UNIVERSITY
OF TRENTO - Italy

Linguistic and Knowledge Resources

Vincenzo Maltese
University of Trento

LDKR course 2014

Roadmap

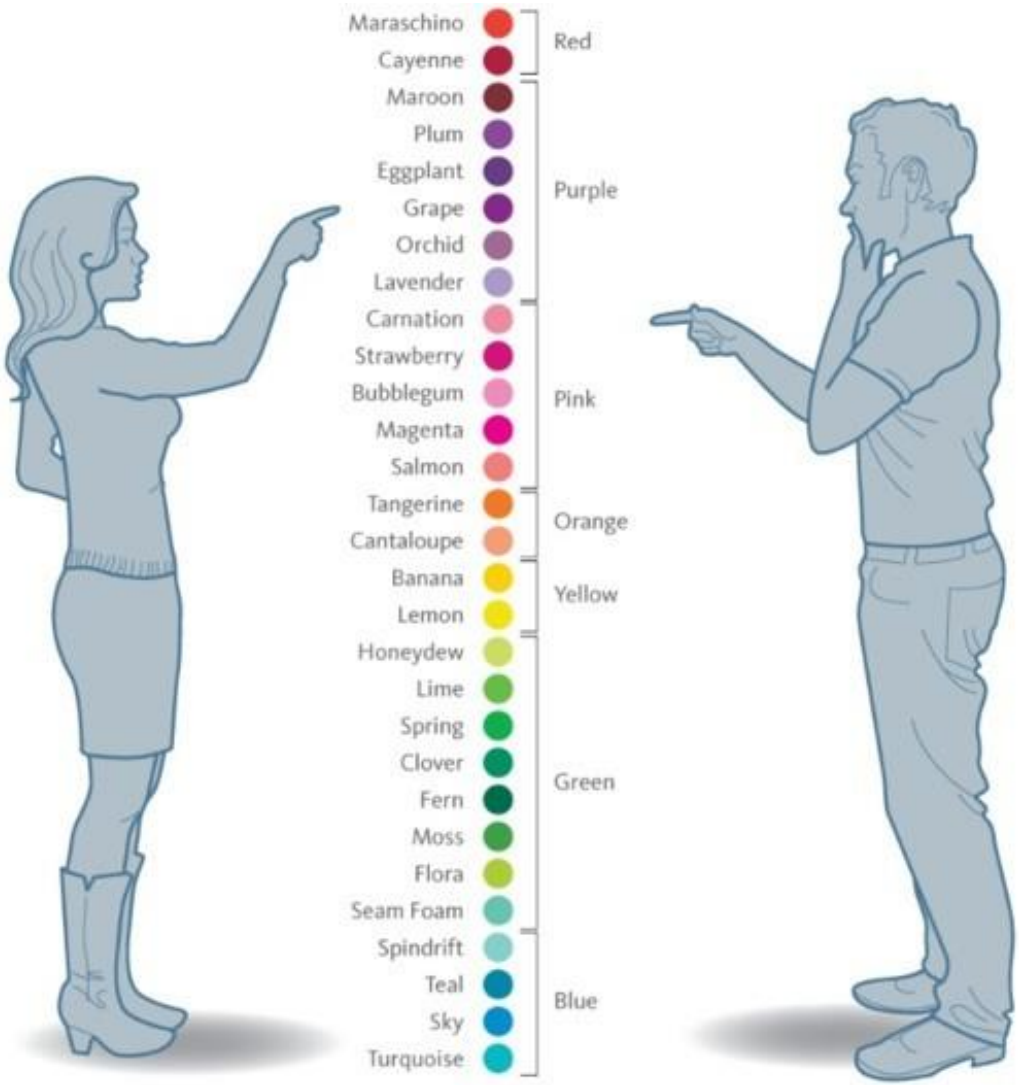
- **Introduction**
- **Linguistic resources**
- **Knowledge resources**
- **Capturing diversity with the UKC and Entitypedia**
- **The DERA methodology**

Introduction

Roadmap

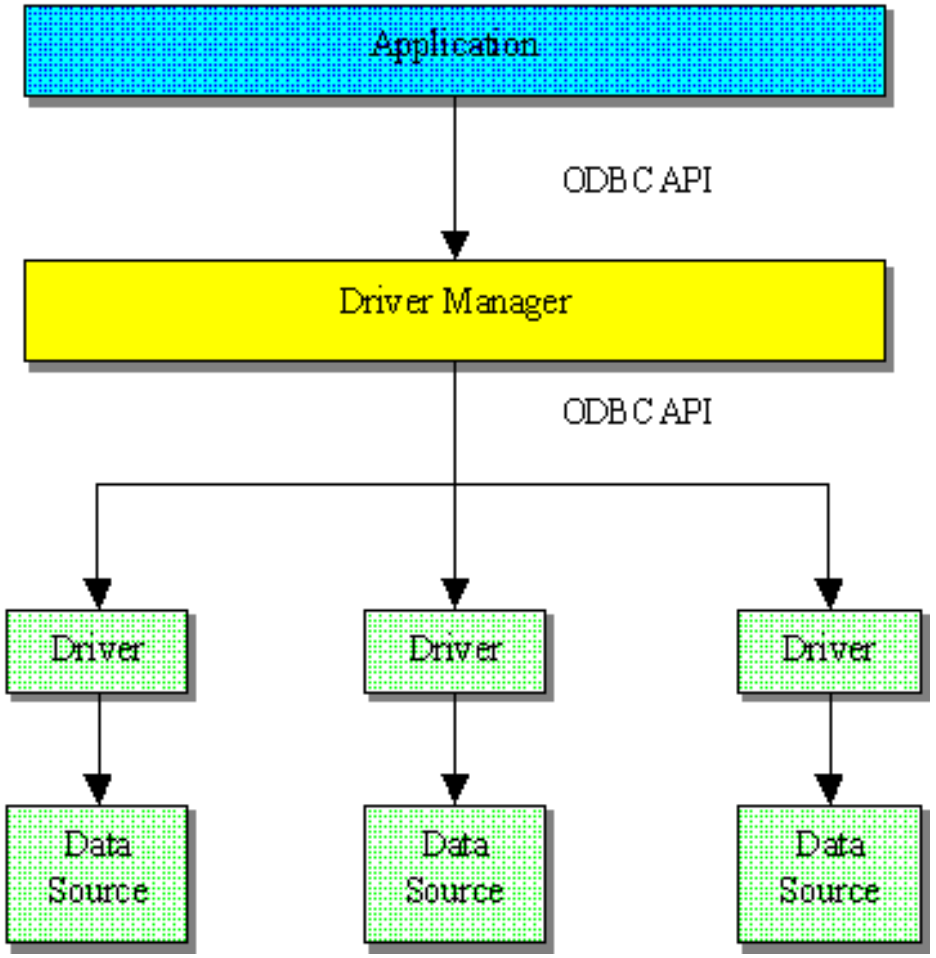
- **Problem:The semantic heterogeneity problem**
- **Solution: Current approaches to interoperability**
- **Ontologies**
- **Linguistic and knowledge resources: what and why**
- **Exercises**

The semantic heterogeneity problem



The difficulty of establishing a certain level of connectivity between people, software agents or IT systems [Uschold & Gruninger, 2004] at the purpose of enabling each of the parties to appropriately *understand* the exchanged information [Pollock, 2002]

Early solutions

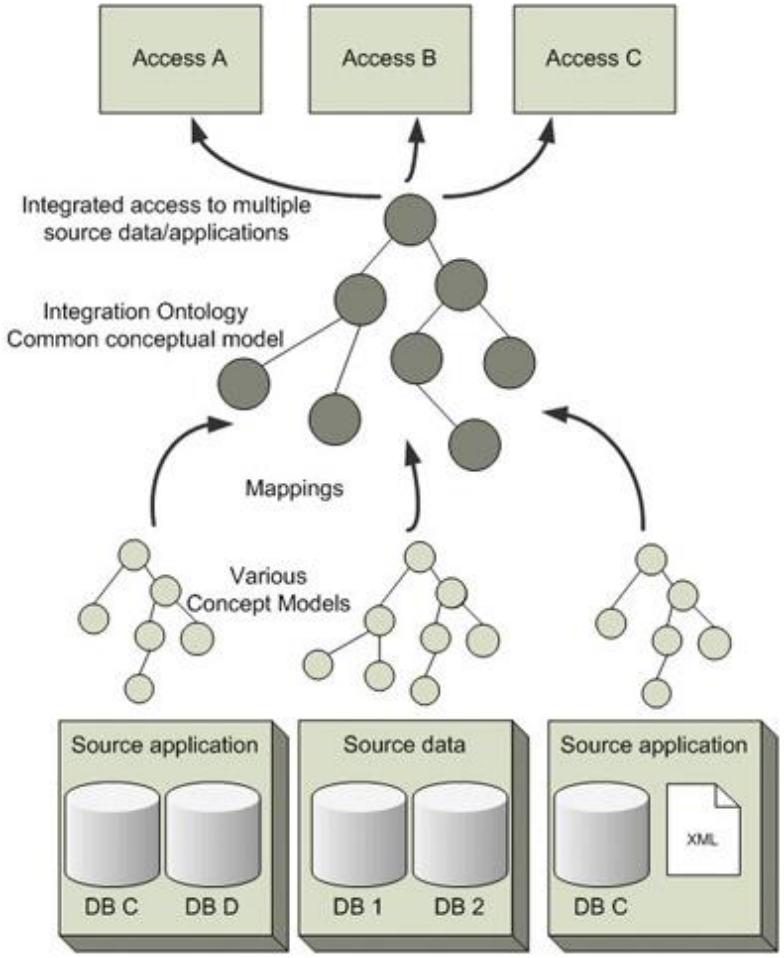


Physical connectivity relies on the presence of a stable communication channel between the parties, for instance ODBC data gateways and software adapters.

Syntactic connectivity is established by instituting a common vocabulary of terms to be used by the parties or by point-to-point bridges that translate messages written in one vocabulary in messages in the other vocabulary.

This rigidity and lack of explicit meaning causes **very high maintenance costs** (up to 95% of the overall ownership costs) as well as **integration failure** (up to 88% of the projects) **[Pollock, 2002]**

The semantic interoperability solution

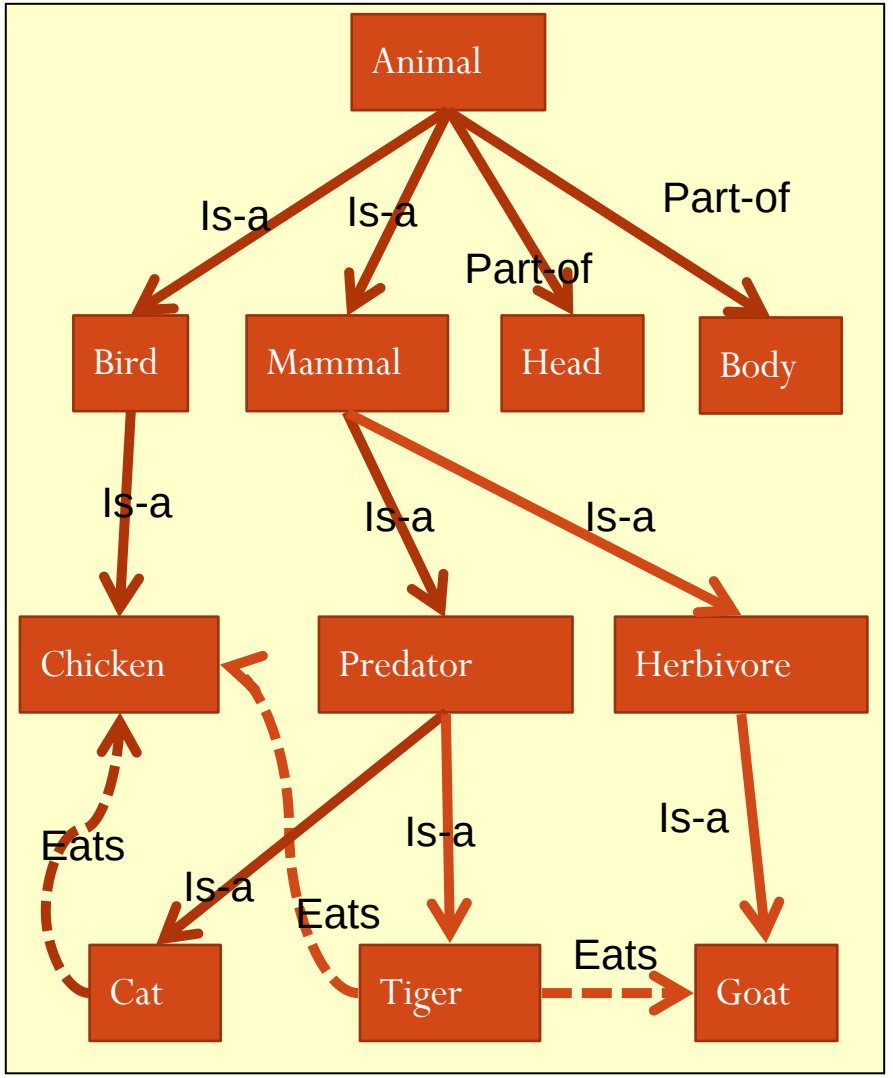


The solution in three points:

- ***Semantic mediation***: the usage of an ontology, providing a shared vocabulary of terms with explicit meaning.
- ***Semantic mapping***: using the ontology, the *establishment of a mapping* constituted by a set of correspondences between semantically similar data elements independently maintained by the parties.
- ***Context sensitivity***: the mapping has *contextual validity*, i.e. it has to be used by taking into account the conditions and the purposes for which it was generated.

Ontologies

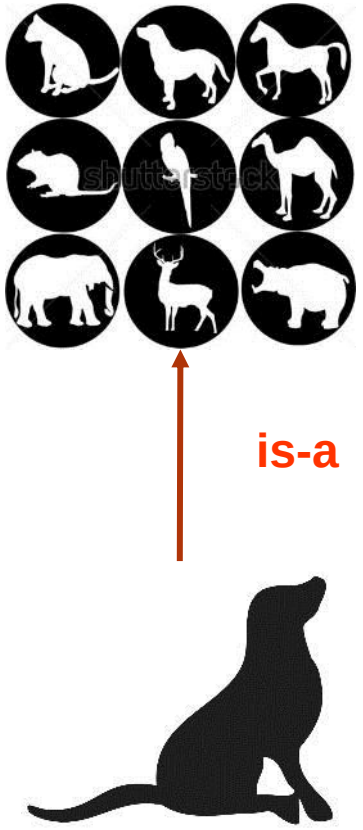
- An explicit specification of a shared conceptualization [Gruber, 1993]
- Directed graphs
- Nodes represent concepts
- Edges represent relations between concepts
- They provide a common (formal) terminology and understanding of a given domain of interest
- They allow for automation (logical inference), support reuse and favor interoperability across applications and people.



Concepts and relations (I)

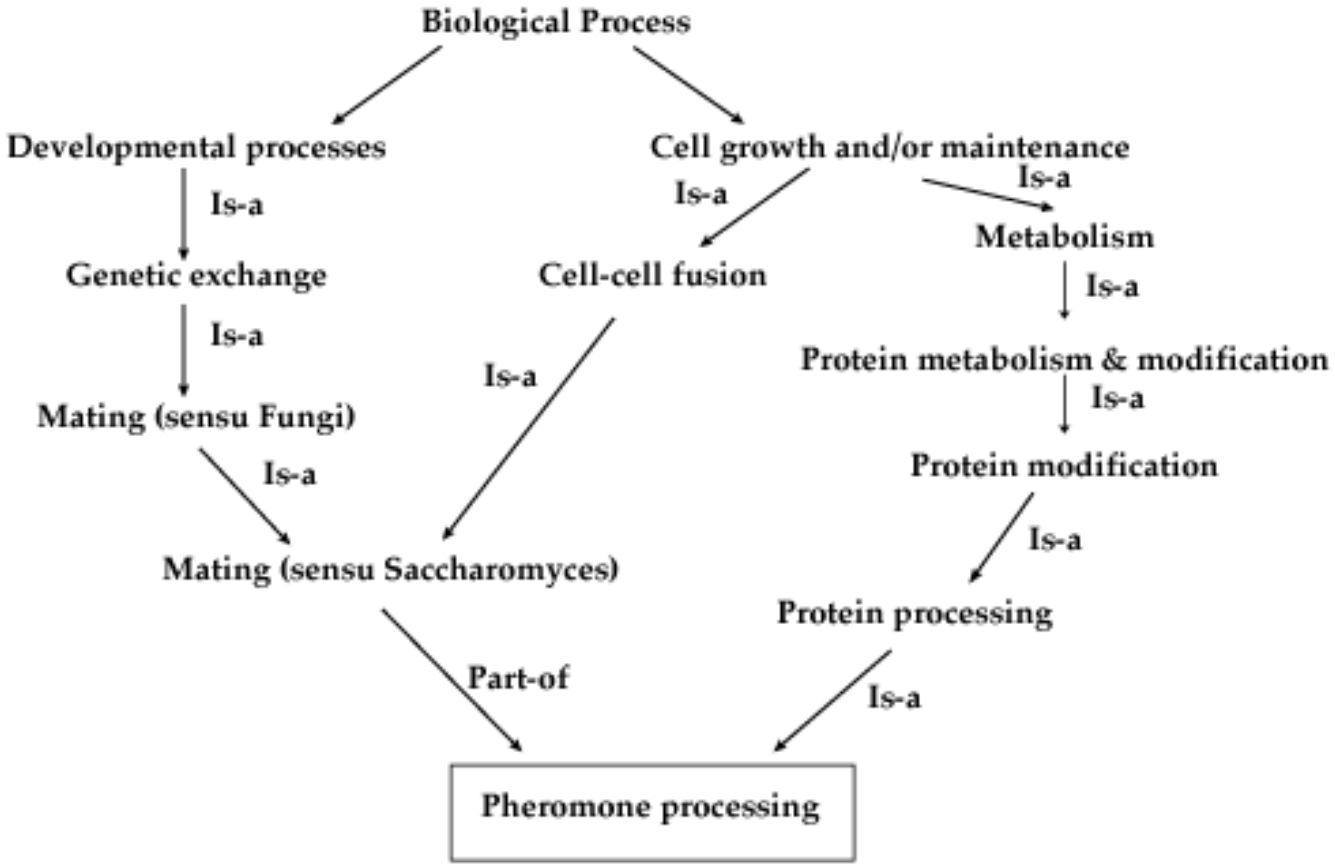
- CONCEPT: it represents a set of objects or individuals
- EXTENSION: the set of individuals is called the concept extension or the concept interpretation
- RELATION: a link from the source concept to the target concept

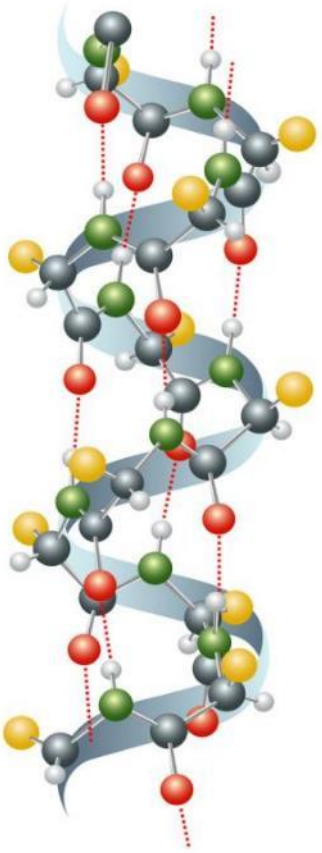
- Concepts are often lexically defined, i.e. they have natural language labels which are used to describe the concept extensions, often with an additional description or gloss



