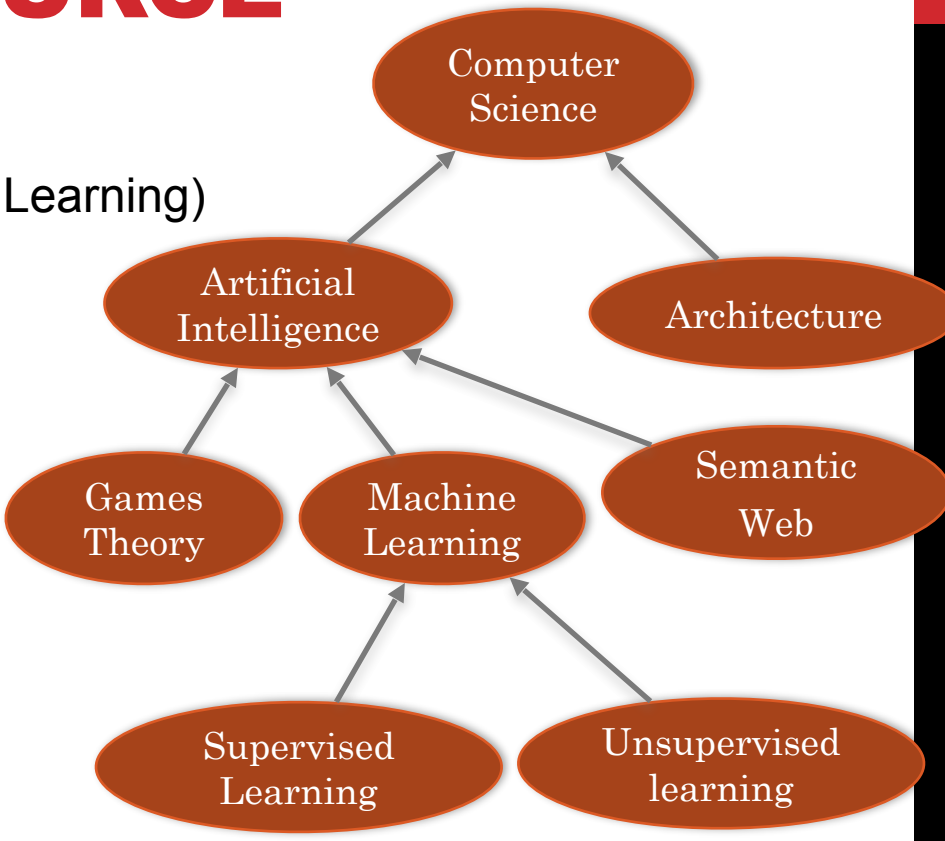
= 2* level(Machine Learning)/
level(Supervised Learning)+
level(Unsupervised Learning)
= (2*3)/(4+4)=6/8=0.75

Sim(Art Intelligence, Architecture)

=(2*1)/(2+2)=0.5

Sim(Supervised Learning, Architecture)

= (2*1)/(4+2)=0.33



EXTERNAL RESOURCE

Semantic closeness

Results that can be obtained using Wordnet [Madche et Zacharias 2002]

	illustrator	**author**	**creator**	**person**
illustrator	1	0.5	0.67	0.4
author		1	0.67	0.4
creator			1	0.67
person				1

HOW TO USE CONCEPT INSTANCES (EXTENSIONAL INFORMATION)?

Naive way: compute the intersection of two concept extensions.

- Equivalence: if C1 ∩ C2 = C1 = C2
- Subsumption: if C1 ∩ C2 = C1
- Disjunction: if C1 ∩ C2 is empty

And more generally, define and use a similarity measure based on the **size of this intersection.**

- Example: (jaccard) (C1 ∩ C2)/(C1 ∪ C2)

[RIMOM14] [PARIS12]

EXTENSIONAL APPROACHES

- **Problem**: how do you know that two instances are the same.

- Possible if instance URI are the same or if different but identical instances share the same document URL that describes them (ex: DBPedia/Yago).

- ……. If not, a linking method is needed (that detect identity links) .

SCALABILITY STRATEGIES

- **Computation of the similarity scores first** then selection of the proposed mappings
- Combinaison of **different strategies in parallel**
- Sequential application of different strategies
- Ontology **partionning**

MAPPING SELECTION

Thresholds are defined ….

