UNIVERSITY
OF TRENTO - Italy

KNOW
DIVE

Linguistic and Knowledge Resources

Vincenzo Maltese
PhD, University ofTrento

ESSENCE AUTUMN SCHOOL

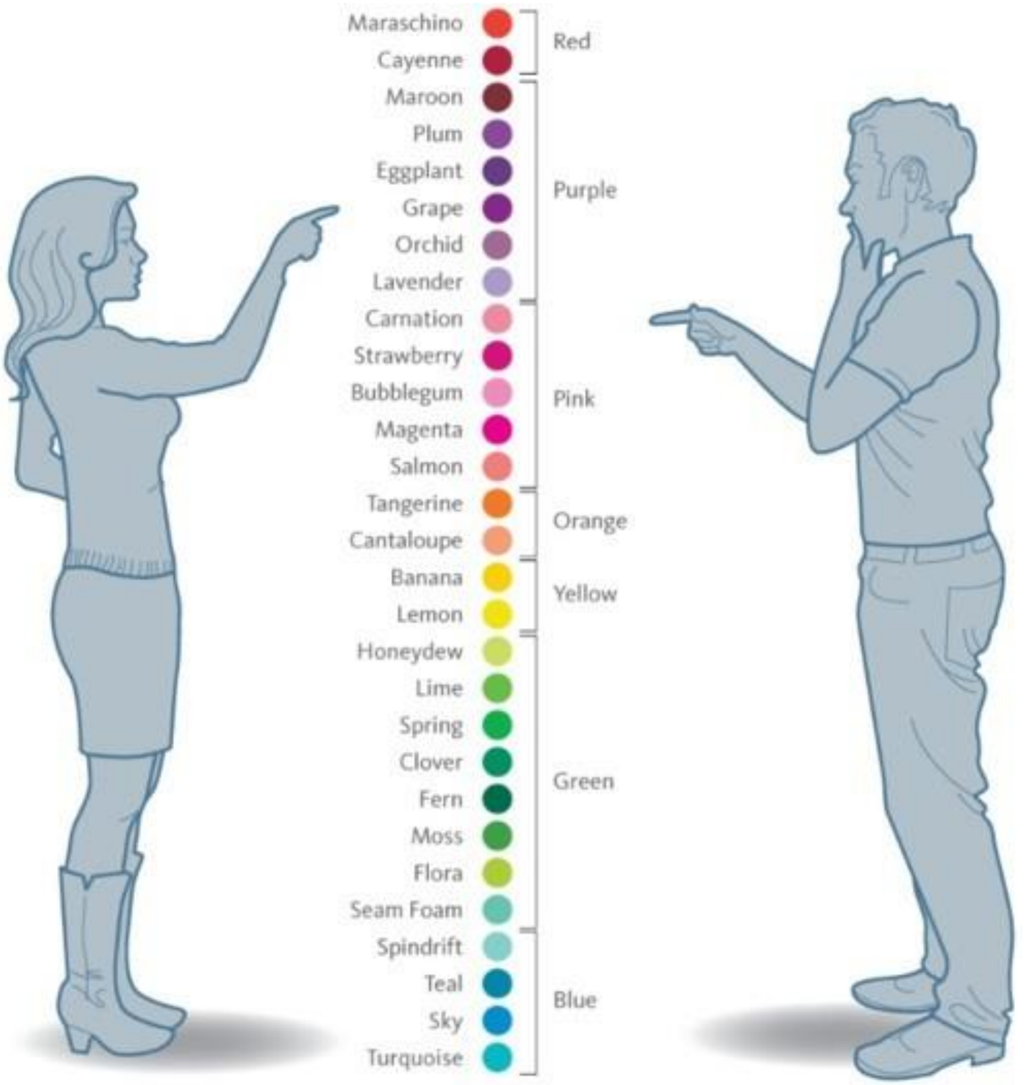
Ischia, Italy - October 2014

Roadmap

- Motivation and use-cases
- Solutions to interoperability
- Linguistic and knowledge resources
- Our approach in Trento
- Methodologies for content generation