Concepts and relations (II)

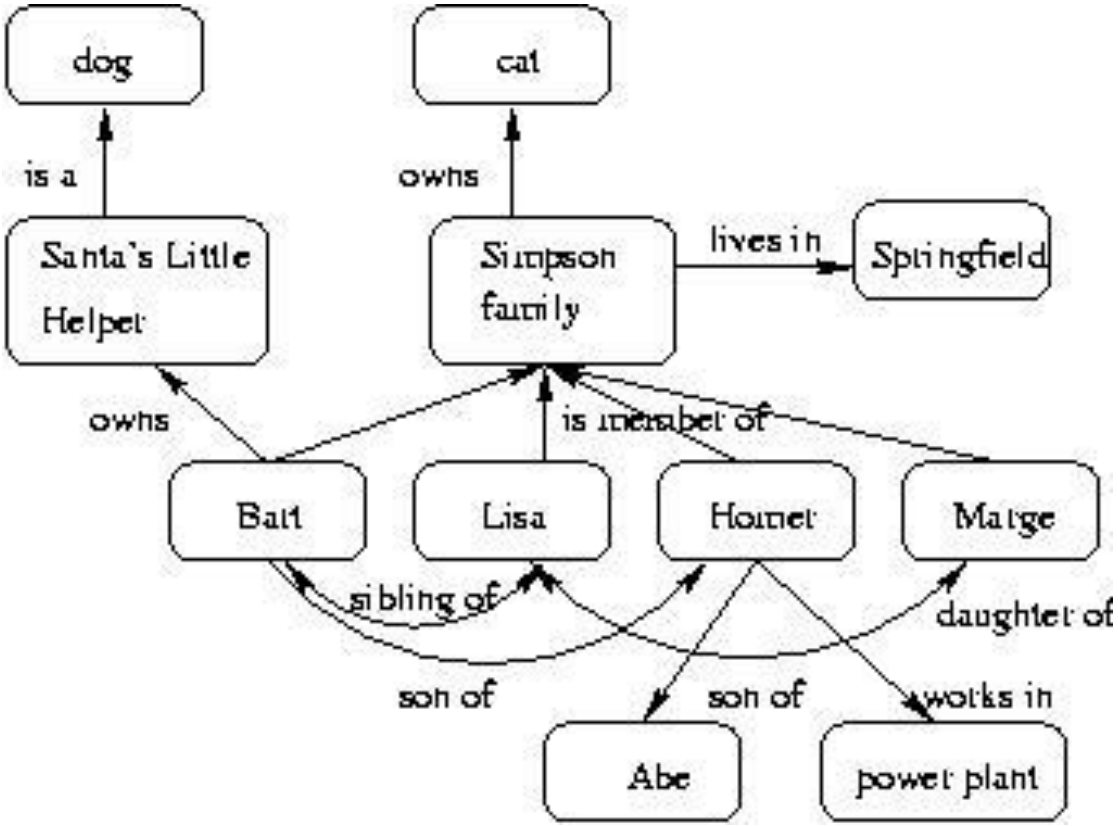
The backbone structure of an ontology graph is a taxonomy in which the ontological relations are genus-species (is-a and instance-of) and whole-part (part-of).

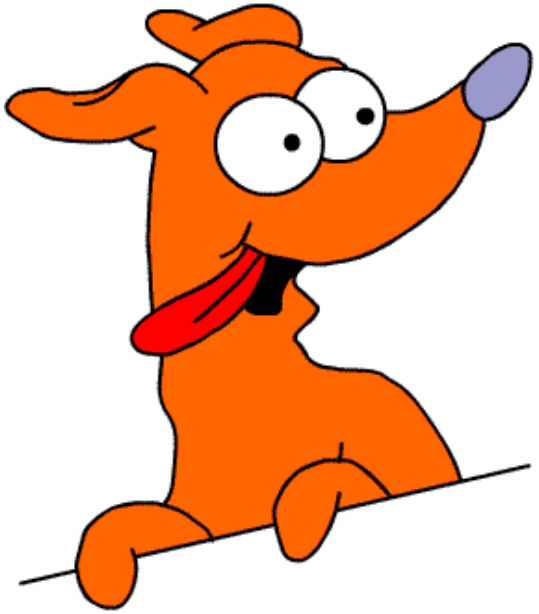




Concepts and relations (III)

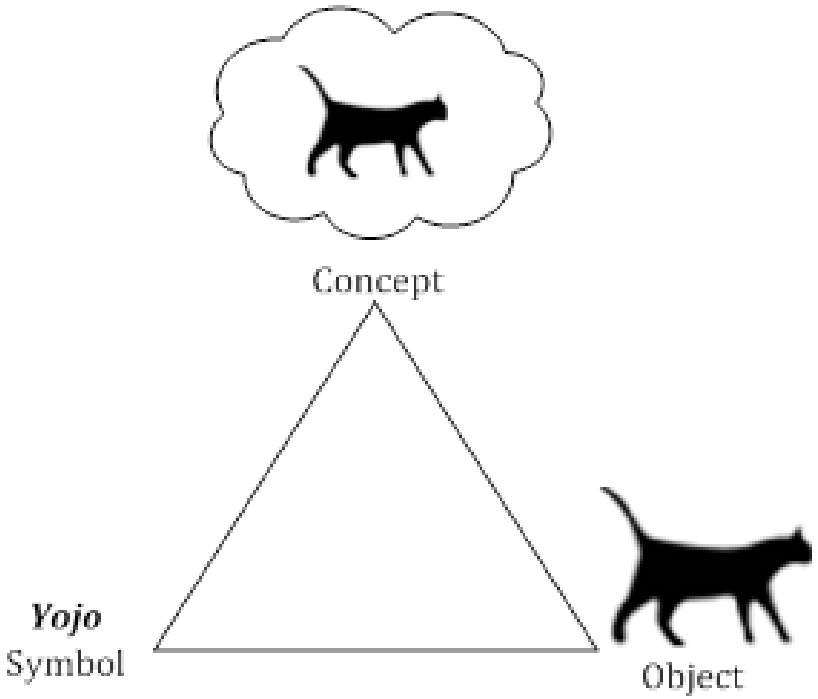
The remaining structure of the graph supplies auxiliary information about the modeled domain and may include relations of any kind.





Conceptualization

An abstract model of how people theorize (part of) the world in terms of basic cognitive units called *concepts*. Concepts represent the intention, i.e. the set of properties that distinguish the concept from others, and summarize the extension, i.e. the set of objects having such properties.



Explicit specification

the abstract model is made explicit by providing *names* and *definitions* for the concepts, i.e. the name and the definition of the concept provide a specification of its meaning in relation with other concepts.

DOG

a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds

Formal specification

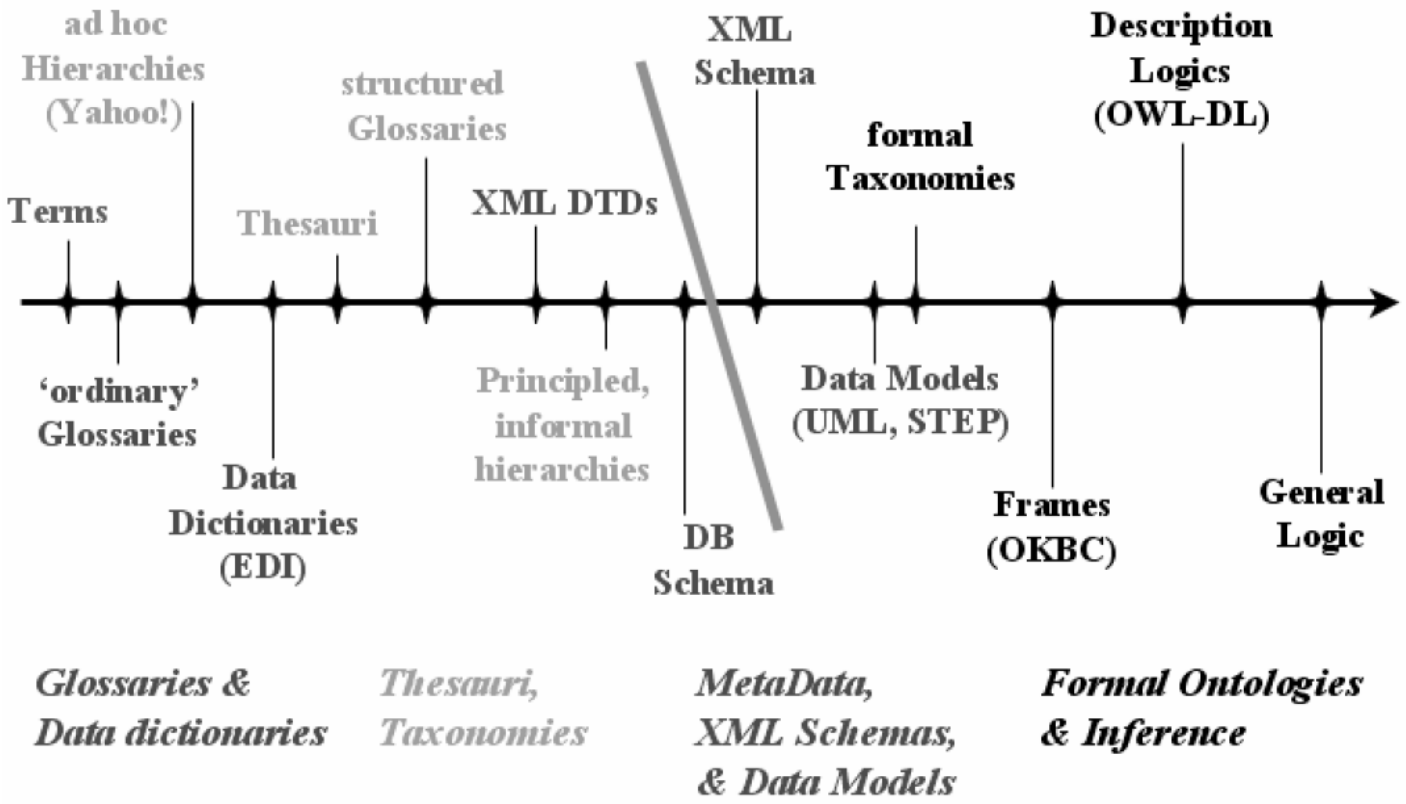
The abstract model is formal when it is written in a language with *formal syntax and formal semantics*, i.e. in a logic-based language.

Shared conceptualization

It captures knowledge which is common to a community of people and therefore represents concretely the level of agreement reached in that community.

Kinds of ontologies

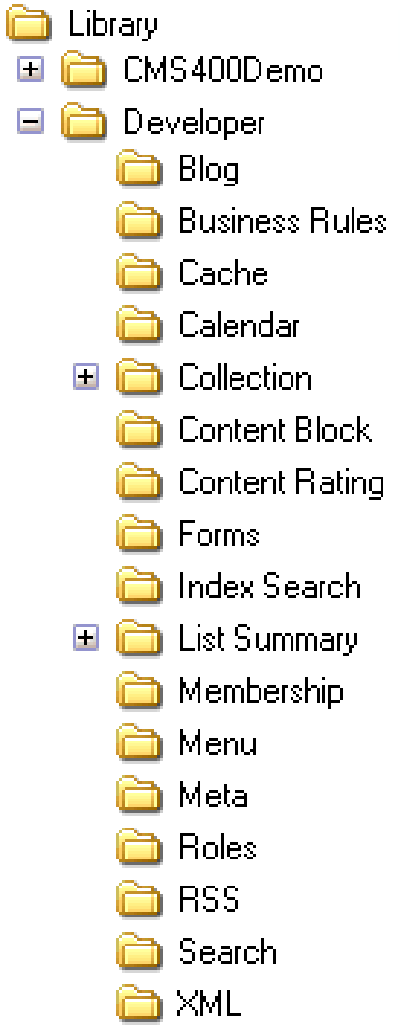
- Ontologies differ according to the purpose, the syntax and the semantics
- There is also a tension between expressivity and effectiveness



[Uschold and Gruninger, 2004]

Informal ontologies

- User classifications
- Folders in a file system
- Web directories
- Business catalogs

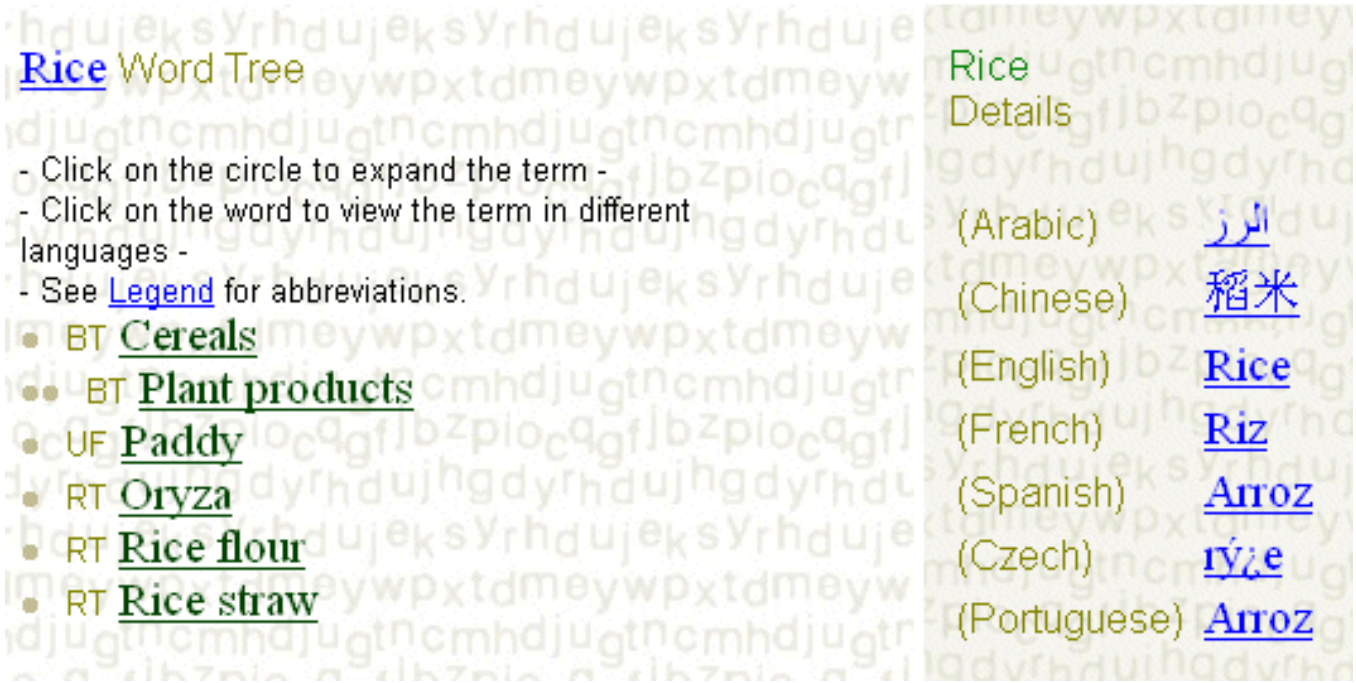


Semi-formal ontologies (I)

- Knowledge Organization Systems: Library classifications, Thesauri

Semi-formal ontologies (II)

In Knowledge Organization Systems (KOS) there are two main kinds of relations: hierarchical (BT/NT) and associative (RT) relations.



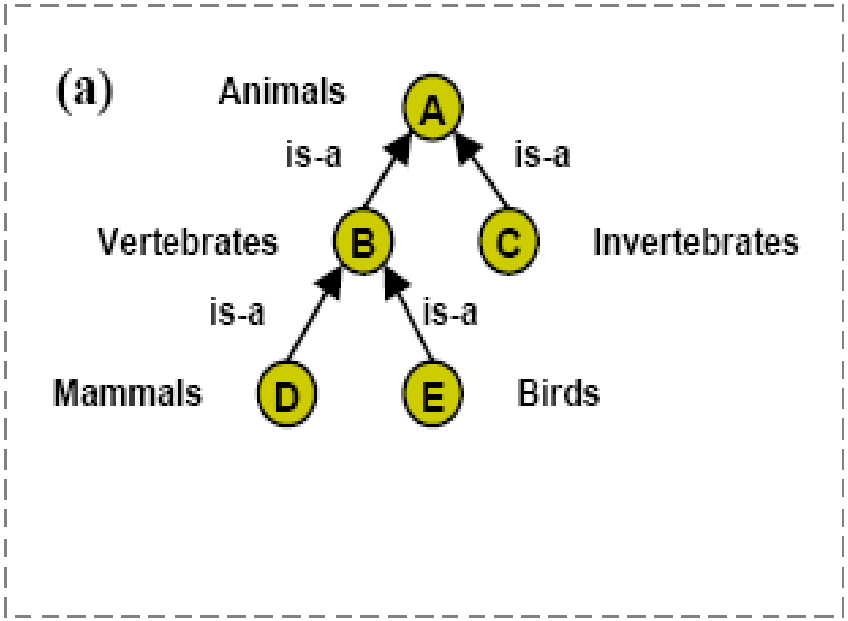
Formal ontologies

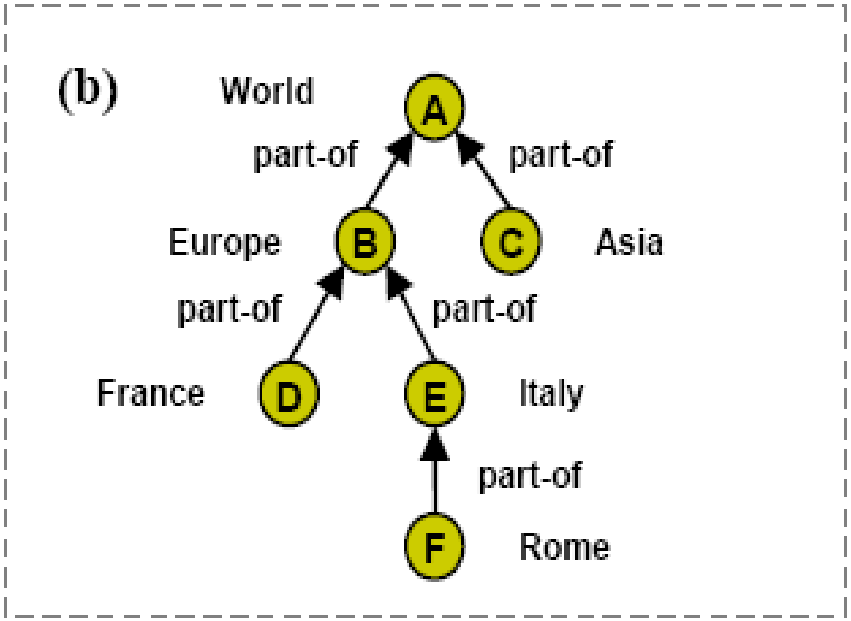
Formal ontologies are expressed into a formal logic language (in syntax and semantics) and represented via formal specifications (e.g. OWL)

```
1  <?xml version = '1.0' encoding = 'ISO-8859-1' ?>
2  <rdf:RDF

3    <owl:Ontology rdf:about = 'Museum'/>

4    <owl:Class rdf:ID = 'Painter'>
5      <rdfs:label>Painter</rdfs:label>
6      <rdfs:subClassOf rdf:resource = '#Artist'/>
7      <rdfs:subClassOf>
8        <owl:Restriction>
9          <owl:onProperty rdf:resource = '#Painter.paints'/>
10         <owl:minCardinality  rdf:datatype = '&xsd#nonNegativeInteger'>0</owl:minCardinality >
11       </owl:Restriction>
12     </rdfs:subClassOf>
13   </owl:Class>

14   <Painter rdf:ID = 'PabloPicasso'>
15     <Artist.lastName rdf:datatype = '&xsd#string'>Picasso</Artist.lastName>
16     <Artist.firstName rdf:datatype = '&xsd#string'>Pablo</Artist.firstName>
17     <Painter.paints rdf:resource = '#Guernica'/>
18     <Painter.paints rdf:resource = '#Autoportrait'/>
19   </Painter>

20 </rdf:RDF>
```