- **Absolute threshold (at)**: select all mappings such that the similarity score is > at
- **Relative threshold (rt)**: select all mappings such that the **similarity score > MaxScore – rt**
- **Rate (n%)**: select the **n%** best scores in A.

MAPPING SELECTION

Example

	book	translator	editor	author
product	.84	0	.9	.12
provider	.12	0	.84	.60
artist	.60	.05	.12	.84

- **Absolute threshold 0.7**: (product, book), (product, editor) (provider,editor), (artist, author)
- **Relative Threshold delta 0.3**: (product, book), (product, editor) (providor,editor), **(providor,author), (artist, book)**, (artist, author)
- **Rate =30% (4):** (product, book), (product, editor), (provider, editor), (artist, author)

MAPPING SELECTION

- When a 1-1 alignment is wanted, a choice is necessary
- **Greedy algorithm** (best score at each iteration).
- **Stable marriage problem** (local optimum):
  - Given 2 entity (concept) sets E et E' and
  - a similarity function sim : E x E' → [0,1],

  - Selection of an alignment A s.t. for all (e1,e2) in A, and for all (e3,e4) in A, we have:

sim(e1,e2) + sim(e3,e4) ≥ sim(e1,e3) + sim(e2,e4)

MAPPING SELECTION

Maximum weight graphs (global optimum):

Given 2 entity (concept) sets E et E' and a similarity function sim : E x E' → [0,1], selection of an alignment A s.t.:

Sum[(sim(ei, ej) in A] >= Sum[(sim(ei, ej) in A']

OAEI Competition

Ontology Alignment Evaluation Initiative (OAEI)

→ every year since 2006

Different Tracks (ontology size, formalism, domain, instances or not)

Various results

(**< 50%** of precision/recall,

to **100%** when ontology are highly similar)

http://www.ontologymatching.org

ANATOMY ONTOLOGIES – RESULTS

OAEI 2024 - ANATOMY BENCHMARK

Matcher	Runtime	Size	Precision	F-Measure	Recall	Recall+	Coherent
Matcha	42	1485	0.951	0.941	0.931	0.82	-
TOMATO	2154	572	0.955	0.523	0.36	0.024	-
MDMapper	121	1441	0.926	0.903	0.881	0.703	-
ALIN	370	1156	0.984	0.851	0.75	0.489	-
LogMap	12	1402	0.917	0.881	0.848	0.602	+
LogMapBio	1346	1549	0.888	0.898	0.908	0.757	+
LogMapLite	2	1147	0.962	0.828	0.728	0.288	-
StringEquiv	-	946	0.997	0.766	0.622	0.000	-

https://oaei.ontologymatching.org/2024/results/anatomy/index.html

ANATOMY ONTOLOGIES – RESULTS

OAEI 2024 - ANATOMY BENCHMARK

```
<?xml version='1.0' encoding='utf-8' standalone='no'?>
<rdf:RDF xmlns='http://knowledgeweb.semanticweb.org/heterogeneity/alignment#'
         xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
         xmlns:xsd='http://www.w3.org/2001/XMLSchema#'>
<Alignment>
  <xml>yes</xml>
  <level>0</level>
  <type>?*</type>
  <onto1>
    <Ontology rdf:about="http://mouse.owl">
      <location>file:/home/minab62/Documents/OAEI2024_copy/anatomy-dataset/mouse.owl</location>
    </Ontology>
  </onto1>
  <onto2>
    <Ontology rdf:about="http://human.owl">
      <location>file:/home/minab62/Documents/OAEI2024_copy/anatomy-dataset/human.owl</location>
    </Ontology>
  </onto2>
  <map>
    <Cell>
      <entity1 rdf:resource='http://mouse.owl#MA_0000029'/>
      <entity2 rdf:resource='http://human.owl#NCI_C12664'/>
      <relation>=</relation>
      <measure rdf:datatype='http://www.w3.org/2001/XMLSchema#float'>1.0</measure>
    </Cell>
  </map>
  <map>
    <Cell>
      <entity1 rdf:resource='http://mouse.owl#MA_0000520'/>
      <entity2 rdf:resource='http://human.owl#NCI_C32040'/>
      <relation>=</relation>
      <measure rdf:datatype='http://www.w3.org/2001/XMLSchema#float'>1.0</measure>
    </Cell>
  </map>
```