Motivation and use-cases

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]

Use-cases

SEMANTIC SEARCH

SEARCH: automobile

1957 Ferrari 625 TRC Spider

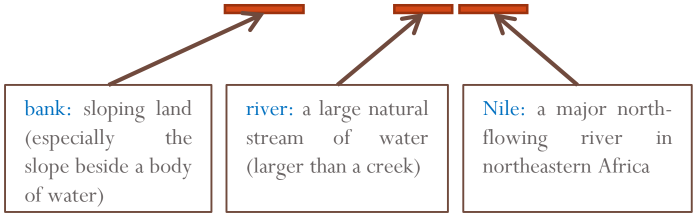
This two-of-a-kind classic Ferrari is lauded by historians as one of the prettiest Ferraris ever built. The 1957 Ferrari 625 TRC Spider is an absolutely stunning automobile, one as dashing in the garage as it is at 120 mph.

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.

NLP

The banks of the river Nile

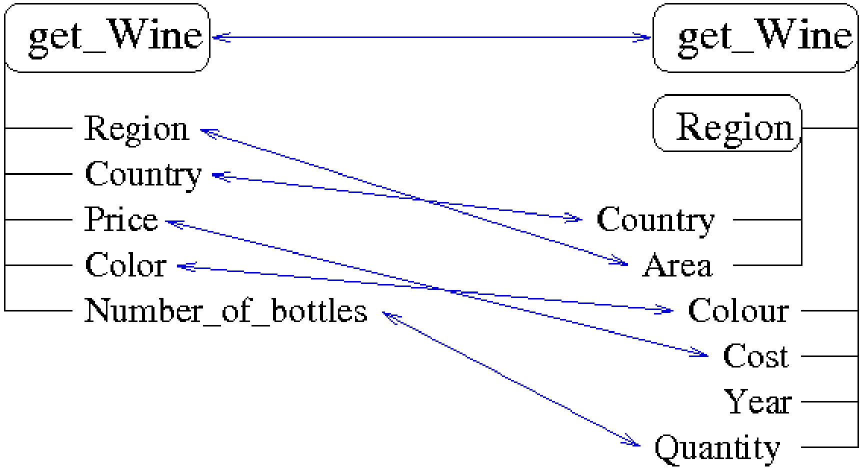


bank: sloping land (especially the slope beside a body of water)

river: a large natural stream of water (larger than a creek)