Descriptive ontologies [Giunchiglia et al., 2009]

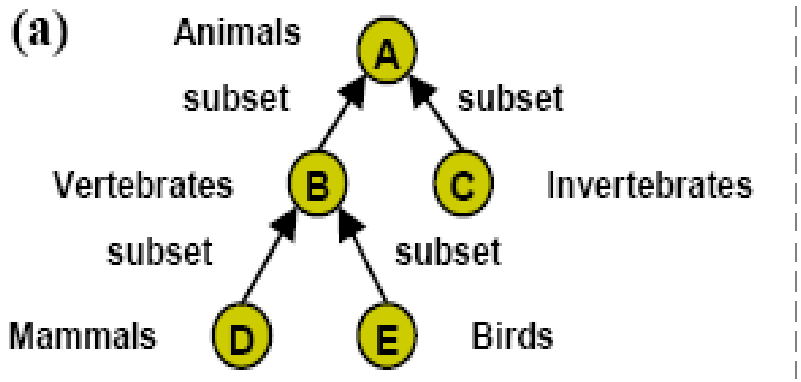
- Used to *describe* objects in a domain
- Real world semantics: the extension of a concept is the set of real world entities about the label of the concept
- We need to distinguish between classes (Animals) and individuals (Italy)
- Is-a relations are translated into DL subsumption (⊑)

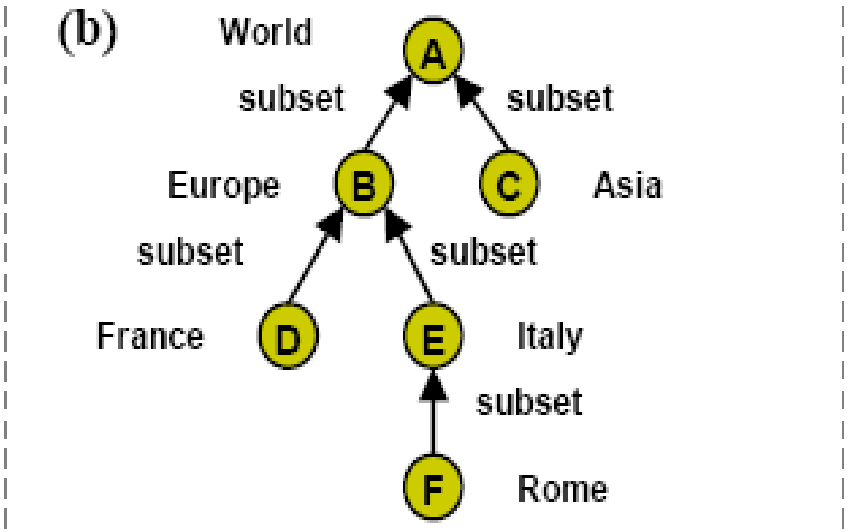




Classification ontologies [Giunchiglia et al., 2009]

- Used to *categorize* objects
- Classification semantics: the extension of a concept is the set of documents about the entities or individual objects described by the label of the concept. The semantics of the links is “subset”.
- No distinction between classes (Animals) and individuals (Italy)
- Subset relations are translated into DL subsumption ($\sqsubseteq$)





Converting ontologies

FROM DESCRIPTIVE TO CLASSIFICATION ONTOLOGY

- convert instances into classes
- convert instance-of, is-a and transitive part-of into NT/BT relations
- convert other relations into RT relations

FROM CLASSIFICATION TO DESCRIPTIVE ONTOLOGY

- each class is mapped to either a real world class or instance
- each NT/BT relation (assuming them to be transitive) has to be converted to either an instance-of, is-a or transitive part-of
- each RT relation has to be codified into an appropriate real world associative relation

The translation process can be easily automated

However, with the translation we have a clear loss of information.

The translation process cannot be automated.

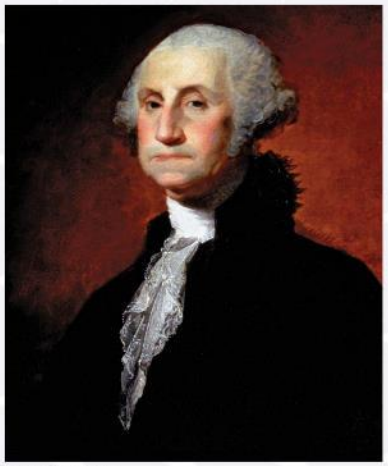
It needs significant manual work to reconstruct implicit information.

What a linguistic and knowledge resource is?

Main Entry:	**great**
Part of Speech:	*adjective*
Definition:	excellent, skillful
Synonyms:	able, absolute, aces, adept, admirable, adroit, awesome, bad*, best, brutal, cold*, complete, consummate, crack*, downright, dynamite, egregious, exceptional, expert, fab, fantastic, fine, first-class*, first-rate, good, heavy*, hellacious, marvelous, masterly, number one, out of sight, out of this world, out-and-out, perfect, positive, proficient, super-duper, surpassing, terrific, total, tough, transcendent, tremendous, unmitigated, unqualified, utter, wonderful
Antonyms:	ignorant, menial, poor, stupid, unskilled, weak

* = informal/non-formal usage

GEORGE WASHINGTON



1st President of the United States (1732–99; president 1789–97)

Vice President: John Adams (1735–1826)

Remembered as the Father of His Country, George Washington stands alone in American history. He was commander in chief of the Continental Army during the American Revolution, chairman of the convention that wrote the U.S. Constitution, and the first president of the United States. He led the people who transformed the United States from a British colony into a self-governing nation. His ideals of liberty and democracy set a standard for future presidents and for the entire country.

Early Life

The eldest child of Augustine and Mary Ball Washington, George Washington was born on Feb. 22, 1732, on the Wakefield plantation in Westmoreland County, Va. His father was a prosperous landowner who managed farms, businesses, and mines. Later the family moved to Ferry Farm on the Rappahannock River, opposite Fredericksburg, Va. Ferry Farm was the setting of George's boyhood adventures as described by Mason Locke Weems in his book *The Life and Memorable Actions of George Washington* (1800). In perhaps the most famous of these stories, George chops down a cherry tree with a hatchet and later admits it to his father, stating that he cannot tell a lie. Today the stories are thought to be fiction rather than fact.

After his father's death in 1743, George lived with his half brother Augustine at Wakefield and attended Henry William's school, one of the best schools in Virginia. By age 15 George was skilled in mathematics and mapmaking and had developed an interest in surveying.

In 1748 George went to live with his other half brother, Lawrence, at an estate on the Potomac River named Mount Vernon. There he met a wealthy landowner who hired him to help survey his holdings in Virginia. George excelled at his new profession. In July 1749 he was appointed surveyor of Culpeper County, his first public office. When Lawrence died in 1752, Washington inherited Mount Vernon, thus becoming a landowner.

French and Indian War

Inspired by his brother's experiences in the British Navy, Washington pursued a military career. He was made a lieutenant colonel in 1754, as tensions rose between the British and the French over control of the Ohio River valley. Washington was sent with nearly 200 troops to take possession of Fort Duquesne, at the site of present-day Pittsburgh, Pa., but he found that the French had claimed it first. Washington built Fort Necessity nearby. From there, his troops and their Native American allies ambushed a French scouting party, killing the commander and nine others. The French and Indian War had begun.

First Lady Martha Washington (1731–1802)

Martha Washington set many of the standards for the proper behavior and treatment of the president's wife. Born Martha Dandridge, she grew up among the wealthy plantation families of eastern Virginia. Her education, traditional for young women of her class and time, emphasized domestic skills and the arts. At age 18 she married Daniel Parke Custis, an heir to a neighboring plantation. His death in 1757 made her wealthy. Two years later she married George Washington and moved with her children to Mount Vernon, where she became known for her graciousness and hospitality. By nature a private person, Martha reluctantly assumed a public role after George became president. In doing so, she contributed to the eventual strength and influence of the position of first lady.

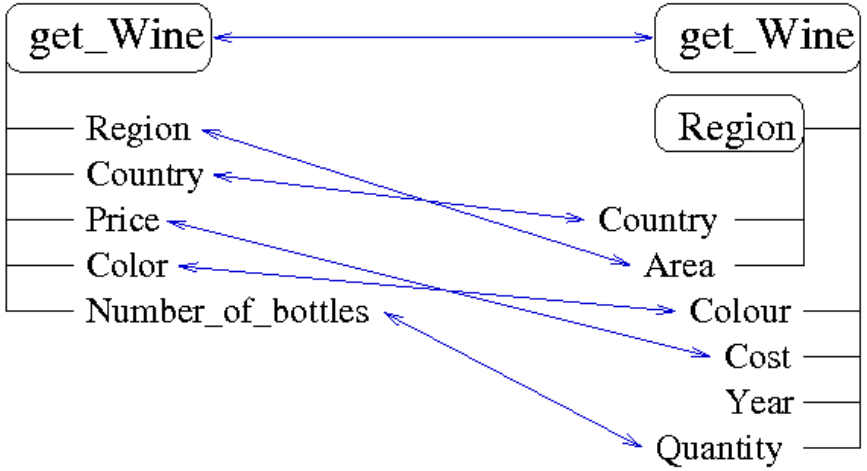
Why do we need linguistic and knowledge resources?

NLP

The banks of the river Nile

bank: sloping land (especially the slope beside a body of water)	river: a large natural stream of water (larger than a creek)	Nile: a major north-flowing river in northeastern Africa

SEMANTIC MATCHING



SEMANTIC SEARCH

SEARCH: automobile

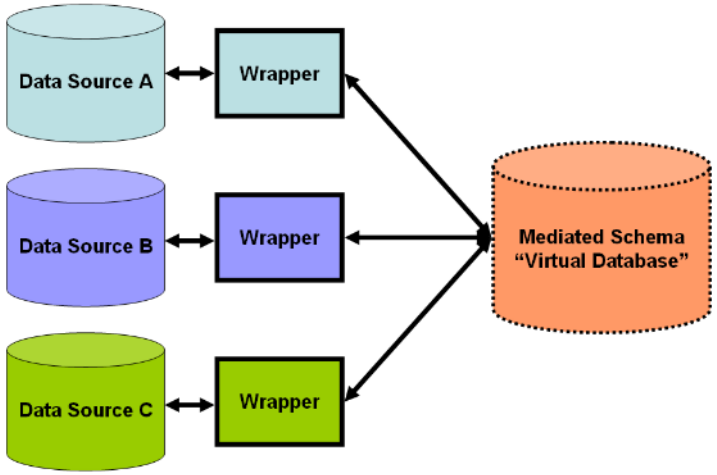
1957 Ferrari 625 TRC Spider

This two-of-a-kind classic Ferrari is lauded by historians as one of the prettiest Ferraris ever built. The 1957 Ferrari 625 TRC Spider is an absolutely stunning automobile, one as dashing in the garage as it is at 120 mph.

Back in the Saddle: Presenting our Porsche 911 (997) Carrera S Cabriolet

There's a reason the Porsche 911 is one of the most popular sports cars ever, and after a few minutes behind the wheel of one you'll understand why.

DATA INTEGRATION



Exercises

1. Is a ER diagram a formal ontology? Explain why yes or no.
2. Is a database schema a formal ontology? Explain why yes or no.
3. Create an ontology to describe your family in terms or general classes, relations between them and actual individuals
4. Identify in the web two thesauri in the agricultural domain
5. Identify in the web an OWL ontology
6. Identify a sub-tree in your file system and convert it into a descriptive ontology where each node label is given a definition

Linguistic resources

Roadmap

- **WordNet**
- **MultiWordNet**
- **Weaknesses of existing linguistic resources**
- **Exercises**

WordNet (1985)

WordNet Search - 3.1
- WordNet home page - Glossary - Help

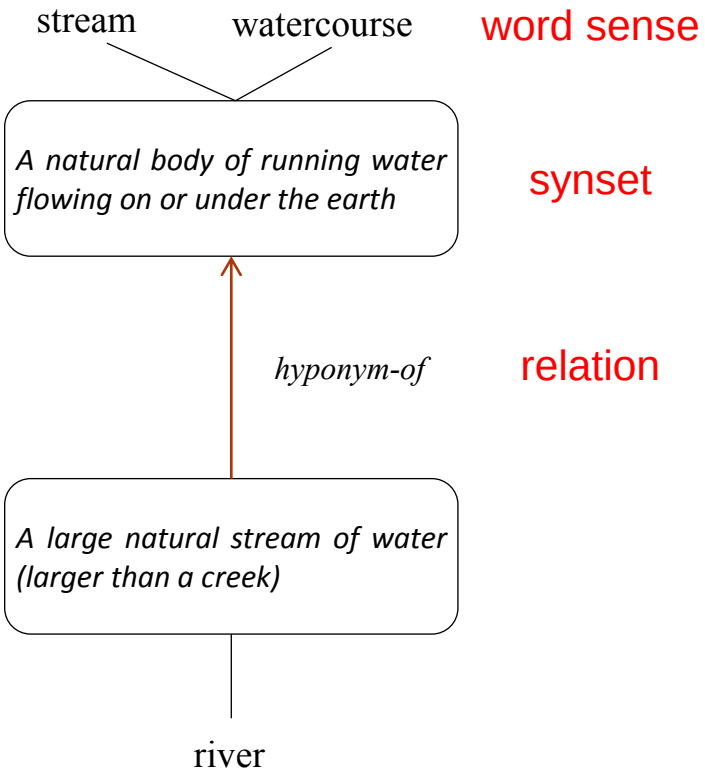
Word to search for: watercourse [Search WordNet]

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) **watercourse** (natural or artificial channel through which water flows)
- S: (n) stream, **watercourse** (a natural body of running water flowing on or under the earth)
  - ***direct hyponym*** / *full hyponym*
    - S: (n) branch (a stream or river connected to a larger one)
    - S: (n) brook, creek (a natural stream of water smaller than a river (and often a tributary of a river)) *"the creek dried up every summer"*
    - S: (n) headstream (a stream that forms the source of a river)
    - S: (n) river (a large natural stream of water (larger than a creek)) *"the river was navigable for 50 miles"*
    - S: (n) rivulet, rill, run, runnel, streamlet (a small stream)
    - S: (n) tidal river, tidewater river, tidal stream, tidewater stream (a stream in which the effects of the tide extend far upstream)
  - *part meronym*
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*
- S: (n) **watercourse**, waterway (a conduit through which water flows)



Words

- Words are the basic constituents of a language
- WordNet focuses on **lemmas,** i.e. the canonical form of a set of words in a language.

  In English, for example, *run*, *runs*, *ran* and *running* are forms of the same lexeme, with the verb *run* as the lemma.

- WordNet also accounts for **exceptional forms**. For nouns, they are usually the irregular plural forms, for adjectives and adverbs irregular superlatives, for verbs irregular conjugations.

  For instance, the noun *wives* is an exceptional form of the noun *wife*.

Senses and synsets

- A (word) **sense** is a word in a language (e.g. English) having a distinct meaning.
- Senses for each word are ranked.
- Words having same sense are grouped together into a **synset**.
- Each synset is associated a **part of speech** (POS) in the set {noun, adjective, verb, adverb} and a **gloss**.

For instance, in English the word *good*:

(noun) good : *an article for commerce*

(adjective) good : *having positive qualities*.

Lexical relations

- Lexical relations are between word senses.

- Synonymy is a symmetric relation connecting two senses of two different words with same POS and same meaning. WordNet implements synonymy through the notion of synset.

  stream and watercourse are synonym

- Antonym is a symmetric relation connecting two senses of two different words with same POS and opposite meaning.

  black is antonym of white.

