COMP318: Ontology alignment

www.csc.liv.ac.uk/~valli/Comp318

Dr Valentina Tamma
Room: Ashton 2.12
Dept of computer science
University of Liverpool
V.Tamma@liverpool.ac.uk

LIVERPOOL

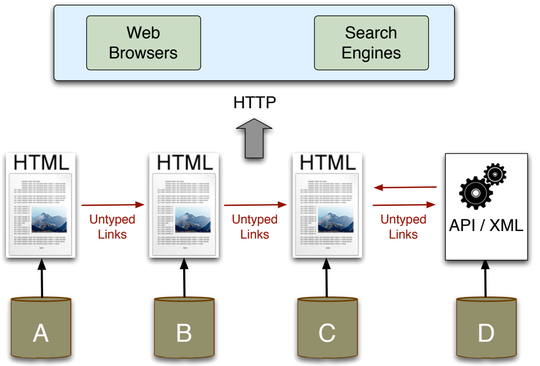
based on slides by V. Tamma, T.R. Payne, J. Euzenat, K. Janowicz and G. Schreiber

Where were we

- Ontology engineering
  - methodology
  - representation in OWL
  - violations and use of reasoning services

The global dataspace

- The Web allowed the creation of a global information space
  - where structured, semi structured and unstructured information is shared
- By 2012 over 3,000 exabytes of data were be available (IDC and UC Berkeley)
  - Mainly generated through transactions, but later from interactions
- By the end of 2016, global Internet traffic reached 1.1 zettabytes per year, according to Cisco
  - 1 zettabyte = 1 sextillion bytes, or 1,000 exabytes
- Things are now connected online
  - Sensors, devices, appliances… all publishing data
  - “A cross country flight from New York to Los Angeles on a Boeing 737 plane generates a massive 240 terabytes of data”. GigaOmni Media



Diversity in models: friend or foe?

- Different systems (sensors, services, applications, agents...) that generate or use knowledge usually make use of different ontologies
  - Similar or overlapping information is modelled in diverse ways even inside organisations with strong governance and internal communication
- These differences in modelling are an obstacle to systems' interoperability

Interoperability example

- You own a company selling digital cameras
  - You organise your information according to your own schema

- Task: You want your company to sell in a marketplace, e.g.:
  - Ebay:
    - *Home > Buy > Cameras & Photo > Digital Cameras > Digital SLR > Nikon > D34XX*
  - Amazon marketplace:
    - *Home > Department > Electronic & Computers > Camera & Photo > Digital Cameras > Digital SLR > Nikon > D34XX*
- Mappings between:
  - the entries in your schema
  - to the entries of the common catalogues of the marketplaces

The need for interoperability

- Interoperability measures the extent to which different systems are able to meaningfully exchange information
  - with the aim of reaching some common goal.

- Some (subtle) considerations
  - Data does not interoperate, but is used to support interoperation.
  - (Inter)operability assumes intention and purpose, e.g., a goal.
  - Interoperability is about a degree of meaningful exchange

- Syntactic vs semantic interoperability

The need for interoperability

- **Syntactic interoperability**:
  - If two or more systems are capable of communicating with each other, they exhibit syntactic interoperability when using specified data formats and communication protocols.
    - XML or SQL standards are among the tools of syntactic interoperability

- **Semantic interoperability**:
  - the ability to automatically interpret the information exchanged meaningfully and accurately in order to produce useful results
    - as defined by the end users of both systems.
  - The meaning of the information exchanged is unambiguously defined: what is sent is the same as what is understood.

Standards!

- Standards for representing, publishing, sharing, querying facts, knowledge, data, software and processes

- Provide a scalable approach for the discovery of knowledge that is
  - formulated in different ways, from independent actors
  - distributed physically and logically

Reusing vocabularies

FOAF Core

- Agent
- Person
- name
- title
- img
- depiction (depicts)
- familyName
- givenName
- knows
- based_near
- age
- made (maker)
- primaryTopic (primaryTopicOf)
- Project
- Organization
- Group
- member
- Document

Social Web

- nick
- mbox
- homepage
- weblog
- openid
- jabberID
- mbox_sha1sum
- interest
- topic_interest
- topic (page)
- workplaceHomepage
- workInfoHomepage
- schoolHomepage
- publications
- currentProject
- pastProject
- account
- OnlineAccount
- accountName
- accountServiceHomepage
- PersonalProfileDocument
- tipjar
- sha1
- thumbnail
- logo

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: process Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) procedure, **process** (a particular course of action intended to achieve a result) *"the procedure of obtaining a driver's license"; "it was a process of trial and error"*
- S: (n) **process**, cognitive process, mental process, operation, cognitive operation ((psychology) the performance of some composite cognitive activity; an operation that affects mental contents) *"the process of thinking"; "the cognitive operation of remembering"*
- S: (n) summons, **process** (a writ issued by authority of law; usually compels the defendant's attendance in a civil suit; failure to appear results in a default judgment against the defendant)
- S: (n) **process**, unconscious process (a mental process that you are not directly aware of) *"the process of denial"*
- S: (n) **process**, outgrowth, appendage (a natural prolongation or projection from a part of an organism either animal or plant) *"a bony process"*
- S: (n) **process**, physical process (a sustained phenomenon or one marked by gradual changes through a series of states) *"events now in process"; "the process of calcification begins later for boys than for girls"*

Verb

- S: (v) **process**, treat (subject to a process or treatment, with the aim of readying for some purpose, improving, or remedying a condition) *"process cheese"; "process hair"; "treat the water so it can be drunk"; "treat the lawn with chemicals"; "treat an oil spill"*
- S: (v) **process** (deal with in a routine way) *"I'll handle that one"; "process a loan"; "process the applicants"*
- S: (v) **process** (perform mathematical and logical operations on (data) according to programmed instructions in order to obtain the required information) *"The results of the elections were still being processed when he gave his acceptance speech"*
- S: (v) action, sue, litigate, **process** (institute legal proceedings against; file a suit against) *"He was warned that the district attorney would process him"; "She actioned the company for discrimination"*
- S: (v) march, **process** (march in a procession) *"They processed into the dining room"*

DEVELOP

GETTING STARTED | DATASETS | SUPPORT | GITHUB | ONTOLOGY

DBPEDIA VERSION 2016-10

DBpedia version 2016-10

Dataset category: DBpedia release
Publication Year: 2017

This release is based on updated Wikipedia dumps dating from October 2016.