Nile: a major north-flowing river in northeastern Africa

SEMANTIC MATCHING



get_Wine

Region

Country

Price

Color

Number_of_bottles

get_Wine

Region

Country

Area

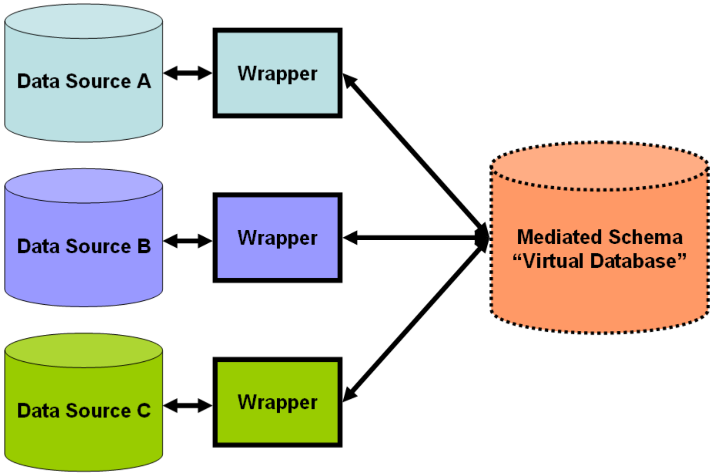
Colour

Cost

Year

Quantity

DATA INTEGRATION



Data Source A

Data Source B

Data Source C

Wrapper

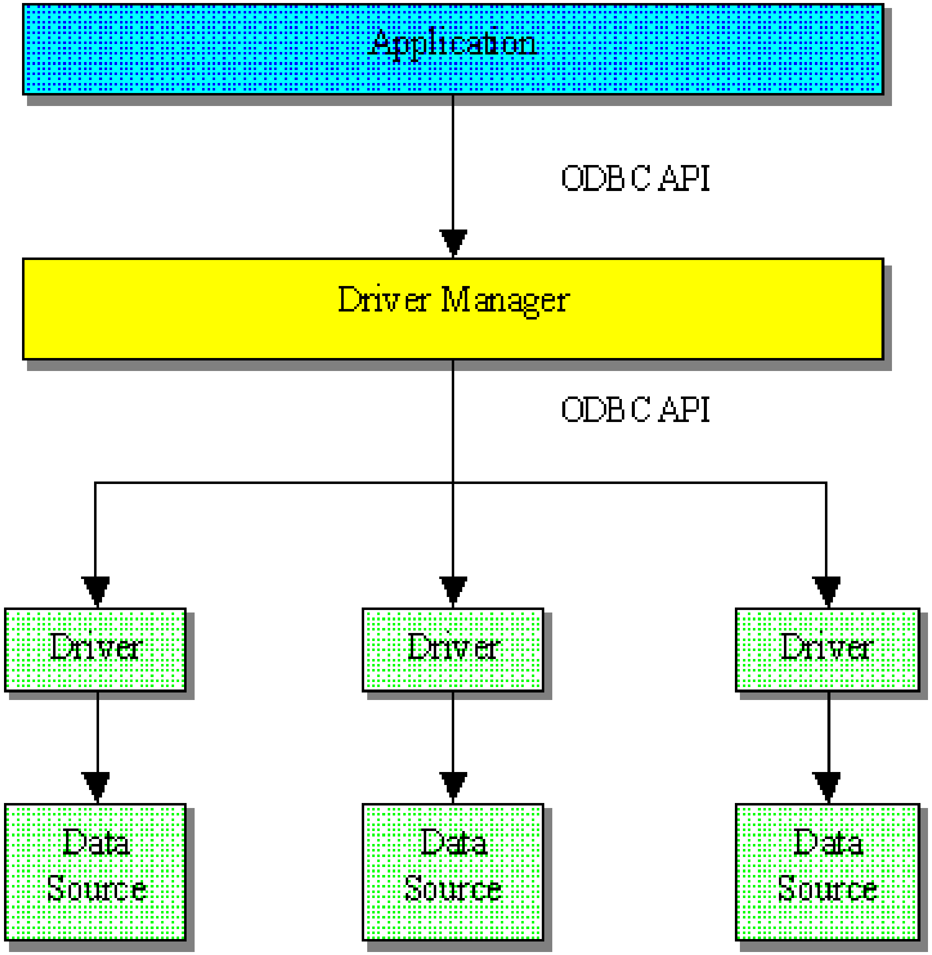
Wrapper

Wrapper

Mediated Schema
“Virtual Database”