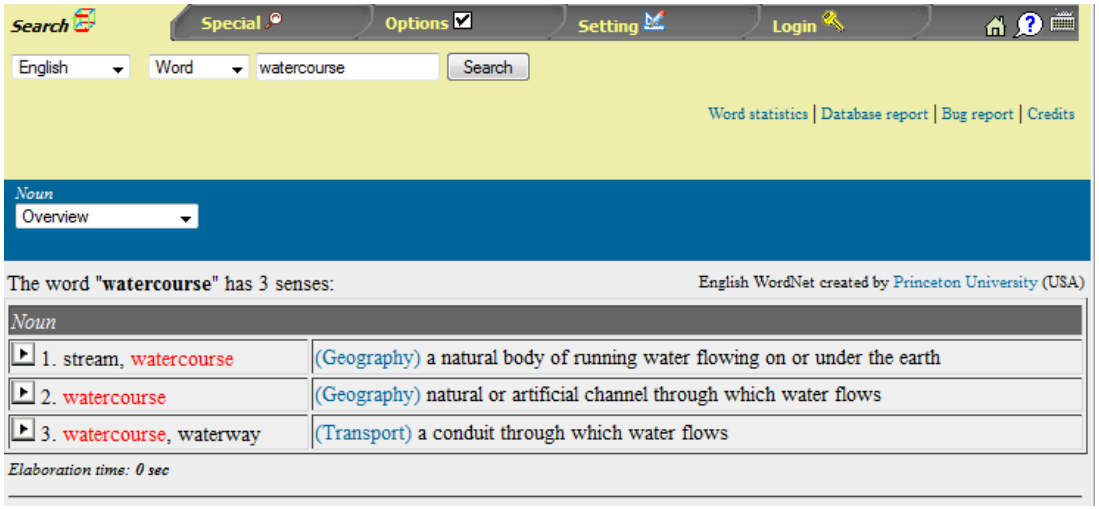
Semantic relations

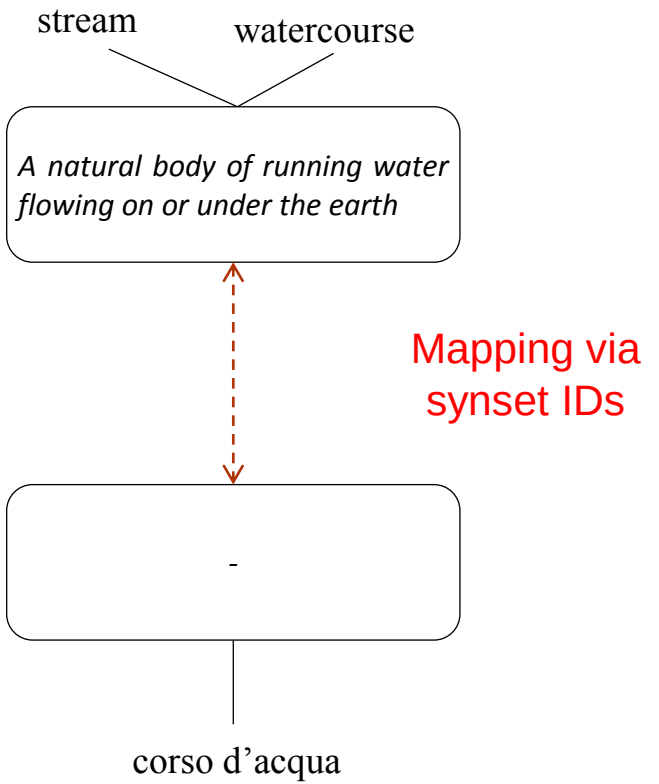
- Semantic relations are between synsets.
- Y is a **hypernym** of X (and X is hyponym of Y) if every X is a (kind of) Y

  canine is a hypernym of dog
- Y is a **meronym** of X (and X is holonym of Y) if Y is a part of X

  window is a meronym of building

MultiWordNet (2002)





Mapping via synset IDs

Strengths

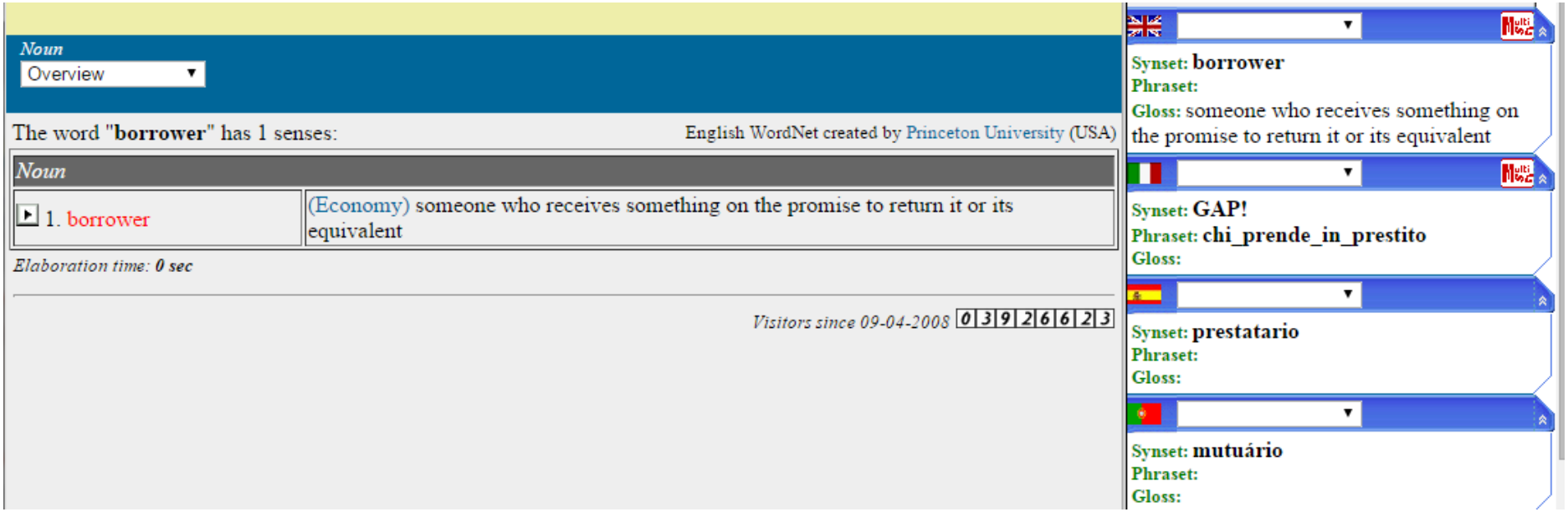
- Mapping with 6 languages
- Lexical GAPs can be defined

Weaknesses

- Only a partial coverage
- A few glosses available
- Biased towards English

Lexical GAPs and phrasets

The fact that a language (e.g. English) expresses in a lexical unit what the other language (e.g. Italian) expresses with a free combination of words (e.g. borrower = chi prende in prestito)



Problems with WordNet-like resources (I)

- S: (n) **educational institution** (an institution dedicated to education)
  - S: (n) school (an educational institution) "the school was founded in 1900"
    - S: (n) dance school (a school where students are taught to dance)
    - S: (n) dancing school (a school in which students learn to dance)
    - S: (n) religious school (a school run by a religious body)

Nodes in similar position do not share same ontological properties

  - S: (n) grade school, grammar school, elementary school, primary school (a school for young children)
    - S: (n) infant school (British school for children aged 5-7)
    - S: (n) junior school (British school for children aged 7-11)
  - S: (n) correspondence school (a school that teaches nonresident students by mail)

Glosses exhibit space and time bias

- S: (n) preschool (an educational institution for children too young for elementary school)
  - S: (n) kindergarten (a preschool for children age 4 to 6 to prepare them for primary school)
  - S: (n) nursery school (a small preschool for small children)
  - S: (n) playschool, play group (a small informal nursery group meeting for half-day sessions)

Some concepts are too similar in meaning

- S: (n) public school (private independent secondary school in Great Britain supported by endowment and tuition)
  - S: (n) eton college (a public school for boys founded in 1440) located in Berkshire
  - S: (n) winchester college (the oldest English public school) located in Winchester

Some concepts are actually individuals

Problems with WordNet-like resources (II)

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

Polysemy – too fine grained distinctions in meaning

Exercises

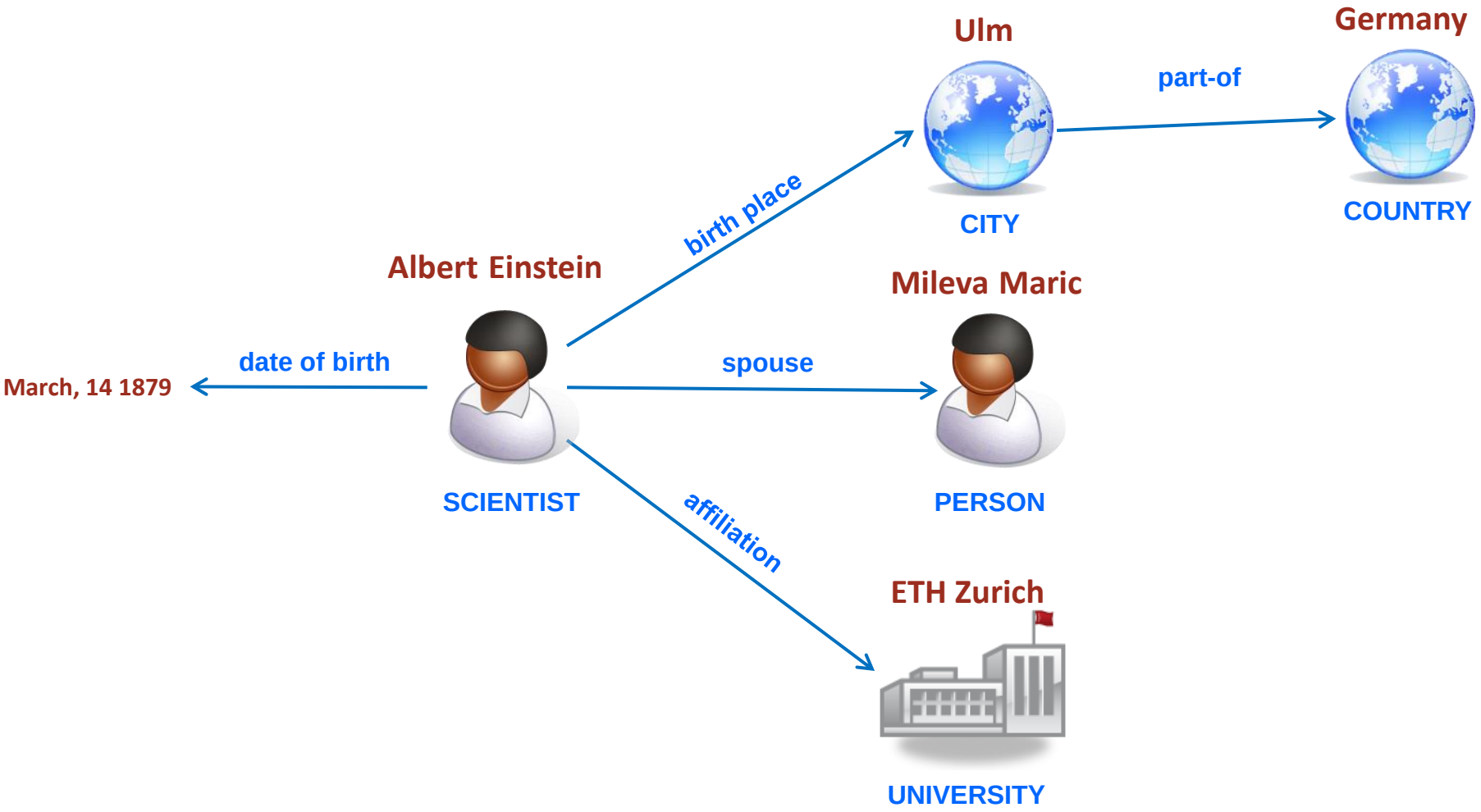
1. Identify in WordNet two synsets denoting individuals
2. Identify in WordNet two equivalent synsets, i.e. two synsets having same meaning
3. Identity in WordNet a word with a polysemy > 10
4. Identity in WordNet the direct hypernym of «museum»
5. Identity in WordNet a word with an antonym
6. Identity in WordNet three cases of space bias and three cases of time bias
7. Identify in MultiWordNet three words having a GAP in another language

Knowledge resources

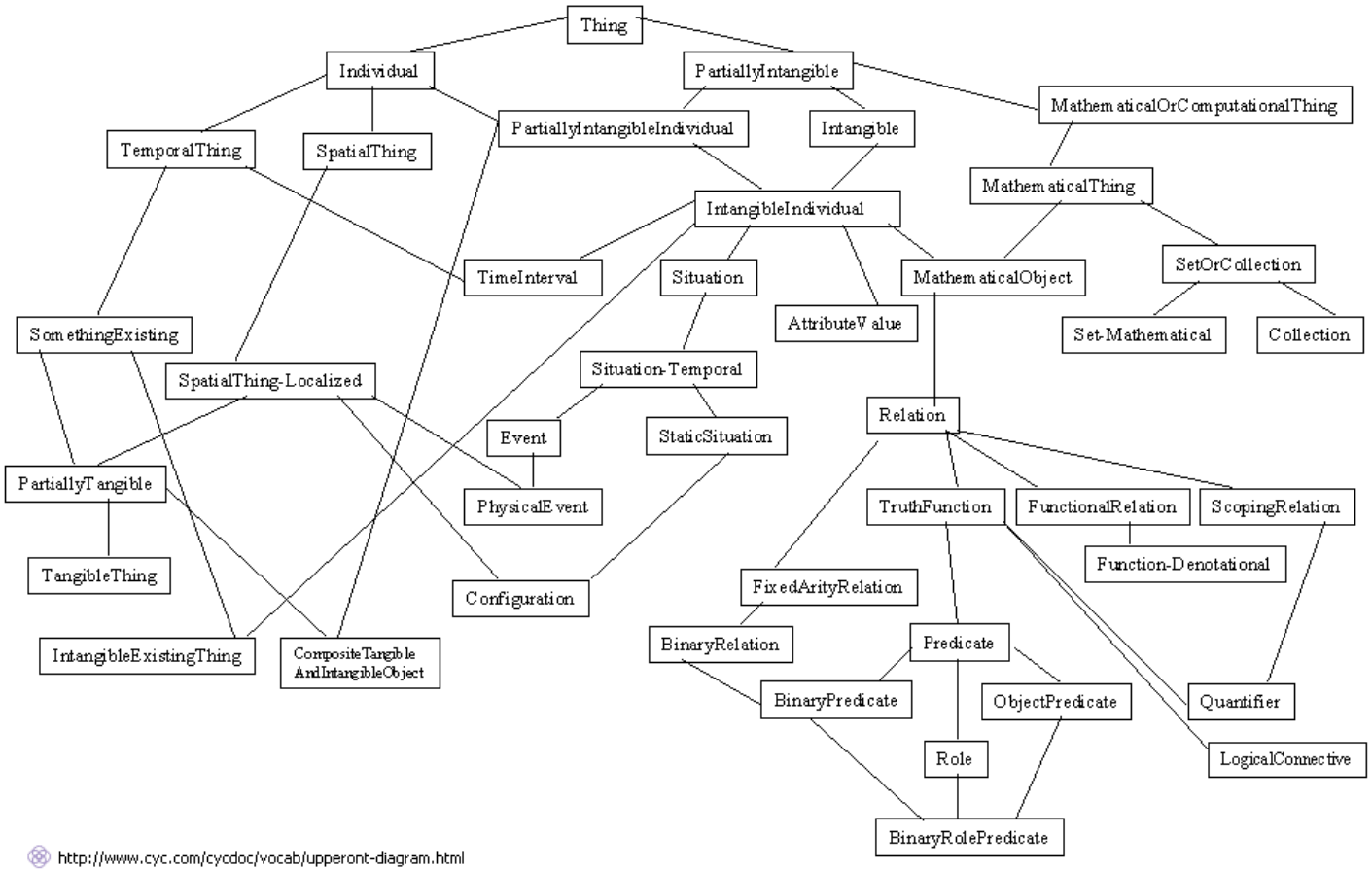
Roadmap

- **Renowned knowledge resources**
- **The (open) linked data initiative**
- **Applications**
- **Exercises**

Example of knowledge content



CYC ontology (1984)



http://www.cyc.com/cycdoc/vocab/upperont-diagram.html

Triples such as:

#$isa
#$BillClinton
#$UnitedStatesPresident

#$capitalCity
#$France
#$Paris

- A general-purpose *common sense* knowledge base
- Hand-crafted
- It contains around 2.2 million assertions and more than 250,000 terms
- Content into three levels from broader and abstract knowledge (the **upper ontology**) and widely used knowledge (the **middle ontology**) to domain specific knowledge (the **lower ontology**).

SUMO ontology (2001)

- entity
  - physical
    - object
    - process
      - dual object process
      - intentional process
        - intentional psychological process
        - recreation or exercise
        - organizational process
        - guiding
        - keeping
        - maintaining
        - repairing
        - poking
        - content development
        - making
          - constructing
          - manufacture
            - publication
          - cooking
        - searching
        - social interaction
        - maneuver
      - motion
      - internal change
      - shape change
  - abstract

Suggested Upper Merged Ontology

- A general-purpose common sense knowledge base
- Hand-crafted
- It contains around 1,000 terms and 4,000 definitional statements
- Its extension, called **MILO** (Mid-Level Ontology), covers individual domains

DBPedia (2007)

Berlin at DBpedia.org
http://dbpedia.org/resource/Berlin

Berlin is the capital city and one of the sixteen states of the Federal Republic of Germany. It is the heart of the Berlin-Brandenburg metropolitan region, located in northeastern Germany. With a population of 3.4 million, Berlin is the country's largest city, and the second most populous city in the European Union.

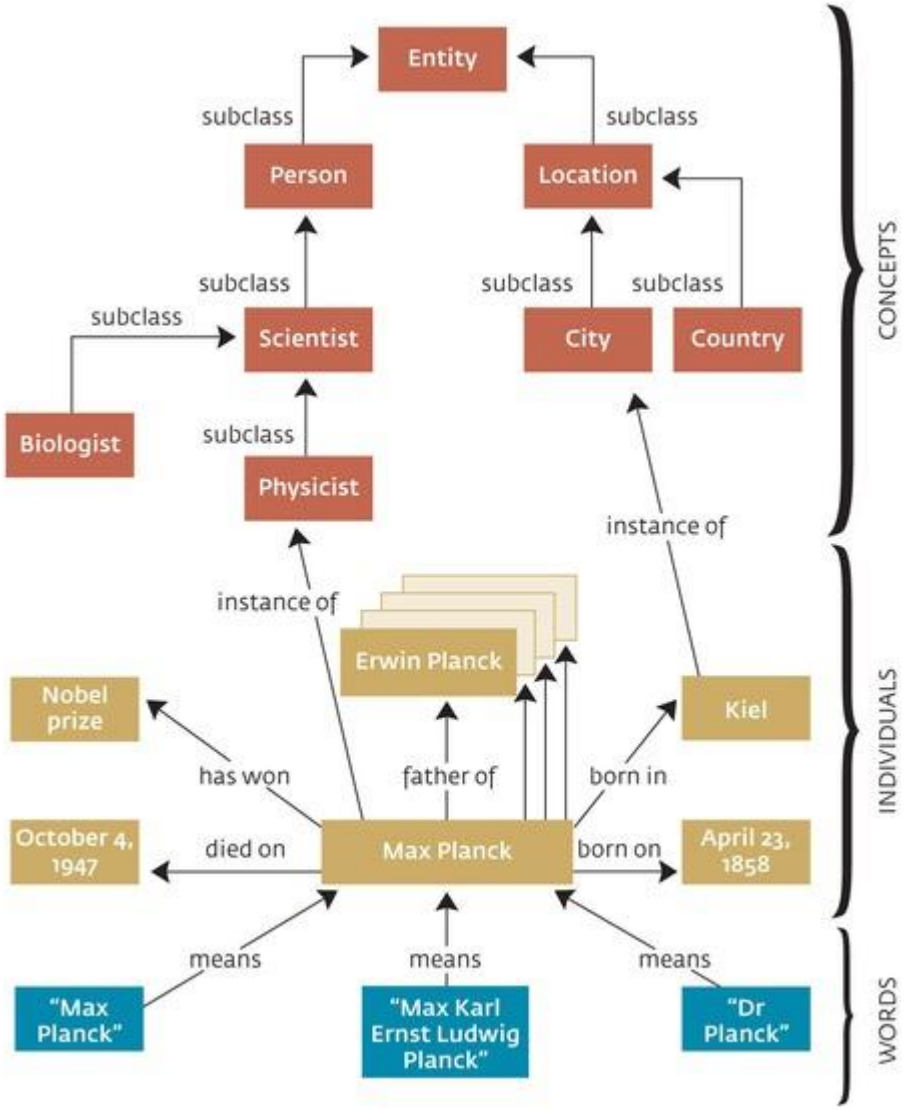
Property	Value
is p:Origin of	dbpedia:Alec_Empire dbpedia:Clara_Hill dbpedia:Frank_Duval
is p:PLACE_OF_BIRTH of	dbpedia:Drafi_Deutscher dbpedia:Hannelore_Kohl dbpedia:Hartmut_Mehdorn dbpedia:Julius_Klaproth dbpedia:Otto_Devrient
is p:PLACE_OF_DEATH of	dbpedia:August_Borsig dbpedia:Heinrich_Gr%C3%BCnfeld dbpedia:Johannes_Rau dbpedia:Ludwig_Suthaus dbpedia:Martin_Heinrich_Klaproth
is p:Recorded of	dbpedia:Benzin dbpedia:K.K.K.K.K._%28album%29 dbpedia:Mann_gegen_Mann dbpedia:Rosenrot_%28song%29
is p:STERBEORT of	dbpedia:Adolph_Wagner dbpedia:Albert_Heilmann dbpedia:Max_Taut dbpedia:Robert_von_Mohl
p:abstract	Berlin is the capital city and one of the sixteen states of the Federal Republic of Germany. It is the heart of the Berlin-Brandenburg metropolitan region, ... »more» (en)