You can download the new DBpedia datasets in N3 / TURTLE serialisation from http://wiki.dbpedia.org/downloads-2016-10 or directly here http://downloads.dbpedia.org/2016-10/.

This release took us longer than expected. We had to deal with multiple issues and included new data. Most notable is the addition of the NIF annotation datasets for each language, recording the whole wiki text, its basic structure (sections, titles, paragraphs, etc.) and the included text links. We hope that researchers and developers, working on NLP-related tasks, will find this addition most rewarding. The DBpedia Open Text Extraction Challenge (next deadline Mon 17 July for SEMANTiCS 2017) was introduced to instigate new fact extraction based on these datasets.

We want to thank anyone who has contributed to this release, by adding mappings, new datasets, extractors or issue reports, helping us to increase coverage and correctness of the released data. The European Commission and the ALIGNED H2020 project for funding and general support.

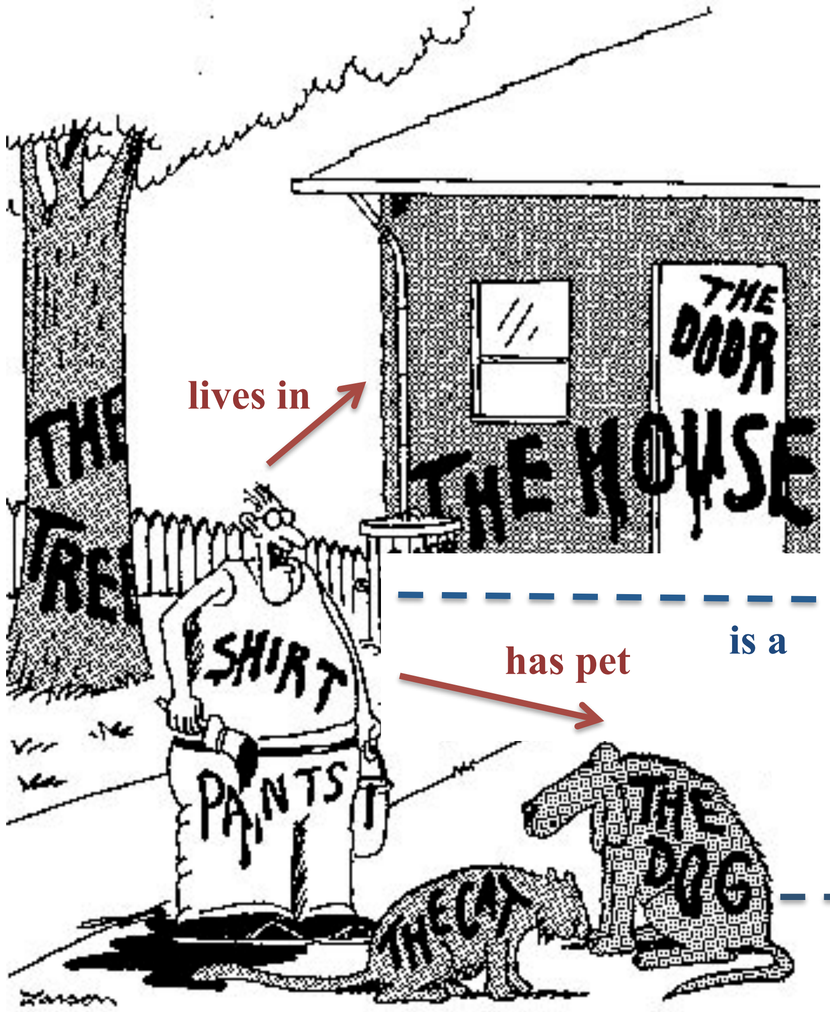
Join and support DBpedia

The active community of developers and engineers comes together in the DBpedia Community Committee. We will extend this Committee with the help of Pablo Mendes and Magnus Knuth. Students wishing to join should be or become a member of the DBpedia Association. Please check all benefits and details on our website.

Every first Wednesday of the month we organise regular development online meetings. You can join the next DBpedia dev telco on Wednesday, 5th of July (@ 2 pm CET). All info regarding the telco can be found here: http://tinyurl.com/DBpediaDevMinutes.

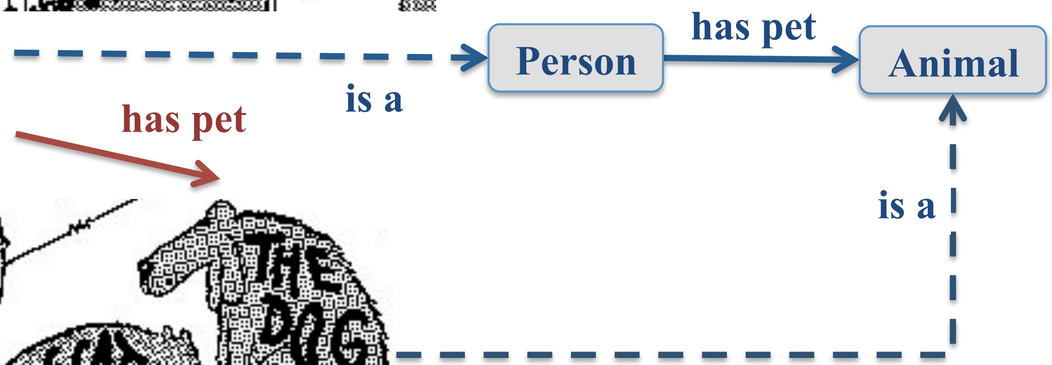
V.Tamma

Semantic Web?