Solutions to interoperability

Early solutions

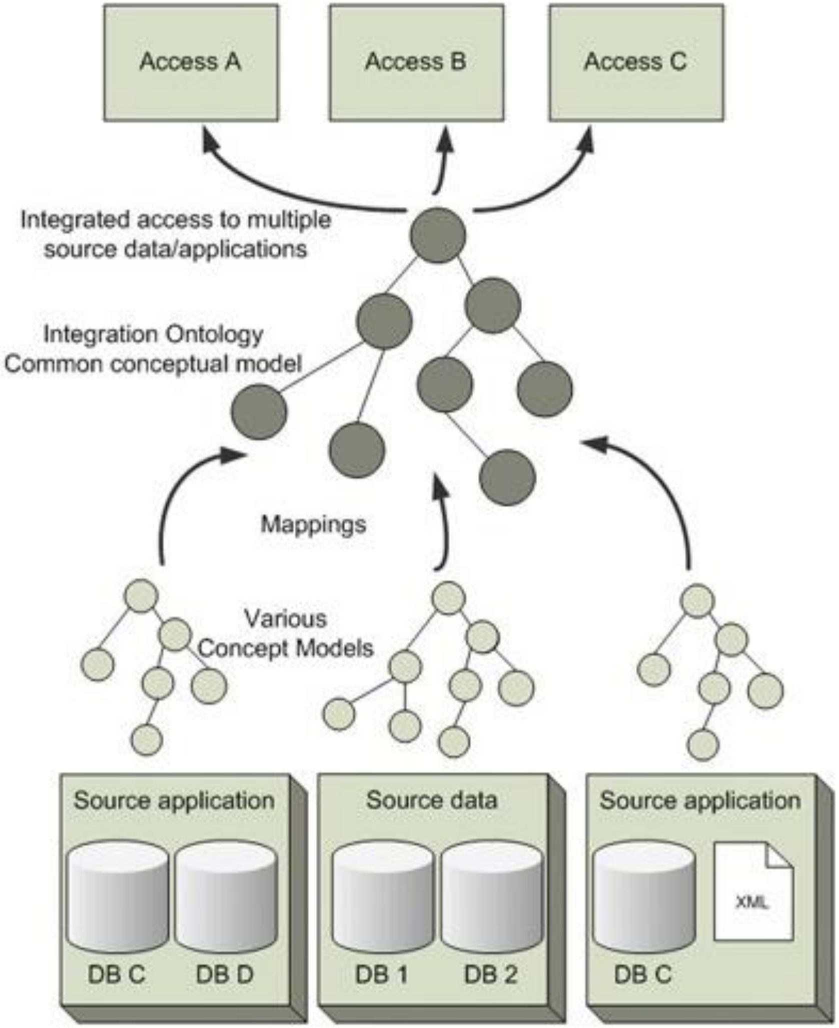


Physical connectivity relies on the presence of a stable communication channel between the parties, for instance ODBC data gateways and software adapters.

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:

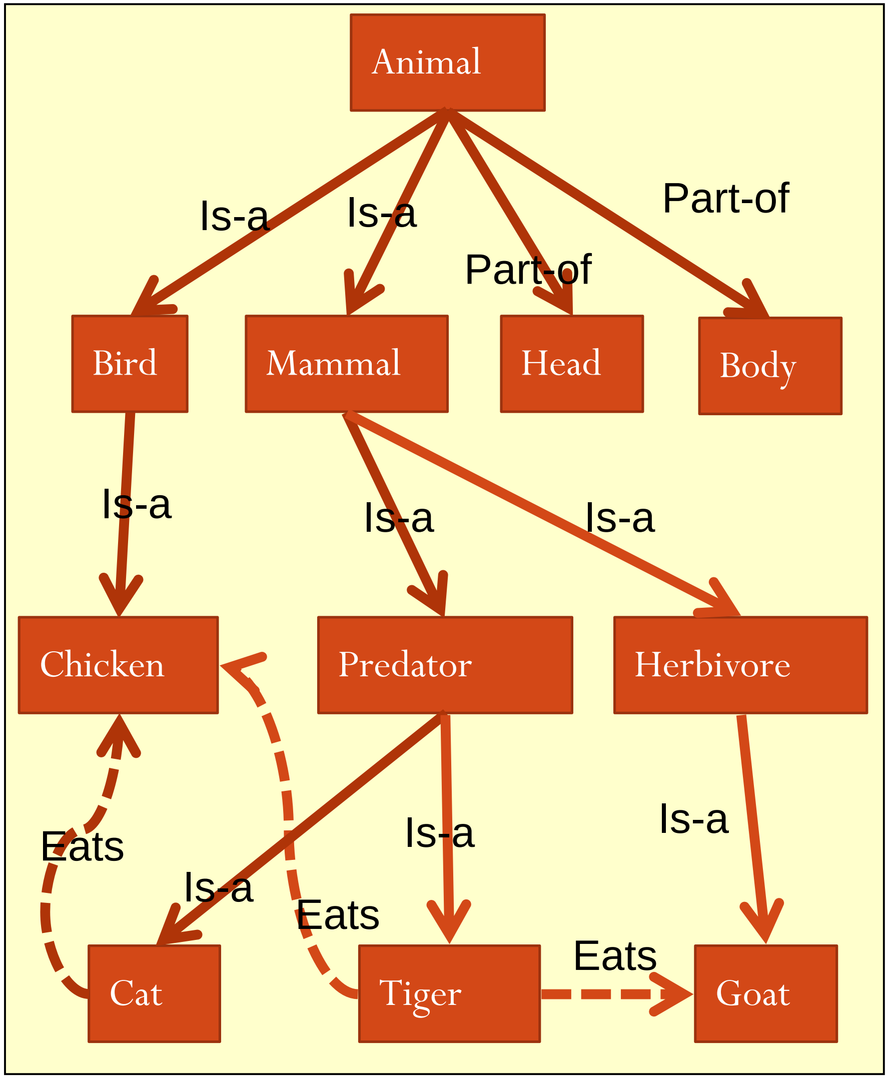
Semantic mediation: the usage of an ontology, providing a shared vocabulary of terms with explicit meaning.