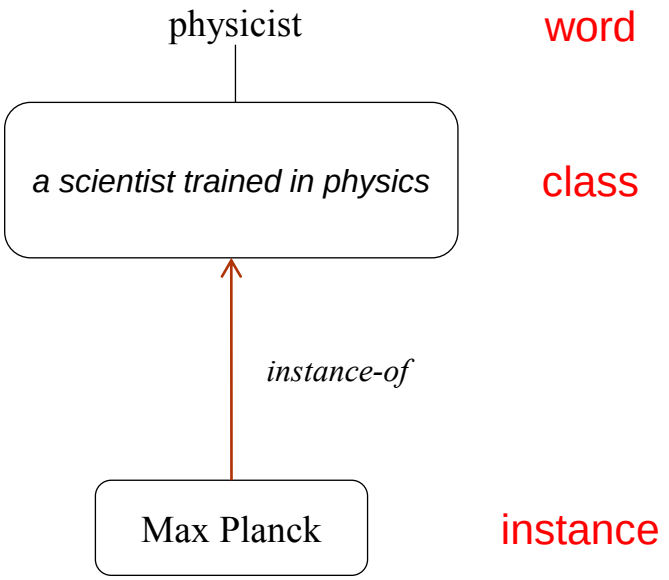
Wikipedia

Country	Germany
Government	
• **Governing Mayor**	Klaus Wowereit (SPD)
• **Governing parties**	SPD / CDU
• **Votes in Bundesrat**	4 (of 69)
Area	
• **City**	891.85 km^2 (344.35 sq mi)
Elevation	34 m (112 ft)
Population (December 2013)[1]	
• **City**	3,517,424
• **Density**	3,900/km^2 (10,000/sq mi)
Time zone	CET (UTC+1)
• **Summer (DST)**	CEST (UTC+2)
Postal code(s)	10115–14199
Area code(s)	030
ISO 3166 code	DE-BE
Vehicle registration	B[2]
GDP/ Nominal	€109.2 billion (2013) [3]
NUTS Region	DE3
Website	berlin.de

- It is automatically built by extracting semi-structured content from Wikipedia
- Text is not semantically analyzed

YAGO ontology (2008)

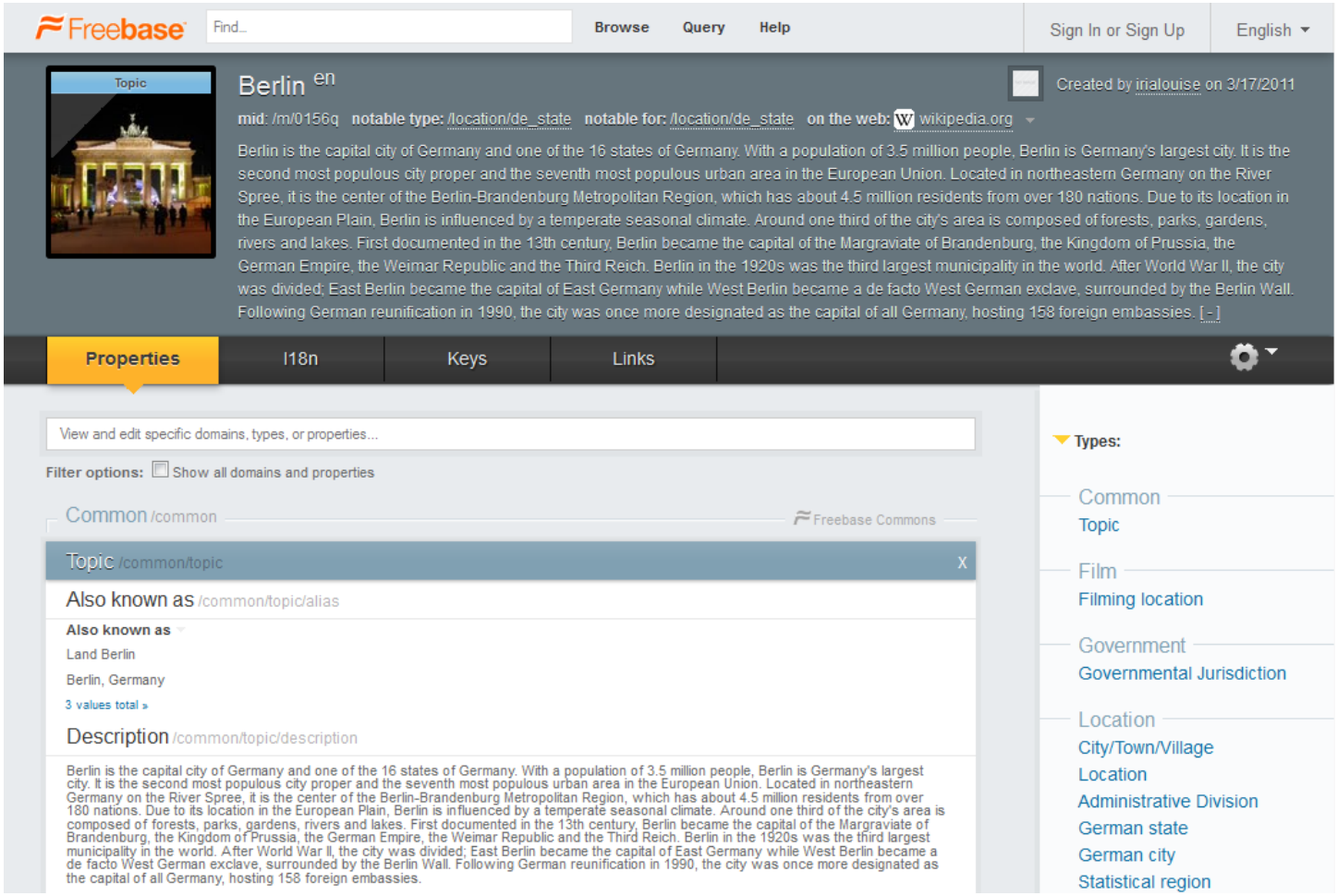




- **Concepts** are taken from noun synsets of WordNet
- **Instances** and their properties are automatically extracted from Wikipedia
- The **linking** of concepts with instances is done via NLP techniques

- Accuracy is claimed to be ~95%
- It is available in triple (RDF) format

Freebase (2010)



- Semi-automatically built
- It contains data harvested from several sources such as Wikipedia, NNDB, FMD and MusicBrainz, as well as individually contributed data from its users.

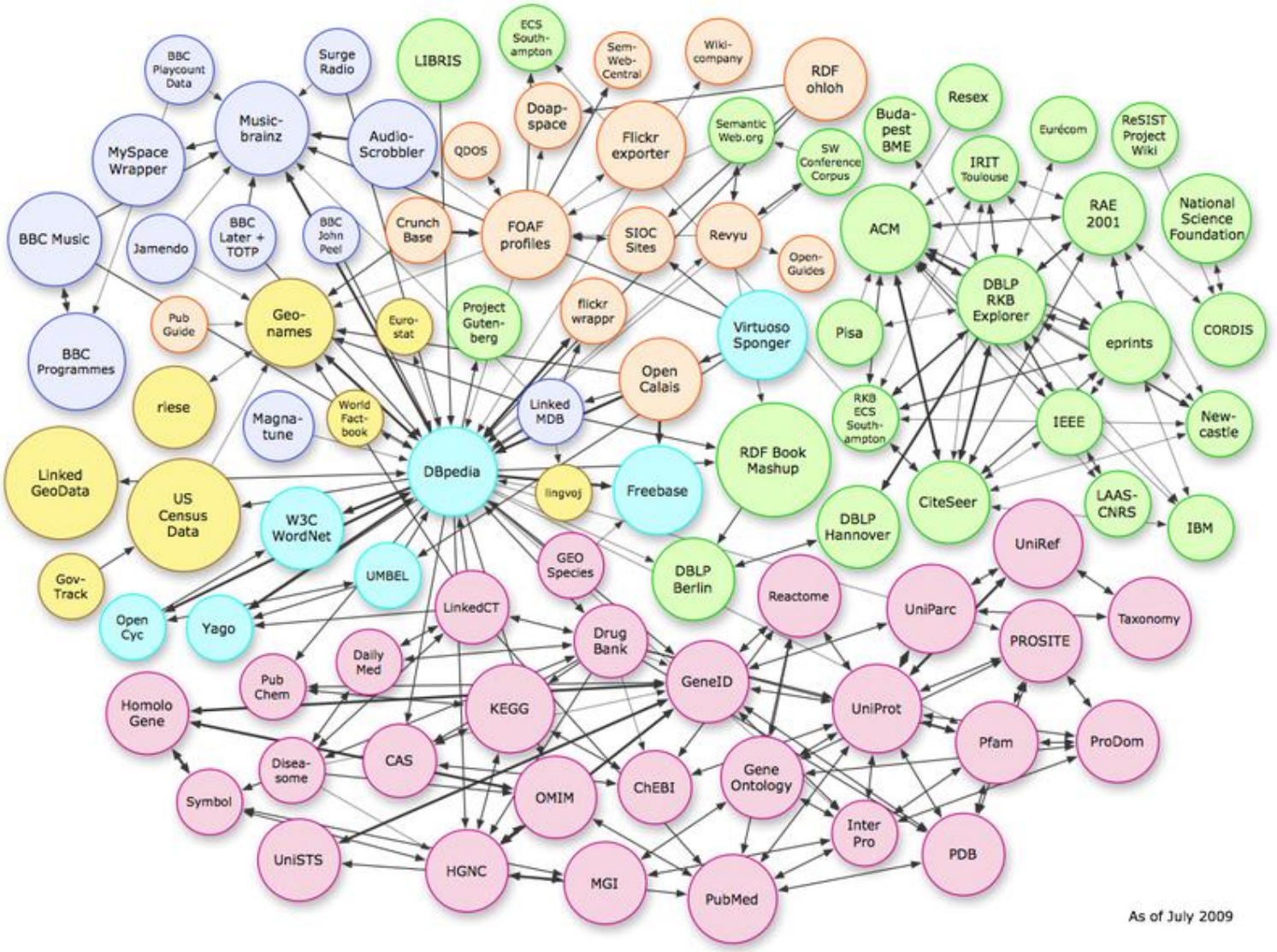
The Schema.org initiative

Thing > Person

A person (alive, dead, undead, or fictional).

Property	Expected Type	Description
Properties from Person		
additionalName	Text	An additional name for a Person, can be used for a middle name.
address	PostalAddress	Physical address of the item.
affiliation	Organization	An organization that this person is affiliated with. For example, a school/university, a club, or a team.
alumniOf	EducationalOrganization	An educational organizations that the person is an alumni of. Inverse property: alumni.
award	Text	An award won by this person or for this creative work. Supersedes awards.
birthDate	Date	Date of birth.
brand	Organization or Brand	The brand(s) associated with a product or service, or the brand(s) maintained by an organization or business person.
children	Person	A child of the person.
colleague	Person	A colleague of the person. Supersedes colleagues.

Linked Data Cloud (since 2007)



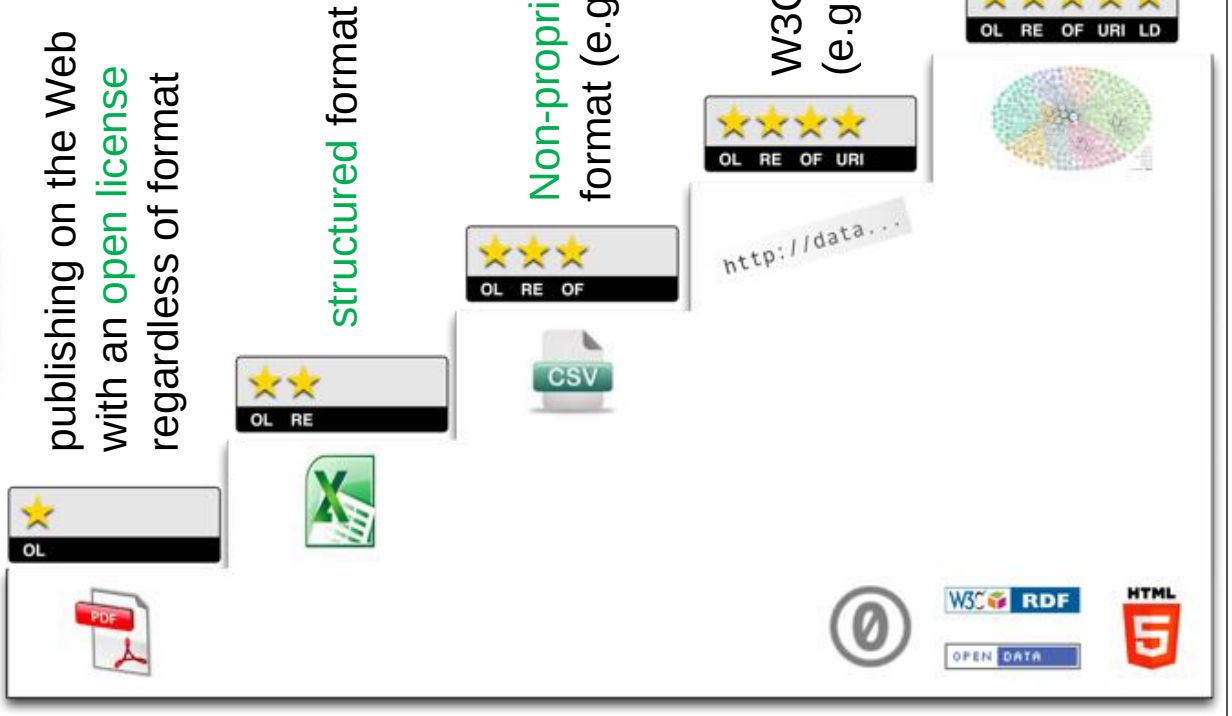
Linked Data

The Linked Data approach forms the basis of data publishing guidelines pinpointing how data from government, public and private sectors can be more valuable for the consumers.

Principles

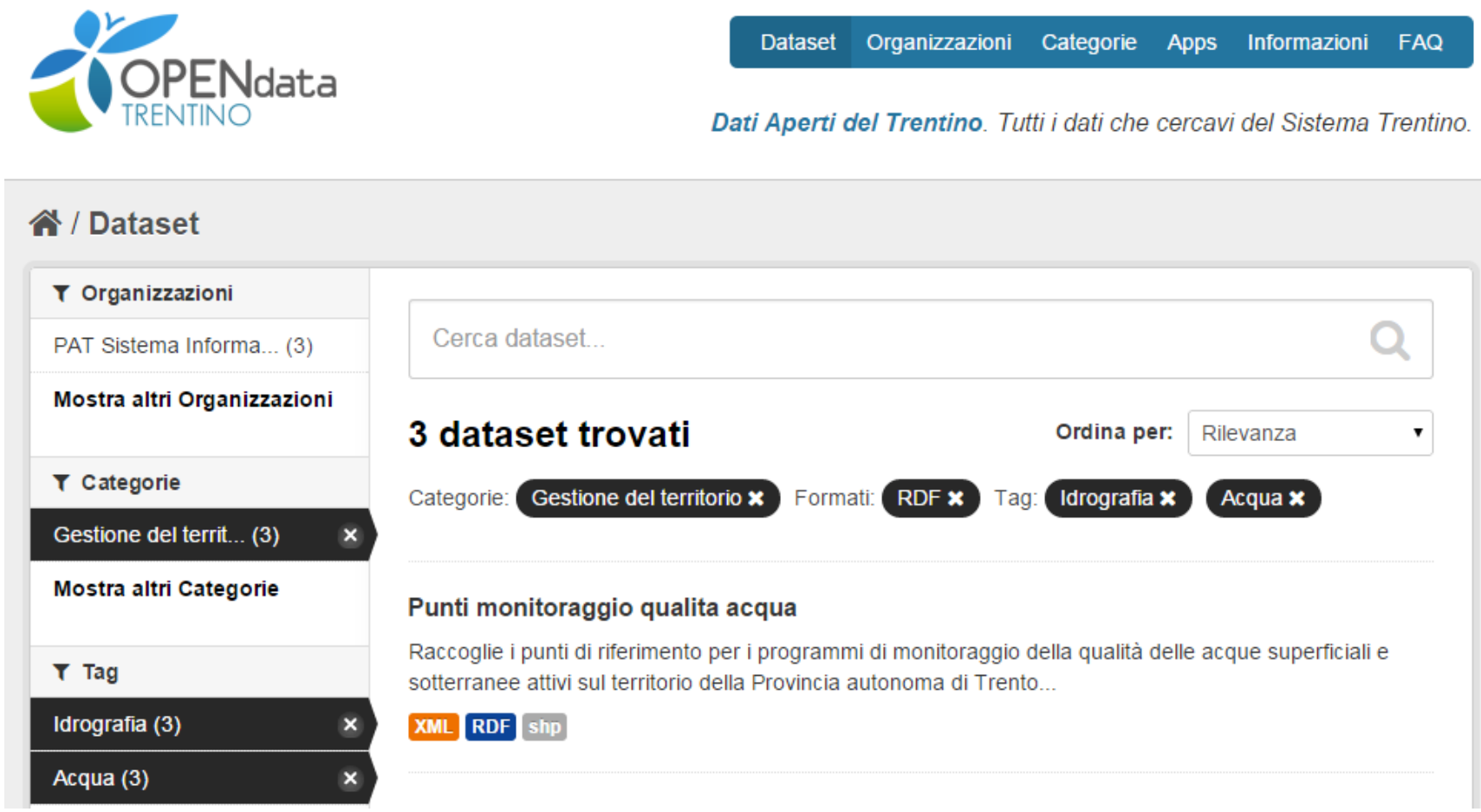
- the use of http URIs as the identifiers of things (concepts, entities and attributes)
- the provision of meaningful content published in open format (RDF) for each URI reference
- the production of navigable content via links