Concrete Facts
Resource Description Framework

General Knowledge
Web Ontology Language



"Now! – *That* should clear up a few things around here!"

8

slide by F. van Harmelen

What does meaning mean

- Semantics is the study of meaning:
  - the study of how and what symbols denote.
  - Meaning can be seen as an emergent property of interaction.

- Humans use complex processes and negotiate the approximate, common meaning of terms during interaction.
  - contextual cues, gestures…
  - Unfortunately, we cannot do so with data.

What does meaning mean here

- The implied assumption behind the call for collecting raw data is that data can be reused outside its original creation context and is free of interpretation:
  - However, there is no such thing as raw data.
    - Data is always created following particular workflows, procedures, sampling strategies,
    - is derived using specific instrumentation, is pre-processes in specific ways.

- Ontologies support interoperability and integration:
  - they cannot fix meaning,
  - but, they formally restrict the possible interpretations of the terms in a domain to intended meaning.

Myth busting slide

- **Myth 1:** Once we have an ontology data becomes interoperable
  - Actually there are already several ontologies for a domain:
    - Ontology alignment;
    - Reference ontologies;
  - It is still meta-data, machines can only interpret it. It still needs intelligent processing, decision making and reasoning to process this data
- **Myth 2:** Semantics makes my data machine-understandable, ergo my system will be intelligent
- **Myth 3:** It's a hype, ontologies and semantics are too much overhead. What about tiny devices?
  - Ontologies are a way to share and agree on a common vocabulary and knowledge in a machine interpretable and reusable format;
  - Semantic metadata does not need to be added in the source, it can be added to the data at a later stage (e.g. in a gateway)
  - Legacy applications can be extended to work with it.
  - And … "a little semantics goes a long way" (Jim Hendler & Tim Berners Lee)

No unified vocabulary

- There is never “the” correct ontology:
  - a number of different ontologies can represent the same domain
  - they all capture different contexts, perspectives, requirements
  - and depend on the task that the ontology should support
    - performed by some (autonomous) system — agent / service / API / …
- These differences in modelling become apparent when
  - These systems must be combined (integration)
  - Or be made to work together (interoperation)

The Architect

When modelling a bridge, important characteristics include:

tensile strength
weight
load
etc

Pat Hayes in conversation with T.R. Payne, 2001

The Military

When modelling a bridge, important characteristics include:

what munitions are required to destroy it!

Semantic integration

- **Semantic Integration** is often treated as a high level cognitive task
  - even when the associated computational artefacts are low level

- Ontologies are a computational representation of the cognitive level, and are the underlying basis for automating semantic data integration.

- **Ontology alignment** is the process of determining correspondences between semantically related entities
  - classes, relationships and instances

Aligning ontologies

- Ontology alignment or mapping:
  - Use ontologies together by defining a set of “links” (mappings or correspondences)
    - Mappings can be of limited types, i.e. only certain logical relations
- Advantages:
  - Benefit from knowledge encoded in the other ontologies models
  - Enable access from different agents/services and across different collections
  - Partial by nature, does not need to cover the entire ontology

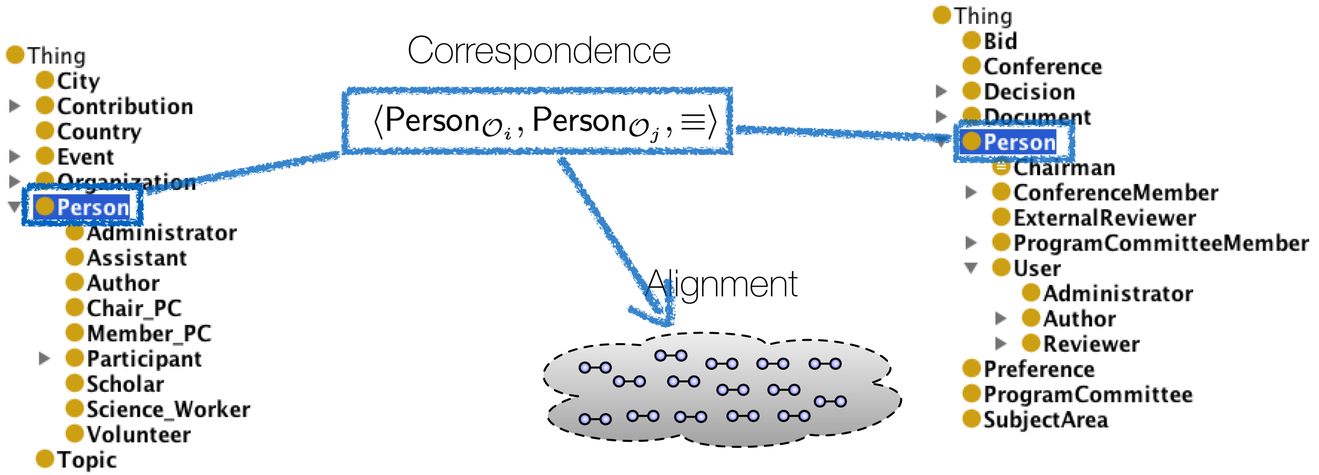
Alignment approaches

- Different alignment approach are available depending on
  - the expressivity of the two ontologies O and O’
  - The availability of additional inputs to the matching process:
    - Oracles, input alignment and external resources, i.e. Wordnet or BabelNet
  - The entities to match:
    - Only the T-box or schema, i.e. classes and possibly properties
    - Instances

- The majority of current ontology alignment systems align classes, and restrict the relationships to equivalence

How to represent a correspondence

- *Given two ontologies O and O', an alignment A is the set of correspondences c between the entities $e \in O$ and $e' \in O'$*
  - *A correspondence c is the tuple* c = <e, e', r, w>
    - $e \in O$ and $e' \in O'$, where e and e' can be classes, properties, individuals
    - r = {≡, ⊑, ⊥} and w ∈ [0,…, 1] is the weight