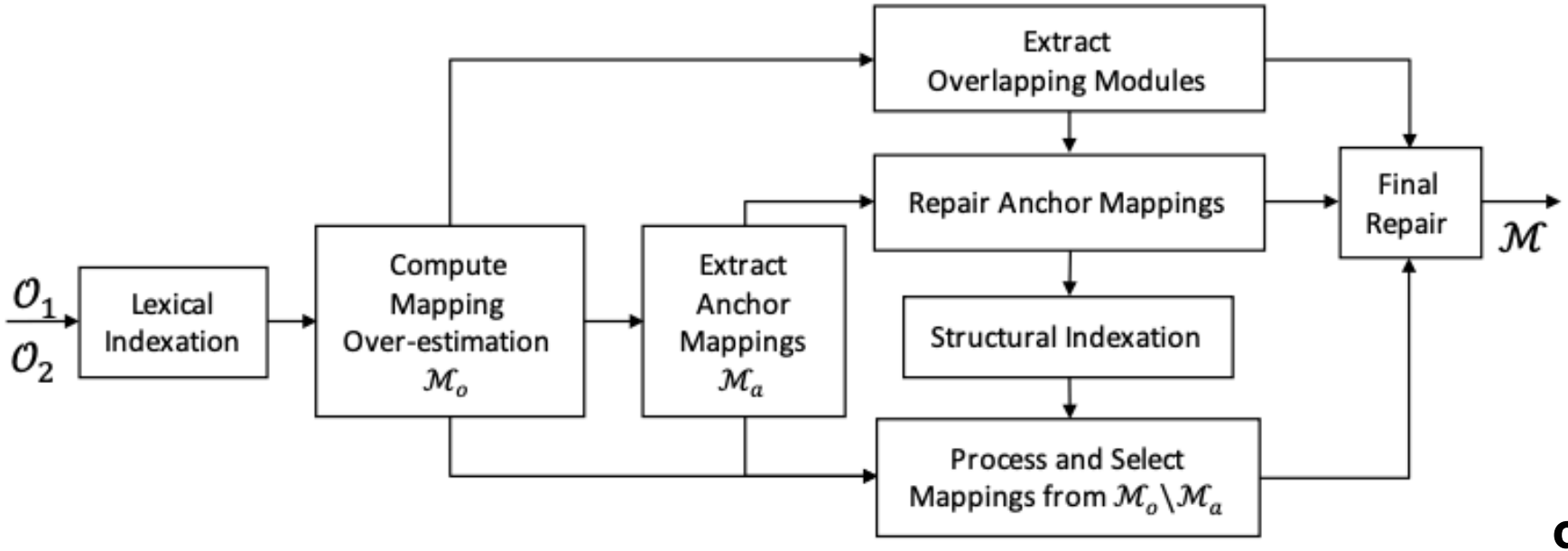
LOGMAP - AN EXAMPLE OF ONTOLOGY ALIGNMENT SYSTEM

https://github.com/ernestojimenezruiz/logmap-matcher

LogMap extracts mappings between classes, properties and instances.

It is able to:

- efficiently match semantically rich ontologies containing hundreds of thousands of classes,
- incorporate sophisticated reasoning and repair techniques to minimise the number of logical inconsistencies, and
- provide support for user intervention during the matching process



Ontology Alignement: Challenges

- How to interact with user experts during the alignment process (validation, updates, suppression)
- How to explain discovered mappings
- How to combine Machine learning and Symbolic approaches
- How to mix data linking and ontology alignment processes
- How confidence degrees that can be associated to mappings can be used
- How to scale
- Ontology evolution → How to update alignments efficiently

REFERENCES

- “Ontology Matching” by Euzenat and Shvaiko , springer, 2007.
- AML (AgreementMakerLight) : D. Faria, C. Martins, A. Nanavaty, D. Oliveira, B. S. Balasubramani, A. Taheri, C. Pesquita, F. M. Couto, and I. F. Cruz. AML results for OAEI 2015. In *Ontology Matching Workshop*. CEUR, 2015.
- Proceedings des conferences **ISWC, ESWC, WWW, EKAW, K-CAP.**
- *Journal of web semantics, Journal on data semantics.*

- *http://www.ontologymatching.org*