Linked Open Data

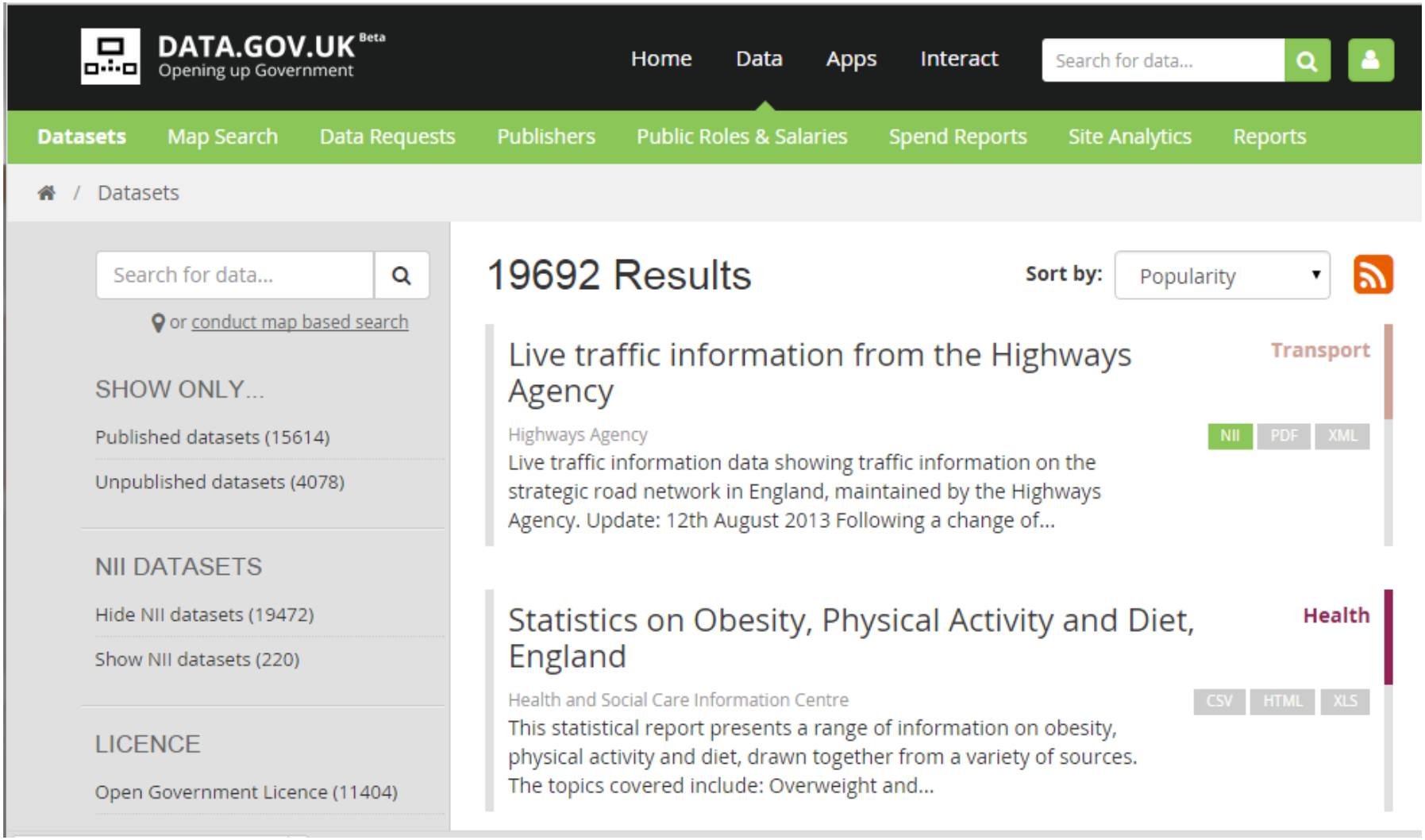


The Semantic Geo-catalogue of the PAT

Open Data Trentino portal



Open Government Data in UK



Exercises

1. Design two small knowledge graphs about a famous person taking information from Wikipedia, and YAGO (use the YAGO browser)
2. Explore Freebase and find information about Trento
3. Explore http://data.gov.uk/ and find useful information about museums
4. Search for the linked data cloud and check how many datasets it currently contains

Capturing diversity with the UKC and Entitypedia

Roadmap

- **Diversity and diversity dimensions**
- **The entity-centric approach**
- **The UKC and Entitypedia**
- **Exercises**

The inherent diversity of the world

What does bug mean?

… goals, culture, belief, personal experience …

Diversity is pervasive in world descriptions

Within a natural language

- "bug as malfunction" vs. "bug as food" *(homonymy)*
- "stream" and "watercourse" have same meaning *(synonymy)*

Across natural languages

- "watercourse" in English is same as "corso d'acqua" in Italian (*concepts*)
- There is no lemma in Italian for "biking" (*lexical GAP*)

In formal language

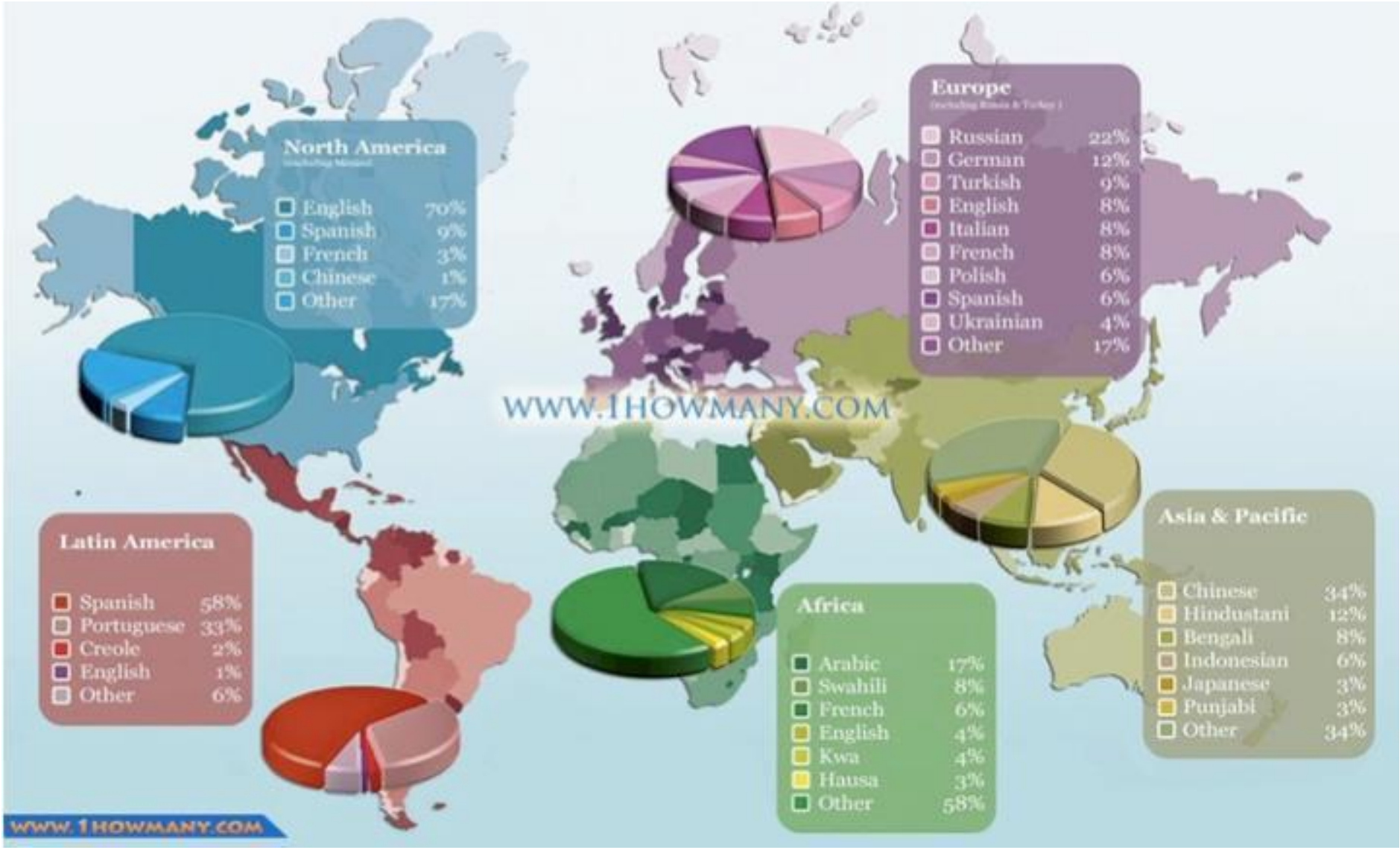
- There are several types of bodies of water (*semantic relations*)
- Rivers have a length, lakes have a depth (*schematic knowledge*)

In data (ground knowledge)

- The Adige river is 410 Km long; The Garda lake is 136 m deep

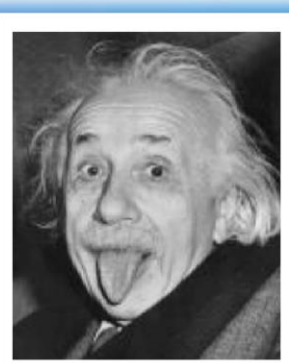
- "Bugs are great food" vs. "how can you eat bugs?" *(the role of culture)*
- "Climate is/is not an important issue" *(the role of schools of thought)*

Diversity in language



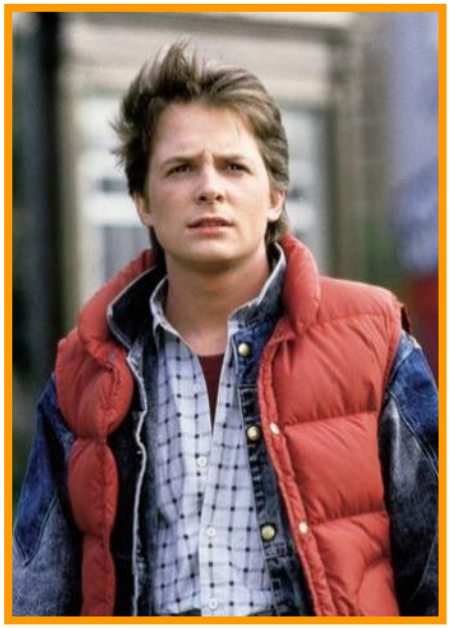
Diversity in Knowledge

- Billions of locations
- Billions of people
- Millions of organizations
- … and events, artifacts, creative works, …



Terminological and ground Knowledge

Actor — **acted in** — **Movie, Film**



Michael J. Fox — **acted in** — **Back to the future II**

An entity-centric vision of the world (I)

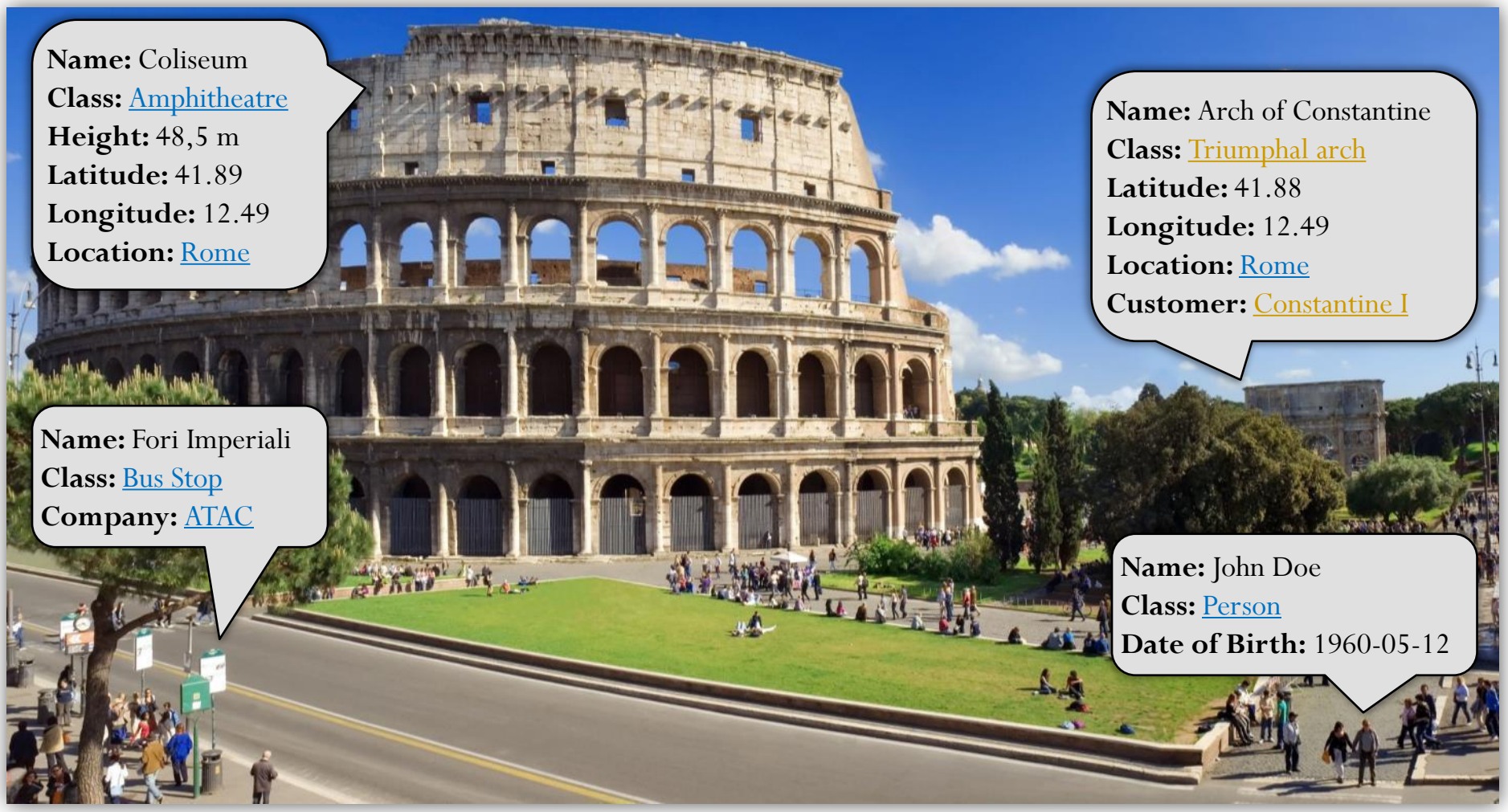
- **Entities are objects which are so important in our everyday life to be referred with a name**
- **Each entity has its own attributes (e.g. latitude, longitude, height…)**
- **Each entity is in relation with other entities (e.g. Eiffel Tower is located in Paris, France)**
- **Each entity as a reference class (e.g. monument) which determines its entity type (e.g. location)**

Eiffel Tower

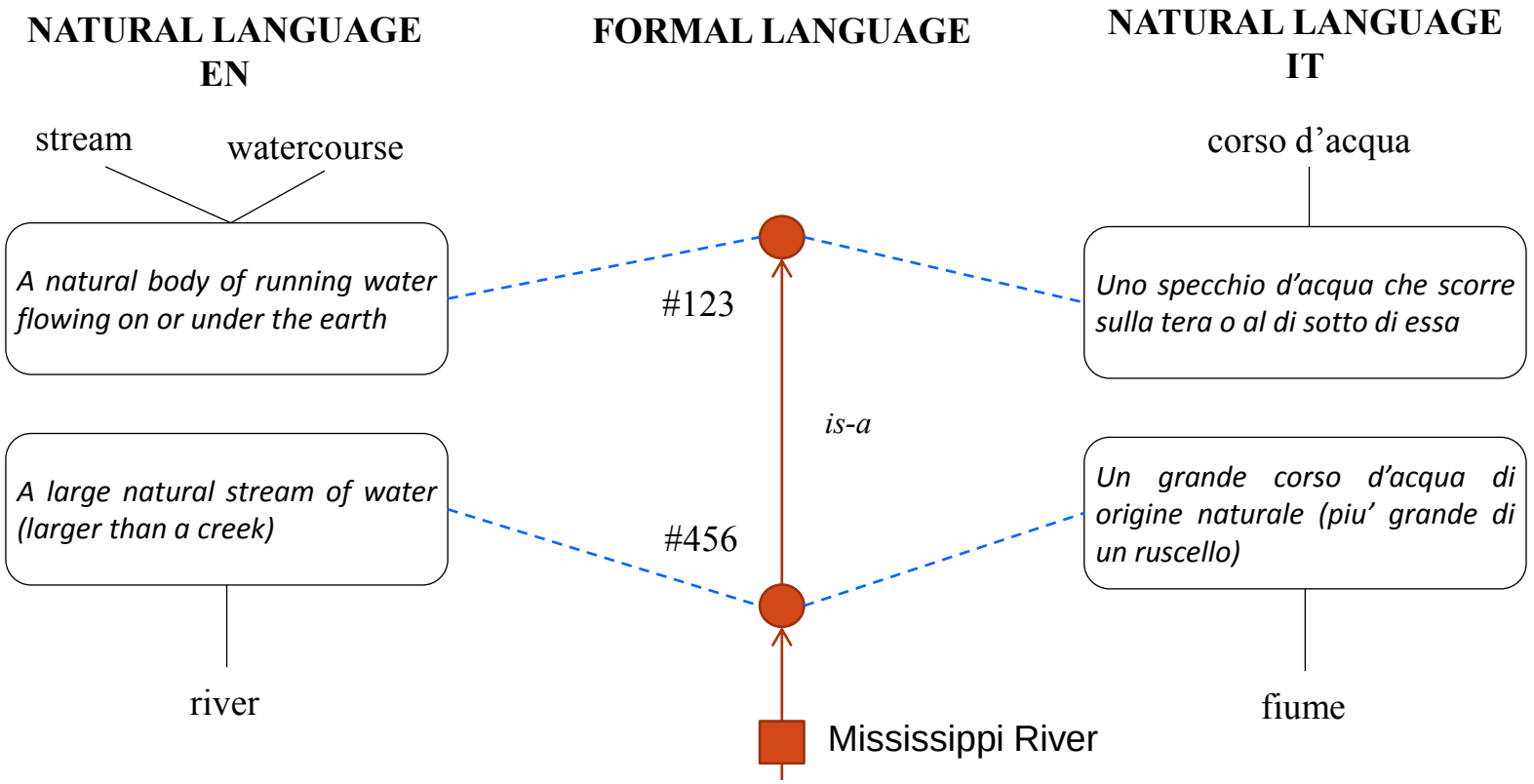
An entity-centric vision of the world (II)

Entities are not all the same; they have different metadata according to the type of entity

What do we aim to? How to achieve that?