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.

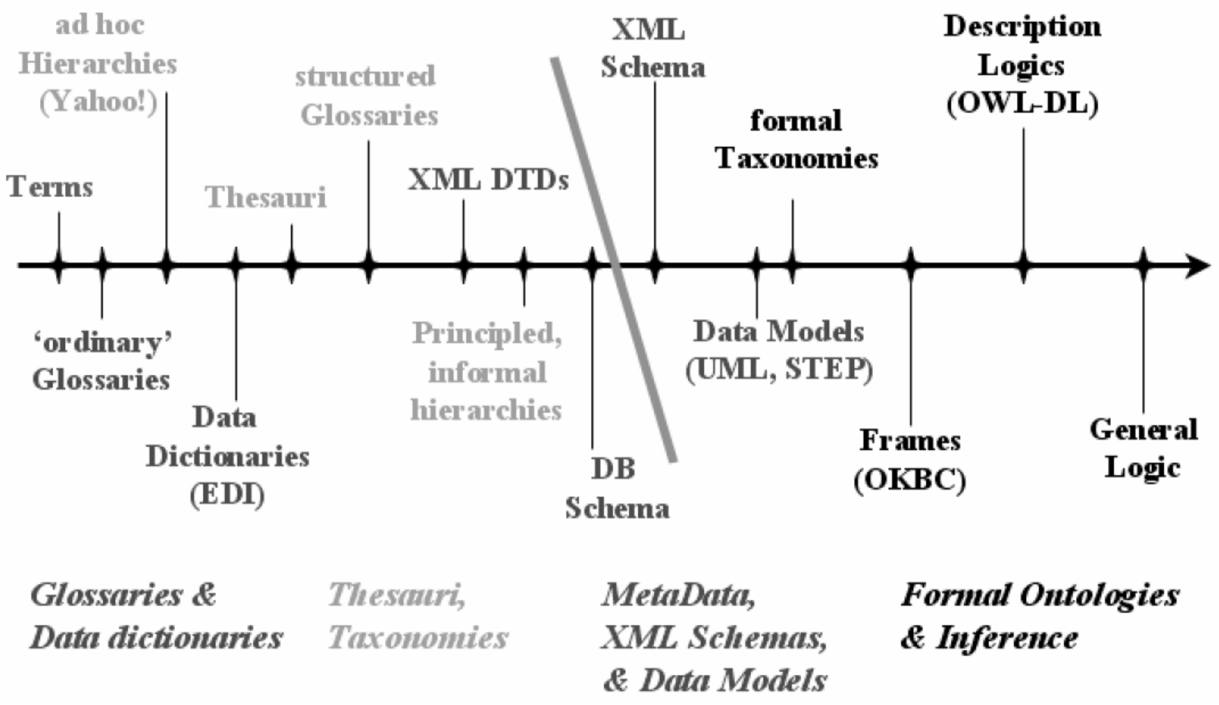
Context sensitivity: the mapping has *contextual validity*, i.e. it has to be used by taking into account the conditions and the purposes for which it was generated.