Types of correspondence relation between classes/ properties

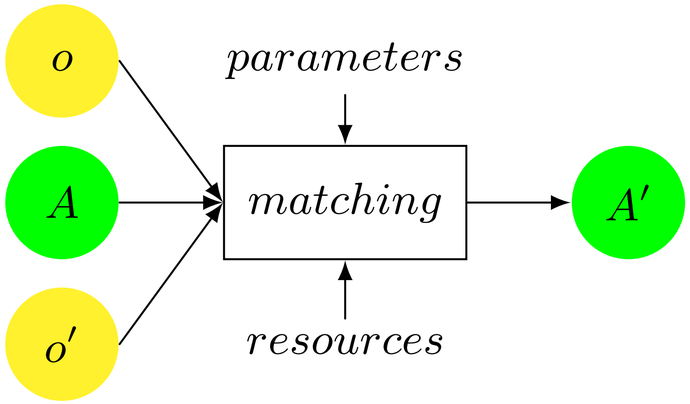
	OWL	Example
≡ Equivalence	`owl:EquivalentClass`	O:Person ≡ O':Person
⊑ Subclass	`rdfs:subClassOf` `rdfs:subPropertyOf`	O:Assistant ⊑ O':User
⊥ Disjointness	`owl:disjointWith`, `owl:allDisjointClasses`	O:Topic ⊥ O':Person

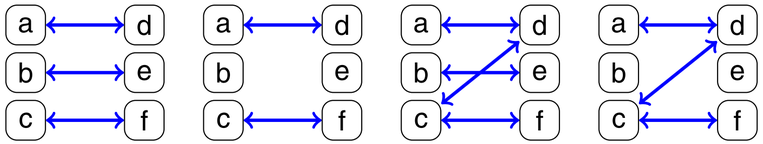
Types of correspondence relation between classes/ properties

	OWL	Example
= Equivalence	`owl:sameAs`	O:Florence = O':Firenze
≠ Difference	`owl:differentFrom`	O:John ≠ O':Ringo
∈ Instance	`rdf:type`	O:Beatles ∈ O':MusicGroup

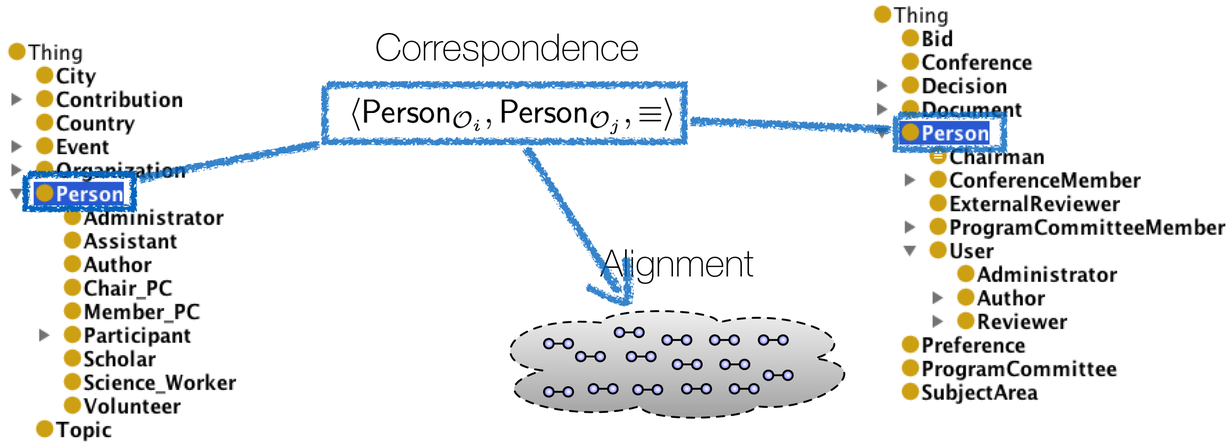
Ontology alignment and the matching process

- Alignments are generated through *a matching process*, a function *f*
  - Input: two ontologies *O and O', and an optional input alignment* A_{input} , *set of parameters par, oracles and resources res*
  - Output: an alignment *A' between O and O'*
    - *A' = f(O, O', A_{input}, par, res)*
    - Set of possible correspondences, with the relationships between entities of O and O'
      - Different multiplicities possible





Alignment approaches



String based techniques

- **Syntactic approaches**
  - based on the comparison of the labels used to denote entities
  - Exploit measures of syntactic distance
  - Assume language pre-processing
    - Stop word removal, Tokenization, Stemming
- **Taxonomy comparison**
  - Identify common parents/children in the taxonomy
    - Semantic similarity in taxonomies: Rada (1989), Resnick (1999)…
- **Language based approaches**
  - Use external thesauri or multi-lingual resources
    - Relate terms in the WordNet hierarchy
- **Instance based mapping**
  - Extensional, based on instance sets

String based techniques

- String based techniques are used to assess the similarity between the labels used to denote classes and properties in the ontologies.

- Exact string match
  - Prefix
    - takes as input two strings and checks whether the first string starts with the second one
      - *The prefix is a substring that is at the beginning of the original string*
    - ***net*** = ***network***; but also ***hot*** = ***hotel***
  - Suffix
    - takes as input two strings and checks whether the first string ends with the second one
      - *The suffix is a substring that is at the end of the original string*
    - ***word*** = ***sword***; but also ***nana*** = ***banana***

String based techniques

- Edit distance
  - takes as input two strings and calculates the number of edit operations
    - counting the minimum number of operations required to transform one string into the other.
      - *e.g., insertions, deletions, substitutions*
    - required to transform one string into another,
    - possibly normalised (divided) by length of the maximum string
    - EditDistance (***Nkn***, ***Nikon***) = 0.4 (2/5)
      - ***Nkn*** → ***Nikn*** *(add i at 1)*
      - ***Nikn*** → ***Nikon*** *(add o at 3)*