The UKC and Entitypedia (since 2010)



- Manually built via collaborative development [Tawfik et al., 2014], bootstrapped from WordNet, MultiWordNet, GeoNames
- Split natural language, formal language and ground knowledge [Giunchiglia et al., 2012b]
- Domain knowledge is created following the DERA methodology [Giunchiglia et al., 2012a] and principles [Giunchiglia et al., 2009] with distinction between entities, **classes**, **relations**, **attributes** and values

The UKC components

The natural language:
our vocabulary in multiple languages

Natural Language Core (NLC)

The fomal language:
our graph of language-independent notions

Concept Core (CC)

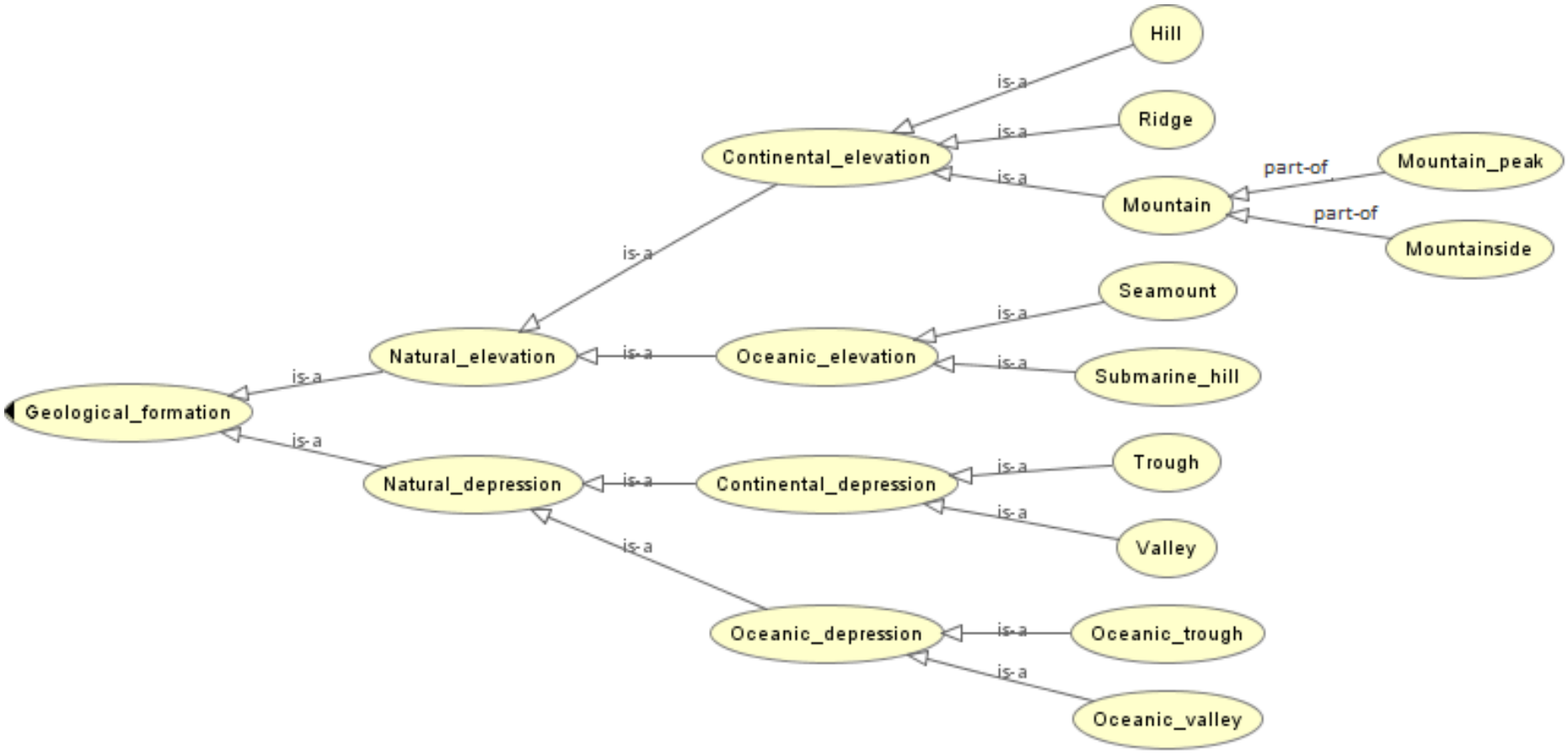
Schematic knowledge:
Our schema of basic entity types

EType Core (ETC)

Domain knowledge:
Domain-specific partition of the language above

Domain Core (DC)

Concept Core

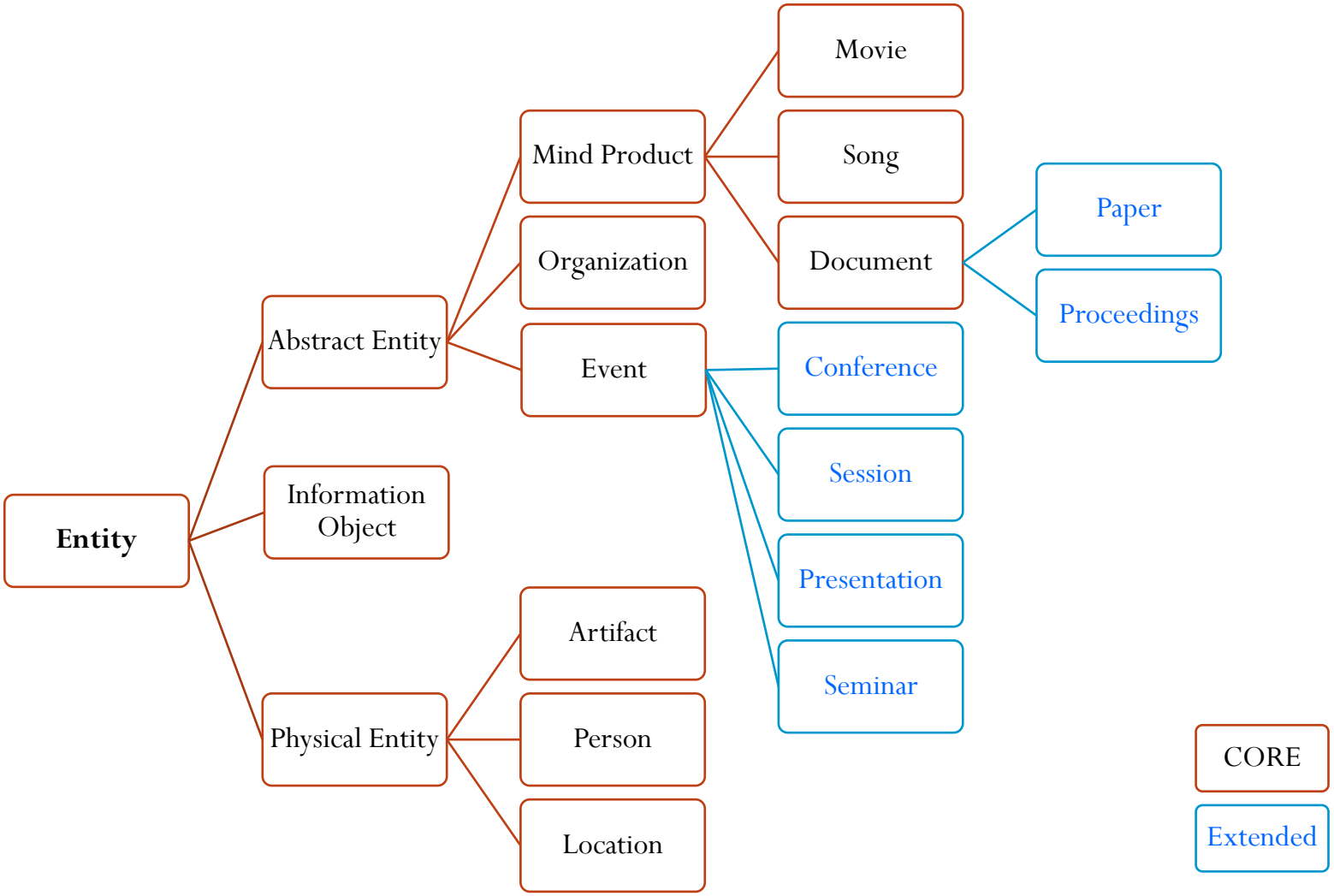


Natural Language Core

Language	Synset	Gloss
en	Canal	long and narrow strip of water made for boats or for irrigation
it	canale; naviglio	corso d'acqua artificiale, costruito per l'irrigazione o la navigazione
mn	суваг	усжуулалт эсвэл завинд зориулсан барьсан усны урт нарийн гудамж
bn	খাল	পানির দীর্ঘ এবং সরু ধারা যা সেচ বা নাব্যতার জন্য তৈরি করা হয়েছে
zh	沟渠; 运河	人工水道或人工修缮的河流，用于旅行、航运或灌溉
hi	नहर; कुलिया	सिंचाई, यात्रा आदि के लिए छोटी नदी के रूप में तैयार किया हुआ जलमार्ग

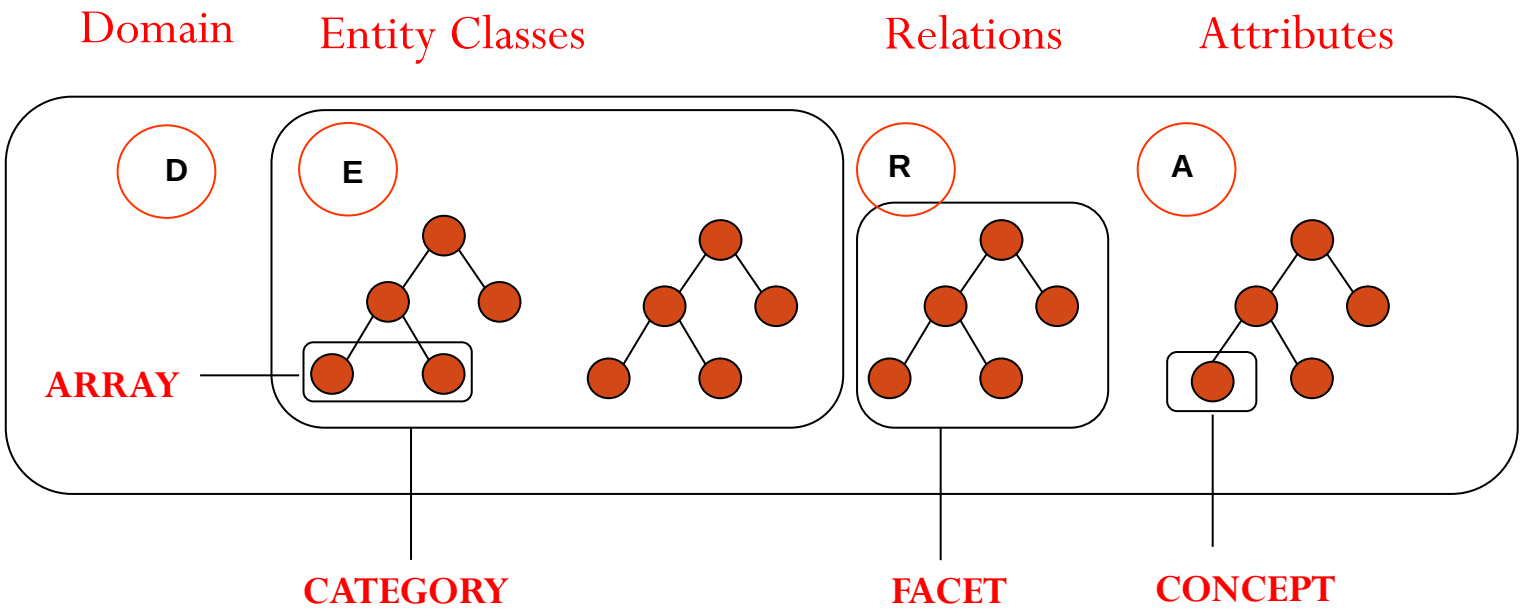
Language	Synset	Gloss
en	Rivulet	A small stream
mn	GAP	छोटी सी धारा

Etype Core: lattice (sample)



Domain Core: the DERA methodology

- To capture terminology relevant to a specific domain
- Based on the faceted approach from Library and Information Science
- Terminology can be directly codified into Description Logic



Entitypedia compared with existing knowledge bases

KB	#entities	#facts	Domains	Distinction classes and instances	Distinction NL/FL	Manual
CYC	250K	2.2 M	Yes	No	No	Yes
OpenCYC	47k	306k	Yes	No	No	Yes
SUMO	1k	4k	No	Yes	Yes	Yes
MILO	21k	74k	Yes	Yes	Yes	Yes
DBPedia	3.5 M	500 M	No	No	No	No
YAGO	2.5 M	20 M	No	No	No	No
Freebase	22 M	?	Yes	Yes	No	Yes
Entitypedia	10 M	80 M	Yes	Yes	Yes	Yes

Exercises

1. Search on the Web information about how many languages are spoken in Europe and in the whole world.
2. What is the most widely spoken language in the world?
3. Provide an example of concept which is heavily cultural dependant.
4. What are the top level entity types (up to 10) that to you are necessary to codify the whole world knowledge?
5. What are the main novelties introduced by the UKC and Entitypedia w.r.t. previous approaches?

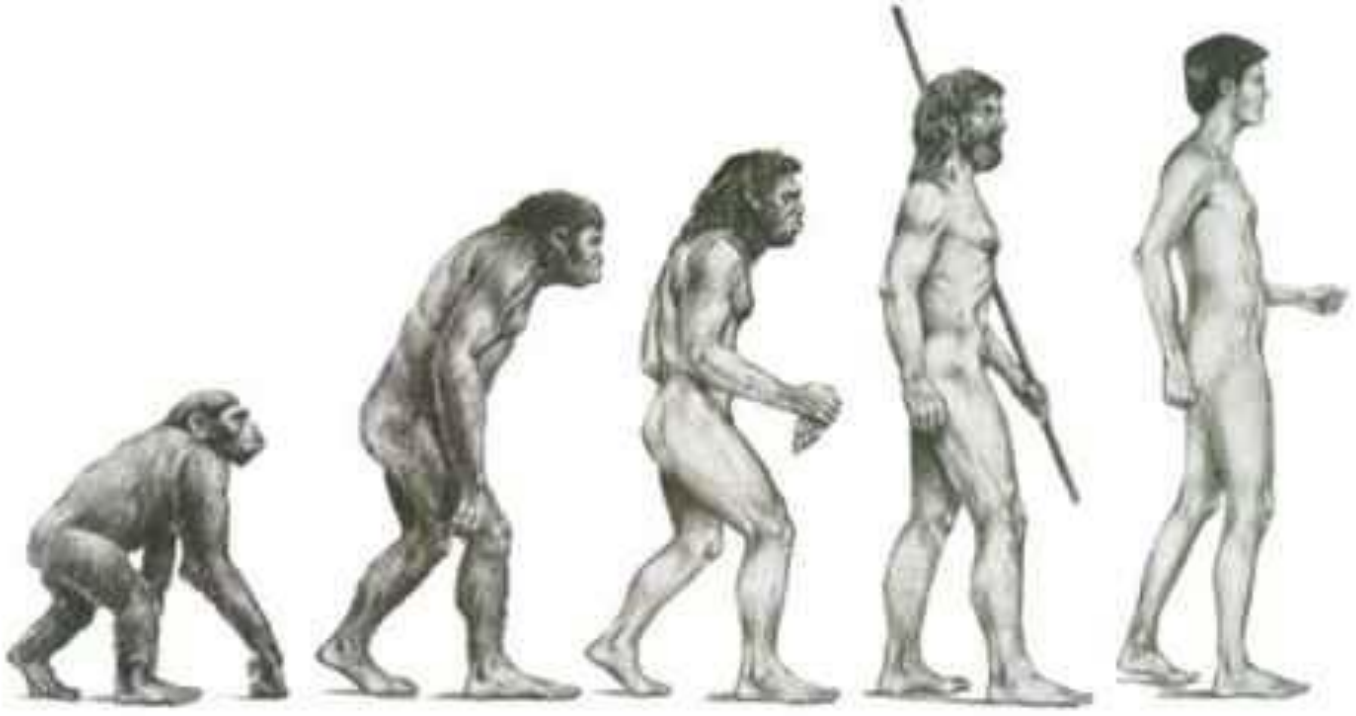
Methodologies for content generation

Roadmap

- **Introduction**
  - **Motivation**
  - **The original faceted approach**
- **Primitive notions in DERA**
- **Steps in the methodology**
- **Guiding principles**
- **Converting DERA ontologies into DL**
- **Applications**
- **Exercises**

WHY DO WE NEED A METHODOLOGY?

BECAUSE SMALL DIFFERENCES MATTER…



Humans and chimps share a surprising 98.8 percent of their DNA.

How to build ontologies which are of the highest quality possible?

Methodologies to ontology development

- Several methodologies have been developed for the construction and maintenance of ontologies (KR) or controlled vocabularies (KO)
- The faceted approach **[Ranganathan, 1967]** from library science is known to have great benefits in terms of quality and scalability
- It is based on the fundamental notions of *domain* and *facets*, which allow capturing the different aspects of a domain and allow for an incremental growth.
- Originally facets were of 5 types (PMEST): Personality, Matter, Energy, Space, Time.
- A key feature is compositionality (meccano property), i.e. the system allows a subject to be constructed by freely combining some basic components (facets).

[D] Medicine

[E] Body Part
. Digestive System
. . Stomach

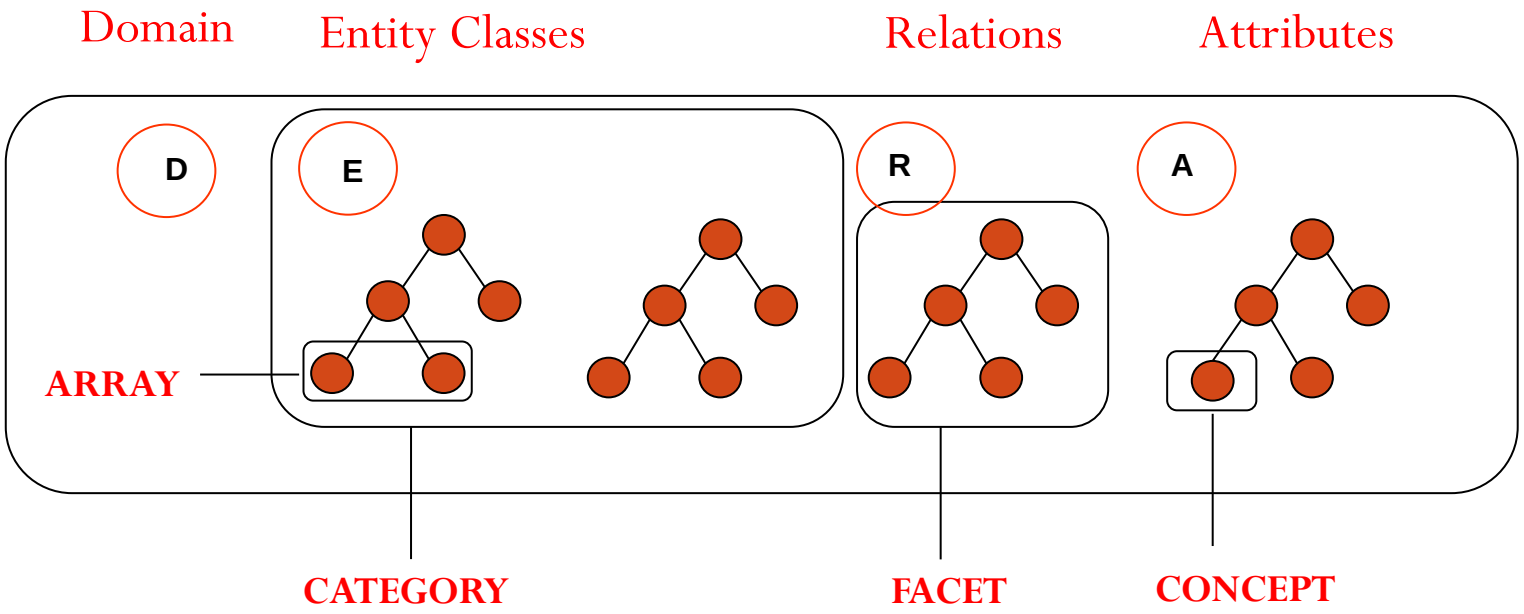
[P] Disease
. Cancer
. . Carcinoma
. . . Adenocarcinoma

[A] Action
. Treatment

[M] Kind (to be applied to [A] Action)
. Chemotherapy

The DERA framework

- To capture terminology relevant to a specific domain
- DERA is **faceted** as it is inspired to the faceted approach
- DERA is a **KR approach** as it models entities of a domain (D) by their entity classes (E), relations (R) and attributes (A)
- Terminology can be directly codified into Description Logic



Domains

- Any area of knowledge or field of study that we are interested in or that we are communicating about that deals with specific kinds of entities:

- Domains are the main means by which the ***diversity of the world*** is captured, in terms of language, knowledge and personal experience.