Ontologies

- An ontology is an explicit specification of a shared conceptualization [Gruber, 1993]
- Ontologies are often thought of as directed graphs whose nodes represent concepts and whose edges represent relations between concepts
- By providing a common formal terminology and understanding of a given domain of interest, it allows for automation (logical inference), supports reuse and favor interoperability across applications and people.
- They differ according to the purpose and the semantics



Kinds of ontologies

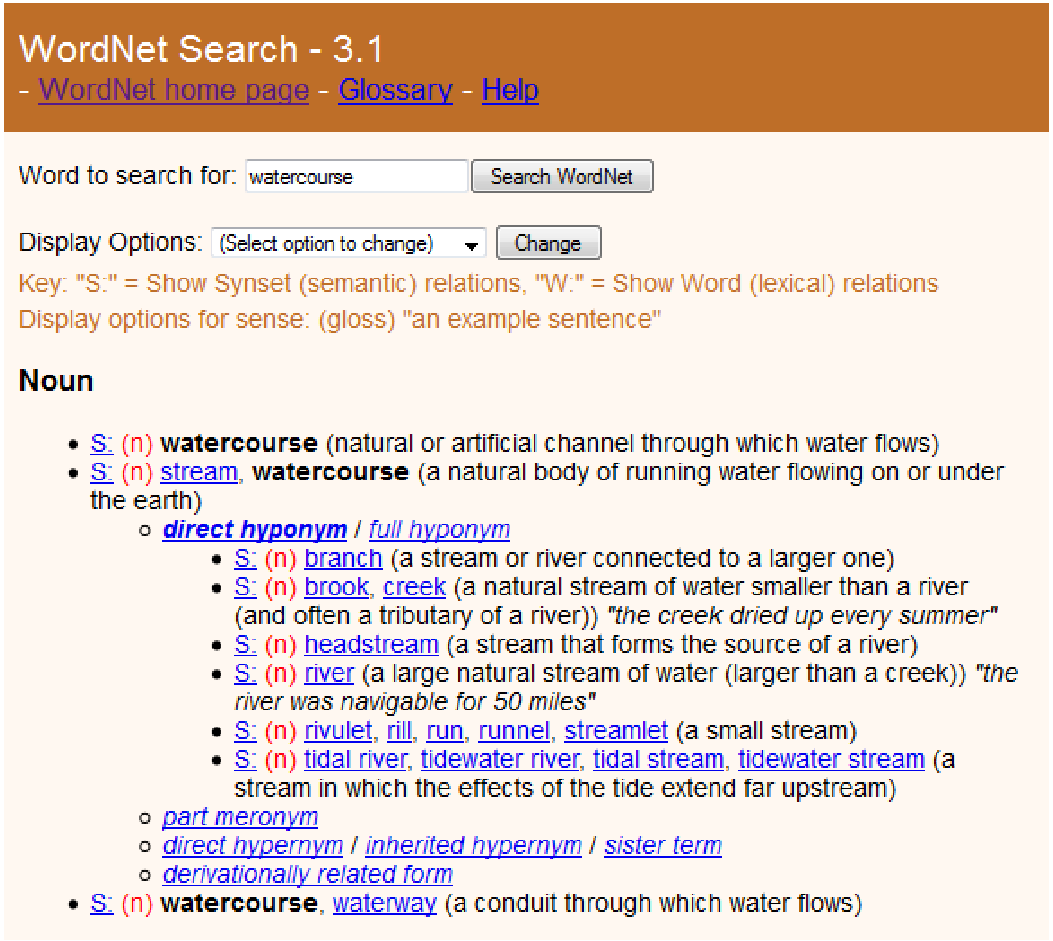


[Uschold and Gruninger, 2004]