Language pre-processing

- Often performed before applying string matching techniques
  - Stop word removal
    - common words, such as articles, prepositions, non-informative adverbs
  - Tokenization
    - extraction of terms from a document
    - text conflation and vocabulary reduction:
      - *from a document (one string) to a sequence of strings*
      - *"One quick brown fox" vs "One" "quick" "brown" "fox"*
  - Stemming
    - reducing words to their root forms
      - *"houses" vs "house"*

Tokenization

- Terms in ontologies are often made up of more than one word
  - e.g. class names
    - Wine, Wine grape, Wine Grape, Wine-Grape, WineGrape
  - e.g. property names
    - madeFromGrape, MadeFromGrape, Made-From-Grape, Made From Grape

- Tokenization
  - Extraction of the individual terms
    - removing punctuation and special characters
    - folding character case (e.g. all to lower case)

Stemming

- Reduces all morphological variants of a word to a single index term
  - stemming allows to recognise variations of the same word and treat them them as if they were the same
  - detects mappings between concepts whose names include different forms of the same word
    - "*Mouse:Intestine_Epithelium*" vs "*NCI_Anat:Intestinal_Epithelium*"
  - Porter stemming algorithm (1980)
    - relies on a preconstructed suffix list with associated rules
      - e.g. if suffix=IZATION and prefix contains at least one vowel followed by a consonant, replace with suffix=IZE
      - "*intestine*" → "*intestinal*"

Wordnet and its senses

- WordNet is a lexical database for the English language. It is often used as a lexical ontology, although it was not developed for this purpose
  - it groups English words into sets of synonyms called ***synsets*** and provides short definitions and usage examples
  - records a number of relations among these synonym sets or their members, e.g. synonym…

- Both nouns and verbs are organised into hierarchies, defined through ***hypernym*** or ***IS-A*** relationships
  - All synsets are connected to other synsets by means of semantic relations, e.g.:
    - hypernyms: ***Feline*** is a hypernym of ***Lion***
    - hyponyms: ***Tiger*** is a hyponym of ***Feline***
    - coordinate terms: ***Tiger*** is a coordinate term of ***Lion***, and ***Lion*** is a coordinate term of ***Tiger***
    - meronym: ***Finger*** is a meronym of ***Hand***
    - holonym: ***Hand*** is a holonym of ***Arm***

Linguistic techniques using Wordnet

- A subClassOf B if A is a hyponym of B
  - ***Tiger*** is a hyponym of ***Feline***
- A hasPart B if A is a holonym of B
  - ***Country*** is a holonym of ***Continent***
- A ≡ B if A is a synonym of B
  - ***Quantity*** is a synonym of ***Amount***
- A disjoint with B if A is an antonym (opposite) of B
  - or A and B are siblings
  - ***Tiger*** is disjoint with ***Cat***

Linguistic techniques gloss-based

- WordNet gloss comparison
  - Glosses model relations obtained from references between synsets by computing semantic relatedness.
    - “*brother*”, “*sibling*”, “*family*”
  - The similarity value increases given the increase in the number of the same words occurring in both input glosses
  - The equivalence relation is returned if the resulting similarity value is above a given threshold
    - “*Maltese dog is a breed of toy dogs having a long straight silky white coat*”
    - “*Afghan hound is a tall graceful breed of hound with a long silky coat*”
  -

Structural techniques: hierarchy comparison

- Compare the structure of entities in ontologies instead or in addition to comparing the labels of the entities *(Euzenat and Shvaiko, 2012)*
  - This looks at the internal structure of an entity
  - its name and annotations,
  - its properties
  - the properties whose value is a data type (OWL ontologies)
  - the comparison of an entity with other related entities (OWL ontologies)