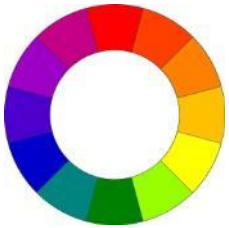
Primitive notions

- **Entity**: a (digital) description of any real world physical or abstract object so important to be denoted with a proper name. A single person, a place or an organization are all examples of entities.

- **Entity Class:** any set of objects with common characteristics.

- **Relation**: any object property used to connect two entities. Typical examples of relations include part-of, friend-of and affiliated-to.

- **Attribute**: any data property of an entity. Each attribute has a name and one or more values taken from a range of possible values.



Elements of DERA

A DERA domain is a triple D = <E, R, A> where:

- ***E** (for Entity)* is a set of facets grouping terms denoting *entity classes*, whose instances (the entities) have either perceptual or conceptual existence. Terms in these hierarchies are explicitly connected by *is-a* or *part-of* relation.
- ***R** (for Relation)* is a set of facets grouping terms denoting relations between entities. Terms in these hierarchies are connected by *is-a* relation.
- ***A** (for Attribute)* is a set of facets grouping terms denoting *qualitative/quantitative* or *descriptive* attributes of the entities. We differentiate between attribute names and attribute values such that each attribute name is associated corresponding values. Attribute names are connected by *is-a* relation, while attribute values are connected to corresponding attribute names by *value-of* relations.

DERA facets

- DERA provides the language required to describe entities of a certain entity type in a given domain (D)
- Language comprises entity classes (E), relations (R) and attributes (A), names and values.
- Concepts and semantic relations between them form hierarchies of homogeneous nature called facets, each of them codifying a different aspect of the domain.
- Each facet is a descriptive ontology [Giunchiglia et al., 2014]

ENTITY CLASS

Location
- Landform
- (is-a) Natural elevation
  - (is-a) Continental elevation
    - (is-a) Mountain
    - (is-a) Hill
  - (is-a) Oceanic elevation
    - (is-a) Seamount
    - (is-a) Submarine hill
- (is-a) Natural depression
  - (is-a)Continental depression
    - (is-a) Valley
    - (is-a) Trough
  - (is-a) Oceanic depression
    - (is-a) Oceanic valley
    - (is-a) Oceanic trough

Body of water
- (is-a) Flowing body of water
  - (is-a) Stream, Watercourse
    - (is-a) River
    - (is-a) Brook
- (is-a) Still body of water
  - (is-a) Lake
  - (is-a) Pond

RELATION

Direction
- (is-a) East
- (is-a) North
- (is-a) South
- (is-a) West

Relative level
- (is-a) Above
- (is-a) Below

Containment
- (is-a) part-of

ATTRIBUTE

Name

Latitude

Longitude

Altitude

Area

Population

Depth
- (value-of) deep
- (value-of) shallow

Length
- (value-of) long
- (value-of) short

Analysis of the term “school”

Term: School			
Source	**Definition**	**Genus**	**Differentia**
WordNet	an educational institution	institution	educational
Oxford dictionary	an institution for educating children	institution	for educating children
Merriam-Webster	an institution for the teaching of children	institution	for the teaching of children
Wikipedia	an institution designed for the teaching of students (or "pupils") under the direction of teachers	institution	for the teaching of students

The term school is in general highly polysemous. Among others, school may denote a building. In the context of educational organizations, as from above, it seems there is quite an agreement about the fact that it indicates a kind of educational institution, but in some cases (such as fore WordNet) the meaning is left very generic. We coined the following definition: *“an educational institution designed for the teaching of students under the direction of teachers”*.

Synthesis of educational organizations

Educational Institution
- *<by level of complexity>*
- Preschool
- School
  - Primary school
  - Secondary school
  - Post-secondary school
    - *<by programme orientation>*
    - Training school
    - Vocational school
    - Technical school
    - Graduate school
- College
- University

Synthesis of educational organizations

Educational Institution (*an institution dedicated to education*)

Preschool (*an educational institution for children too young for primary school*)

School (*an educational institution designed for the teaching of students under the direction of teachers*)

Primary school (*a school for children where they receive the first stage of basic education*)

Secondary school (*a school for students intermediate between primary school and tertiary school*)

Tertiary school (*a school where programmes are largely theory based and designed to provide sufficient qualification for entry to advanced research programmes or professions with high skill requirements and leading to a degree*)

Training school (*a tertiary school providing theoretical and practical training on a specific topic or leading to certain degree*)

Vocational school (*a tertiary school where students are given education and training which prepares for direct entry, without further training, into specific occupation*)

Technical school (*a tertiary school where students learn about technical skills required for a certain job*)

Graduate school (*a tertiary school in a university or independent offering study leading to degrees beyond the bachelor's degree*)

College (*an educational institution or a constituent part of a university or independent institution, providing higher education or specialized professional training*)

University (*an educational institution of higher education and research which grants academic degrees in a variety of subjects and provides both undergraduate education and postgraduate education*)

Guiding principles

Principle	Example
Relevance	breed is more realistic to classify the universe of cows instead of by grade
Ascertainability	flowing body of water
Permanence	spring as a natural flow of ground water
Exhaustiveness	to classify the universe of people, we need both male and female
Exclusiveness	age and date of birth, both produce the same divisions
Context	bank, a bank of a river, OR, a building of a financial institution
Currency	metro station vs. subway station
Reticence	minority author, black man
Ordering	stream preferred to watercourse

Guidelines for the formal language

- ***Concepts:*** facets in UKC are descriptive ontologies where each concept denotes a set of real world entities (classes) or a property of real world entities (relations and attributes).
- ***Look for essential concepts:*** a property of an entity (that we codify as a concept) is essential (as opposite of accidental) to that entity if it must hold for it. As special form of essence, a property is rigid if it is essential to all its instances [Guarino and Welty, 2002].
- ***Avoid complex concepts:*** e.g. "red car".
- ***Avoid redundancies:*** e.g. "nursery school" and "kindergarten" are synonyms
- ***Avoid individuals:*** e.g. "United States military academy"
- ***Pay attention to meronymy relations:*** while *part-of* is assumed to be transitive in general, *substance-of* and *member-of* are not. Therefore, the latter two cannot be considered as hierarchical. In fact, [Varzi, 2006] describes some of the paradoxes that would be generated in assuming otherwise.

Guidelines for the natural language (I)

- ***Terms and synsets:*** terms are grouped into synsets. In UKC multiple languages are accounted for by developing multiple dictionaries, i.e. by assigning either a synset or a GAP to every concept.
- ***Lemmas:*** for the selection of terms we focus on lemmas.
- We do not accept in UKC:
  - ***articles*** (e.g. the) and ***plural forms***;
  - ***capitalization***, except for cases such as acronyms and abbreviations;
  - ***punctuation characters and parenthesis***;
- The following are instead accepted, but not recommended:
  - ***loan terms***, i.e. terms borrowed from other languages, if widely used. For instance, the term *kindergarten* in English is typically well accepted.
  - ***transliterations***, i.e. when a terms is a transcript from one alphabet to another one.

Guidelines for the natural language (II)

- ***Parts of speech:*** noun, adjective, adverb and verb. A lemma can be a single word (e.g. bank), a multi-word (e.g. traffic light) or a prepositional phrase (e.g. place of warship).
- ***Homographs:*** terms which are spelled the same, but have different meaning. The same term can be associated to multiple concepts.
- ***Glosses:*** in line with principle of reticence, a gloss should not convey any cultural, temporal or regional bias.

Primary school: a school for young children; usually the first 6 or 8 grades
Infant school: British school for children aged 5-7
Junior school: British school for children aged 7-11

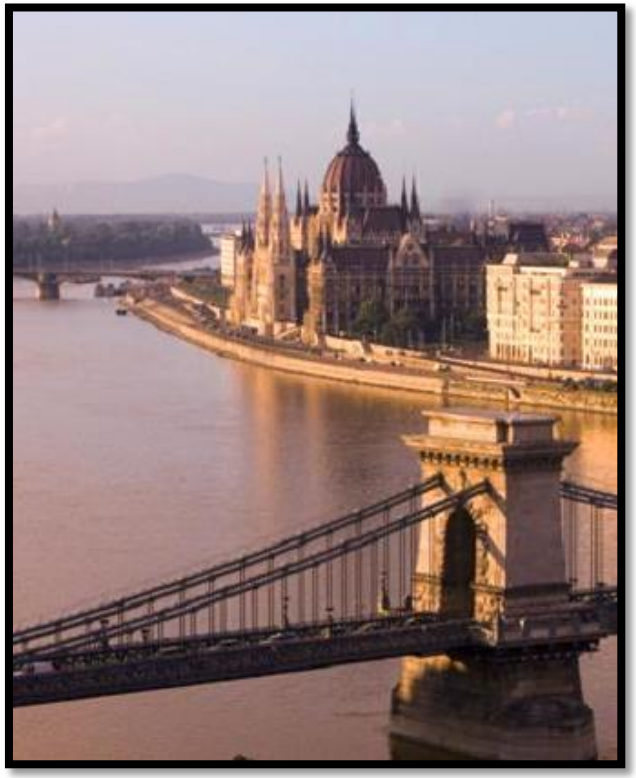
NO

Primary school: a school for children where they receive the first stage of basic education
Infant school: a primary school for very young children where they learn basic reading and writing skills
Junior school: a primary school for young children where they learn basic notions of core subjects such as math, history and other social sciences

YES

Back to entities

Entity Class	**Class:**	**River**
Attributes	**Name:**	**Thames**
	Latitude:	**51.50**
	Longitude:	**0.61**
	Length:	**346 km (long)**
Relations	**Part-of:**	**UK**

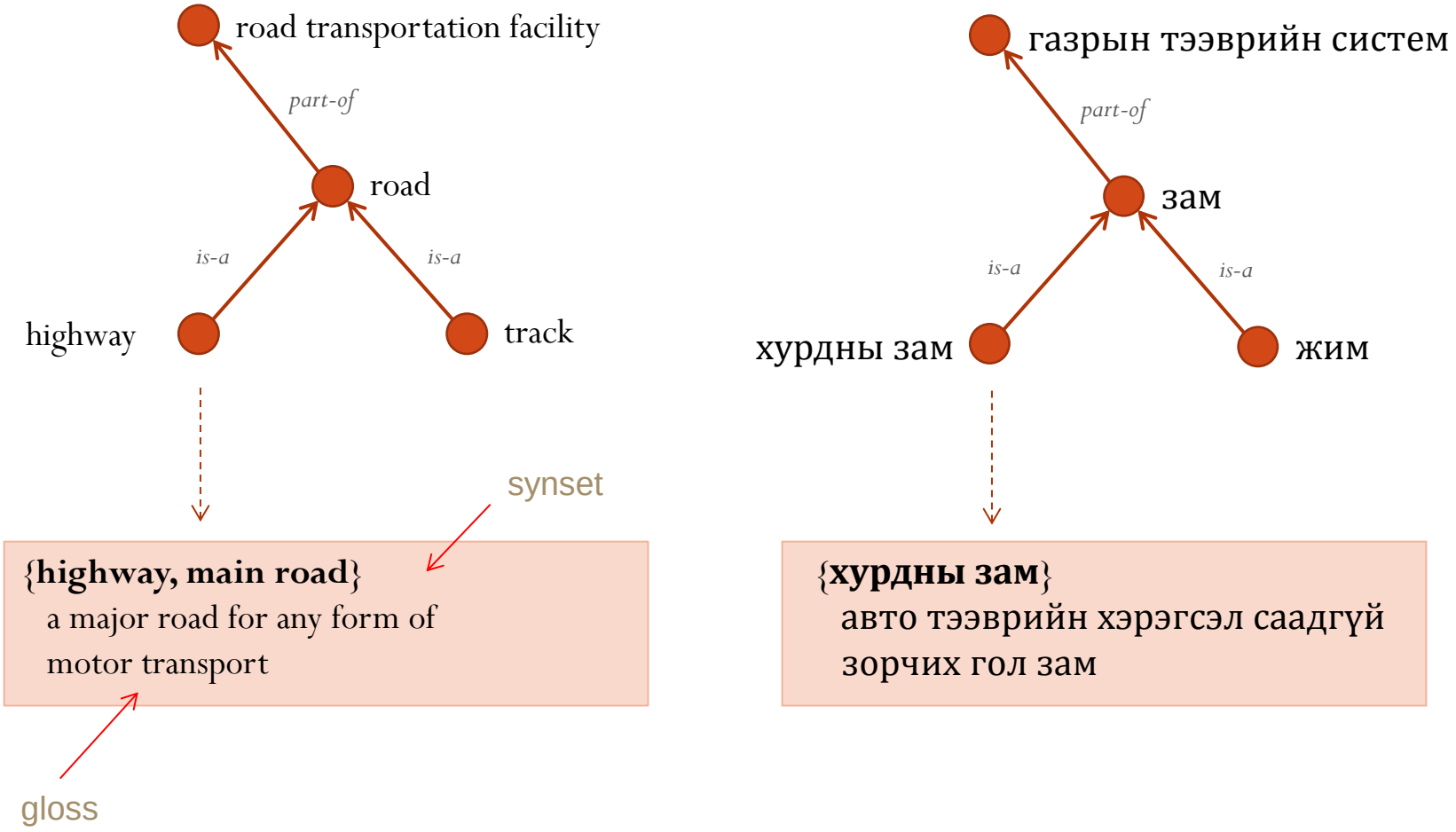


Thames

Each of the terms above comes from a DERA ontology in KB

Localization [Ganbold et. al., 2014]

English translation → **Mongolian**



Formalizing DERA into DL (I)

With the formalization, DL concepts denote either sets of entities or sets of attribute values. DL roles denote either relations or attributes.

A DL *interpretation* $I = <\Delta, I>$ consists of the *domain of interpretation* $\Delta = F \cup G$ where:

- F is a set of individuals denoting real world *entities*
- G is a set of *attribute values*

and of an interpretation function I where:

$$E_i^I \subseteq F \qquad R_j^I \subseteq F \times F \qquad A_k^I \subseteq F \times G \qquad v_r^I \in G$$

Formalizing DERA into DL (II)

	Object	DL formalization	
$E_1, \ldots, E_p$	entity classes	Concepts	TBox
$R_1, \ldots, R_q$	relations between classes	Roles	
$A_1, \ldots, A_s$	Attributes	Roles	
value-of	hierarchical relation	role restrictions	
is-a	hierarchical relation	subsumption ($\sqsubseteq$)	
part-of	hierarchical relation	Roles	
any other relation	associative relations	Roles	
$e_1, \ldots, e_n$	entities instances	individuals in F (entities)	ABox
$v_1, \ldots, v_r$	attribute values	individuals in G (values)	
$r_1, \ldots, r_m$	relations between entities	role assertions	
$a_1, \ldots, a_t$	attributes of entities	role assertions	
instance-of	hierarchical relation	concept assertions	

Advantages of DERA

- DERA facets have **explicit semantics** and are modeled as descriptive ontologies
- DERA facets inherits all the important properties of the faceted approach, such as robustness and scalability
- DERA allows for **automated reasoning** via the formalization into Description Logics ontologies. In particular, DERA allows for a very expressive search by any entity property

The space ontology [Giunchiglia et al., 2012]

- Knowledge is extracted from GeoNames and the Getty Thesaurus of Geographic Names
- Terms are collected, categorized into classes, entities, relations and attributes, and synsets are generated
- Synsets are mapped to and integrated with WordNet
- Synsets are analyzed and arranged into facets
- Terms are standardized and ordered

Objects	Quantity
Entity classes (E)	845
Entities (e)	6,907,417
Relations (R)	70
Attributes (A)	31

Landform
- Natural depression
  - Oceanic depression
    - Oceanic valley
    - Oceanic trough
  - Continental depression
    - Trough
    - Valley
- Natural elevation
  - Oceanic elevation
    - Seamount
    - Submarine hill
  - Continental elevation
    - Hill
    - Mountain

Body of water
- Flowing body of water
  - Stream
    - River
    - Brook
- Stagnant body of water
  - Lake
  - Pond

The semantic-geo catalogue [Farazi et al., 2012]

- Knowledge is extracted from the geographical dataset of the Province of Trento
- The faceted ontology was built in English and Italian
- Usage of the ontology
  - The ontology is used in combination with S-Match within the search component of the geo-catalogue to improve search
  - The evaluation shows that at the price of a drop in precision of 0.16% we double recall

Objects	Quantity
Facets	5
Entity classes (E)	39
Entities (e)	20,162
part-of relations	20,161
Alternative names	7,929

Body of water
- Lake
- Group of lakes
- Stream
- River
  - Rivulet
- Spring
- Waterfall
- Cascade
- Canal

Natural elevation
- Highland
- Hill
- Mountain
- Mountain range
  - Peak
- Chain of peaks
- Glacier

Natural depression
- Valley
- Mountain pass

Exercises

1. Analyse the following terms:
   - (geography) river, lake, salt lake, depth
   - (business) organization, company, business
   - (literature) newspaper, newsletter, book, archive, author, publisher, format, frequency
2. Take one domain of your choice, identify the entity types which are relevant and define corresponding terminology using DERA (concentrate on a few classes, relations and attributes).

Some reference material

[Ranganathan, 1967] S. R. Ranganathan, Prolegomena to library classification, Asia Publishing House.

[Gruber, 1993] A translation approach to portable ontology specifications. Knowledge Aquisition, 5 (2), 199–220.

[Pollock, 2002] Integration's Dirty Little Secret: It's a Matter of Semantics. Whitepaper, The Interoperability Company.

[Guarino and Welty, 2002] Guarino, N., Welty, C. (2002). Evaluating ontological decisions with OntoClean. Communications of the ACM, 45(2), 61-65.

[Uschold and Gruninger, 2004] Ontologies and semantics for seamless connectivity. SIGMOD Rec., 33(4), 58–64.

[Varzi, 2006] Varzi, A. (2006). A note on the transitivity of parthood. Applied Ontology, 1 (2), 141-146.

[Giunchiglia et al., 2009] Faceted Lightweight Ontologies. In: Conceptual Modeling: Foundations and Applications, LNCS Springer.

[Giunchiglia et al., 2012a] A facet-based methodology for the construction of a large-scale geospatial ontology. Journal on Data Semantics, 1 (1), pp. 57-73.

[Giunchiglia et al., 2012b] Domains and context: first steps towards managing diversity in knowledge. Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web.

[Giunchiglia et al., 2014] From Knowledge Organization to Knowledge Representation. Knowledge Organization. 41(1), 44-56.

[Tawfik et al., 2014] A Collaborative Platform for Multilingual Ontology Development. International Conference on Knowledge Engineering and Ontology.

[Ganbold et. al., 2014] An Experiment in Managing Language Diversity Across cultures. eKNOW 2014