- Informal representations
  - User classification
  - Web directories
  - Business catalogs
- Progressive formal
  - Enumerative (e.g. DDC)
  - Knowledge Organization systems
  - Faceted Classification systems
- Formal ontologies
  - Expressed into a formal logic language and represented using formal specifications, such as, OWL)

Linguistic resources

WordNet (1985)



WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: watercourse [Search WordNet]

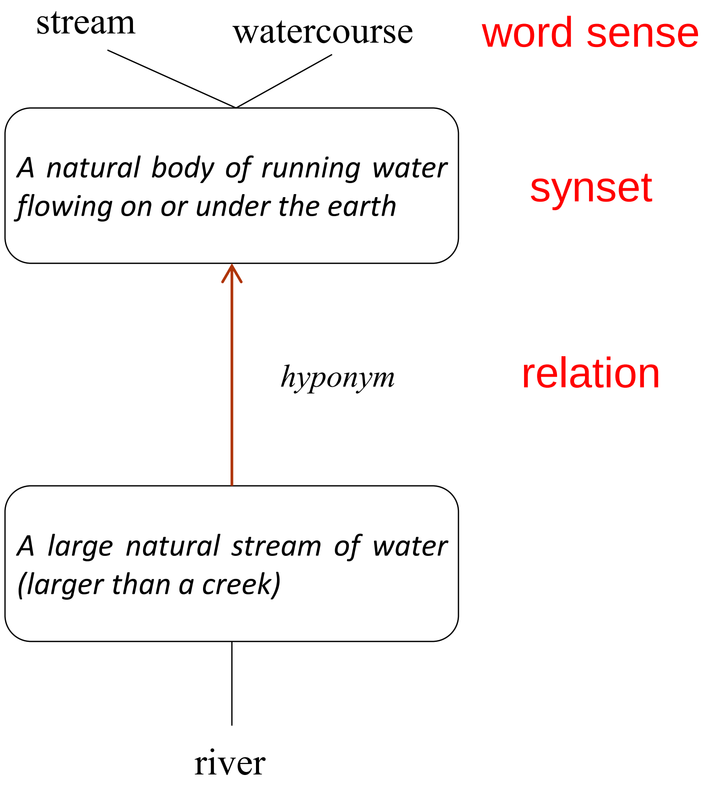
Display Options: (Select option to change) [Change]

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)