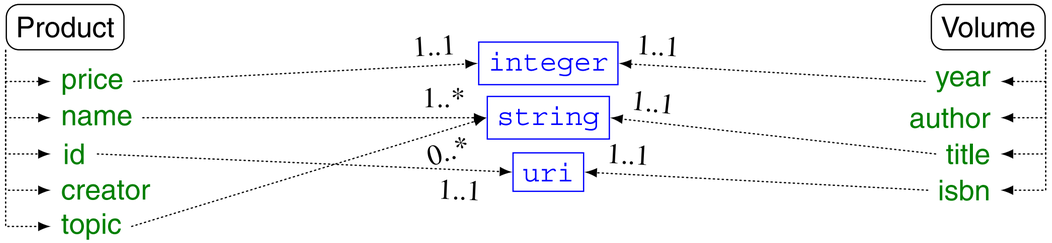
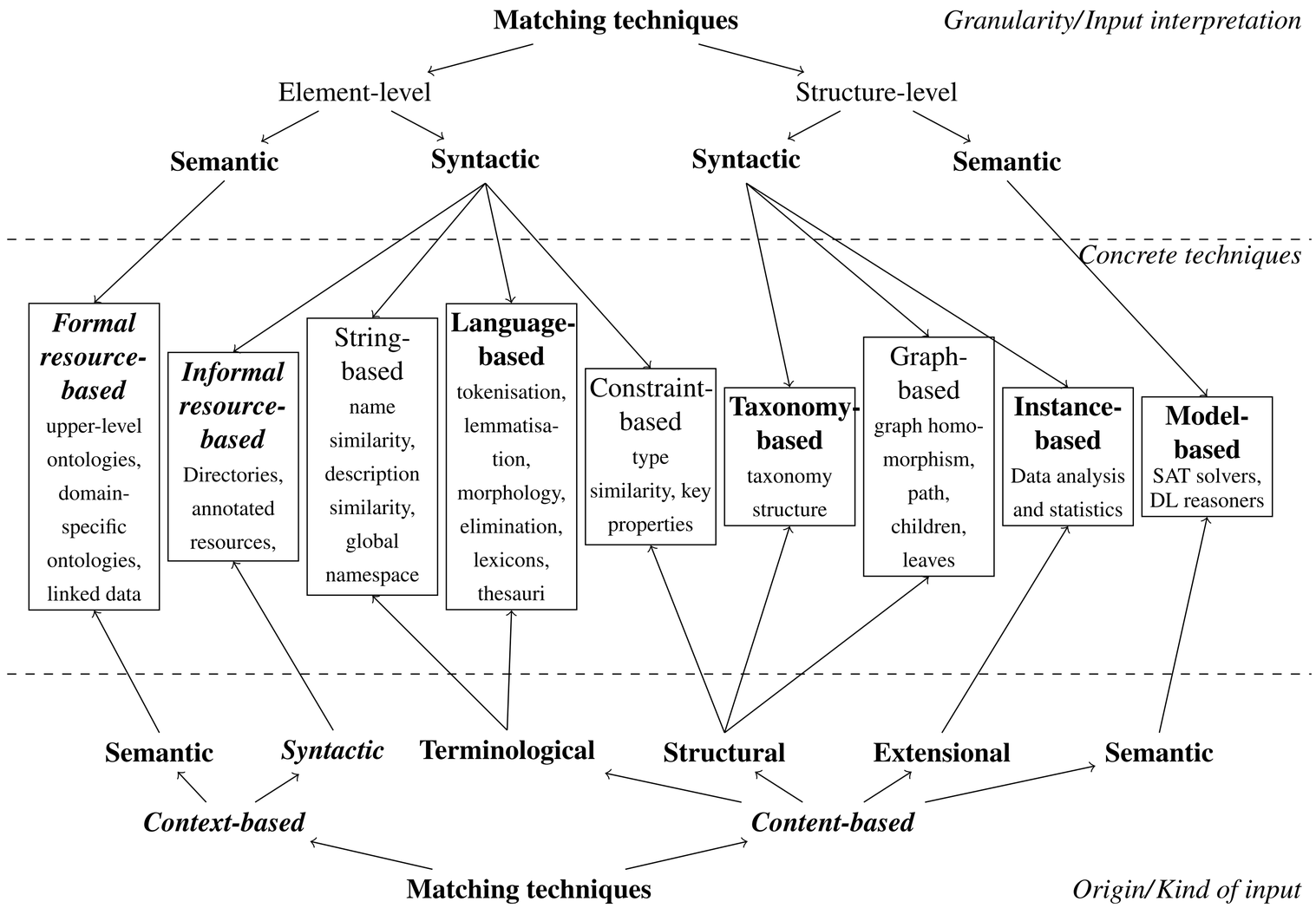


diagram by J. Euzenat

Classification of matching techniques



Challenges in using alignments

- Does the problem affect the way we align?
  - If we are merging ontologies, we could *align the whole ontology*
    - What about FRANKENSTEIN ontologies?
    - Quality vs Quantity
  - If we are using services or sensors, we should only align the necessary concepts
    - Is the alignment fit for purpose?
  - Fragments of the ontological space may be *confidential*, or *commercially sensitive*.
    - Disclosure or exposure is problematic

Challenges in using alignments

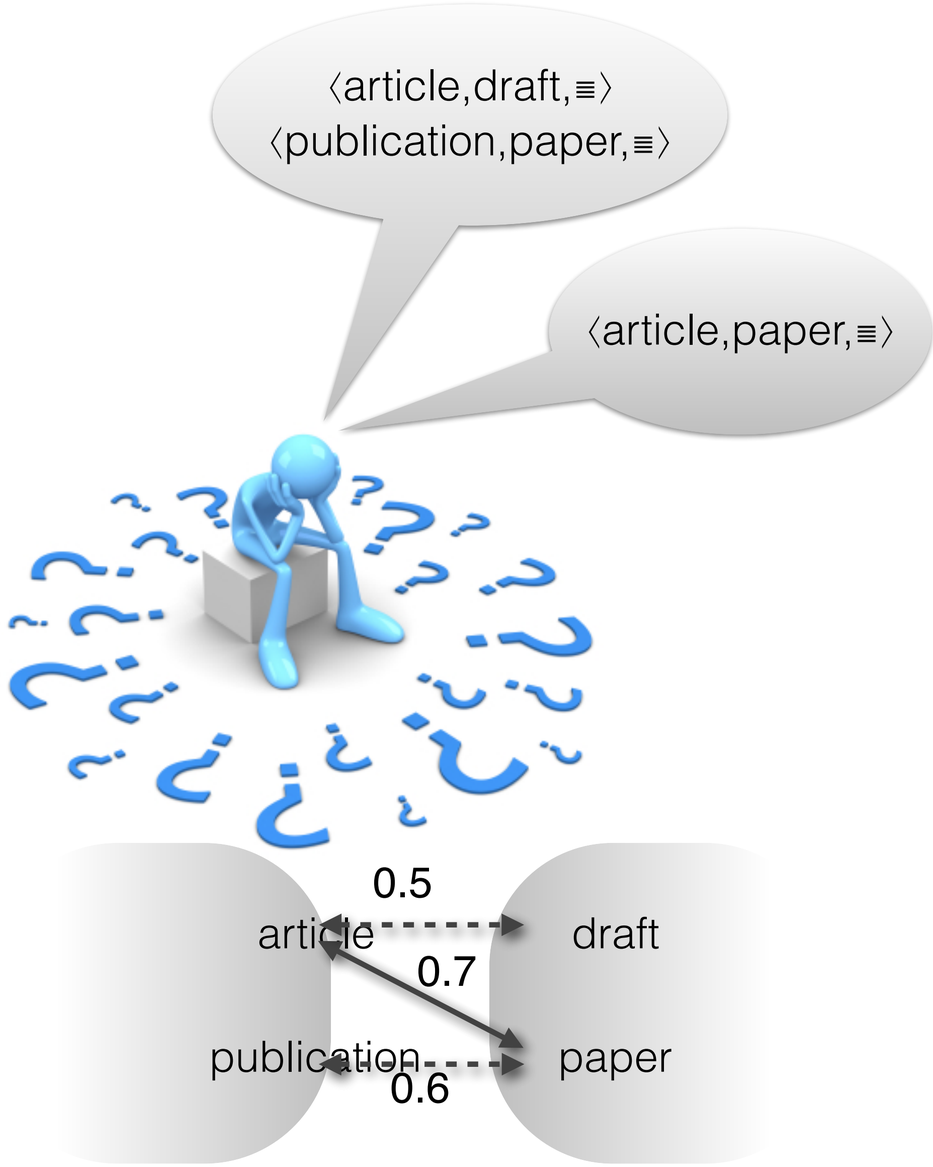
- Embarrassment of riches
  - What if we know many alignments?
    - Some better suited for certain tasks than others
    - Superset of different mappings, resulting in greater coverage
  - However, **aggregating alignments** may cause problems
    - Ambiguous mappings
    - Erroneous mappings

Challenges in using alignments

- Dearth of Knowledge
  - What if have no prior alignments?
    - Need to understand how to align the ontologies
    - Easy if there are ***existing services***, and the ontologies are ***public***
  - What if some of the ontological knowledge is **private**?
    - Disclosure of the full ontology no longer possible
    - Need to collaborate / negotiate to establish the mappings

Picking the right mappings

- Quality vs Quantity
  - Do we maximise coverage
    - Preferable when merging the whole ontology
  - Do we find the “best” mappings
    - Preferable when aligning specific signatures
- Interpretation of the output
  - Approximate vs exact
  - Graded vs absolute confidence
- Performance varies
  - Semi automatic alignment



Alignment evaluation

- Three types of alignment evaluation approaches:
  - Competence benchmarks: allow users to measure the extent of competence (and performance) of a particular system with regard to a set of well defined tasks
    - Usually, tasks are designed to isolate particular characteristics
    - E.g. set of tests designed
  - Comparative evaluation: compare the results of various systems or several versions of the same system on a common task.
    - The rules and the evaluation criteria MUST be clearly specified
    - Blind or nearly blind tests
  - Application-specific evaluation: compare the results of various systems wrt the output of a particular application instead of considering the alignments in isolation
    - Usually, tasks are designed to isolate particular characteristics

OAEI

- Ontology Alignment Evaluation Initiative defines a set of benchmark dataset to be used to evaluate ontology alignment approaches.

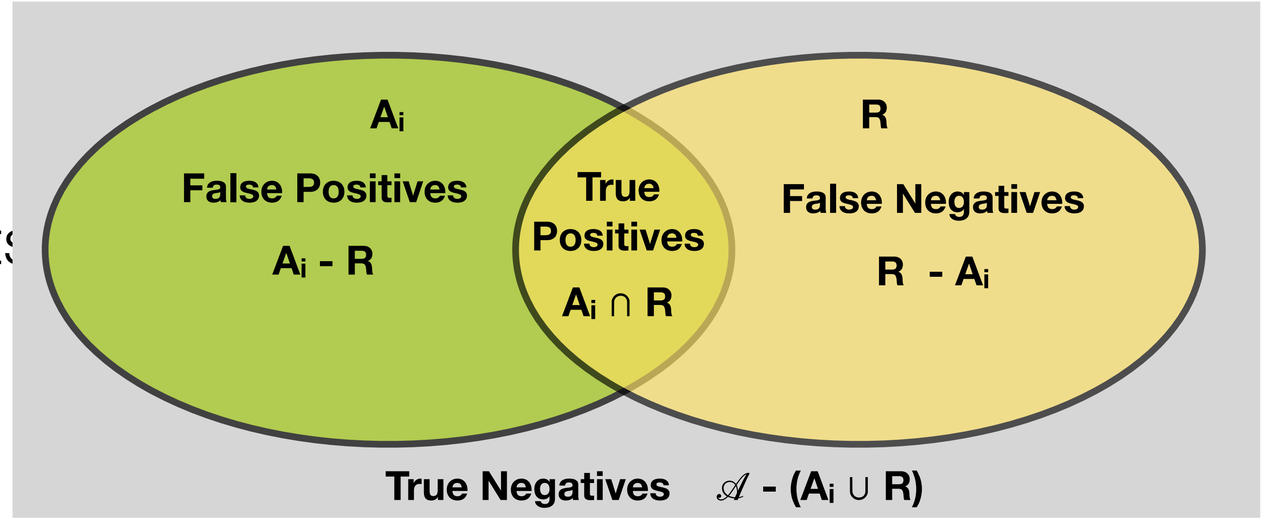
- Annual challenge, started in 2004

- Its aims are
  - *assessing strengths and weaknesses of alignment/matching systems;*
  - *comparing performance of techniques*
  - *increase communication among algorithm developers;*
  - *improve evaluation techniques*
  - *helping improving the work on ontology alignment/matching*

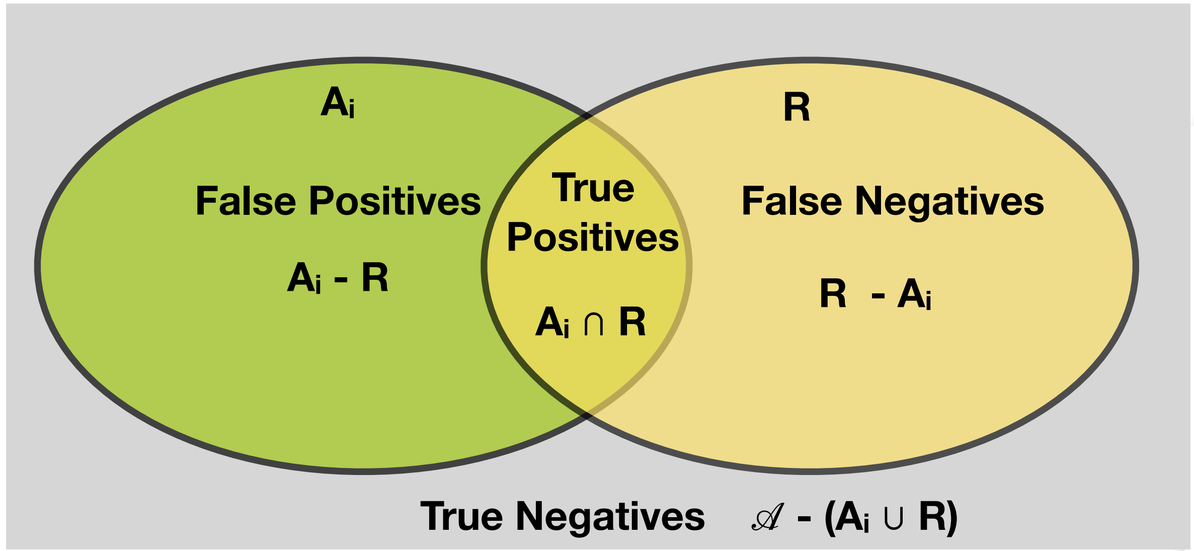
Compliance Measures: Precision and Recall

- Compliance measures evaluate the degree of compliance of a system with regard to some standard
  - typically a reference alignment ***R*** defined as part of the OAEI benchmark

- Given:
  - the set of all possible alignments
  - an alignment to assess $A_i \in \mathscr{A}$
  - and a reference alignment ***R***



Compliance Measures: Precision and Recall



Precision

$$\mathrm{Prec}(\mathrm{A_i}, \mathrm{R}) = \frac{|\mathrm{A_i} \cap R|}{|\mathrm{A_i}|}$$

Recall

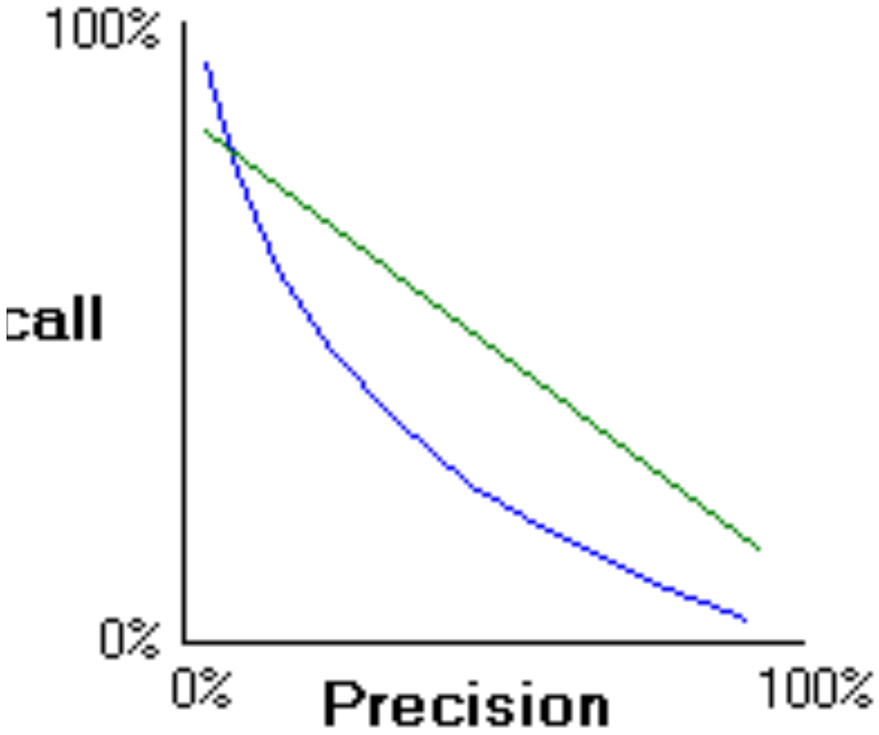
$$\mathrm{Rec}(\mathrm{A_i}, \mathrm{R}) = \frac{|\mathrm{A_i} \cap R|}{|\mathrm{R}|}$$

F-measure

$$\mathrm{F}_{mesure(\mathrm{A_i},\mathrm{R})} = 2 \times \frac{\mathrm{Prec}(\mathrm{A_i}, \mathrm{R}) \times Rec(\mathrm{A_i}, \mathrm{R})}{\mathrm{Prec}(\mathrm{A_i}, \mathrm{R}) + Rec(\mathrm{A_i}, \mathrm{R})}$$

Precision and recall behaviour

- While the exact slope of the curve may vary between systems, the general inverse relationship between recall and precision remains.
- Why is there an inverse relationship? Much of this relationship has to do with language:
  - If the goal of a search is comprehensive retrieval, then the searcher must include synonyms, related terms, broad or general terms, etc. for each concept. They using Boolean operators to combine terms. ISecondary concepts may be omitted. ***As a consequence of these decisions, precision will suffer.***
  - Synonyms may not be exact, so the probability of retrieving irrelevant material increases. Broader terms may result in the retrieval of material which does not discuss the narrower search topic. Using Boolean operators may increase the probability that the terms won't be in context. ***This might cause low recall!***



Precision, recall, F-measure

- ***Precision***: fraction of relevant correspondences (| Ai ∩ R |) amongst all the correspondences generated (| Ai |)

- ***Recall***: fraction of relevant correspondences produced by an alignment system (| Ai ∩ R |) over the total number of relevant correspondences
  - those in the reference alignment

- ***F-measure***: is the harmonic mean of precision and recall
  - score reaches its best value at 1 (perfect precision and recall) and worst at 0.

- Precision and recall should not be considered in isolation,
  - as considering just one out of precision and recall can lead to extreme but unhelpful solutions.
    - A system that returns every correspondence indiscriminately has 100% recall;
      - ***Recall without precision will provide too many results.***
    - A system that returns only a single correct correspondence is 100% precise.
      - ***However, only precision will limit the number of results (e.g. finding only one of the possible correct results will yield a precision of 100%).***

Summary

- The need of knowledge integration and sharing
  - support for interoperability
- Ontology alignment
  - Definition
  - Techniques
  - Evaluation metrics