stream

watercourse

word sense

A natural body of running water flowing on or under the earth

synset

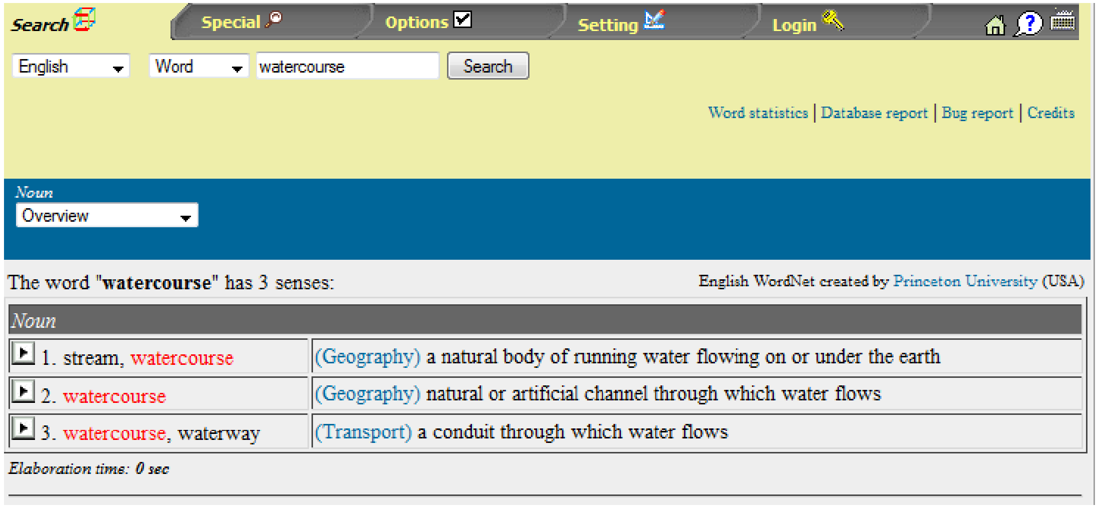
hyponym

relation

A large natural stream of water (larger than a creek)

river

MultiWordNet (2002)

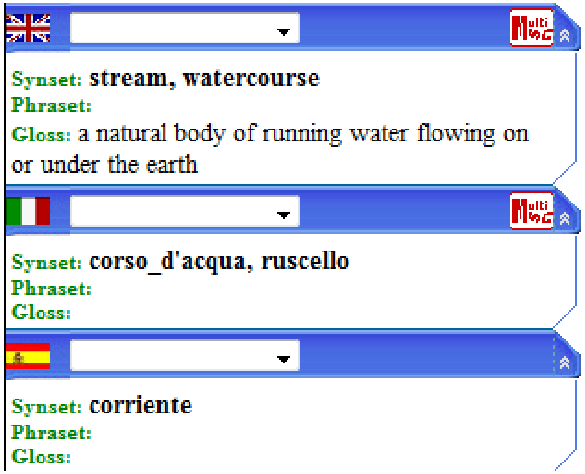


The word "**watercourse**" has 3 senses:

English WordNet created by Princeton University (USA)

Noun	
1. stream, watercourse	(Geography) a natural body of running water flowing on or under the earth
2. watercourse	(Geography) natural or artificial channel through which water flows
3. watercourse, waterway	(Transport) a conduit through which water flows

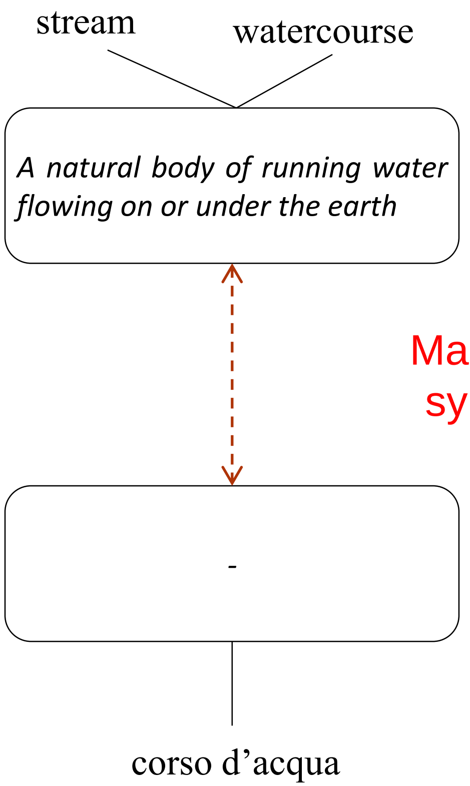
Elaboration time: 0 sec



Synset: **stream, watercourse**
Phraset:
Gloss: a natural body of running water flowing on or under the earth

Synset: **corso_d'acqua, ruscello**
Phraset:
Gloss:

Synset: **corriente**
Phraset:
Gloss:


stream
watercourse
A natural body of running water flowing on or under the earth
-
corso d'acqua

Mapping via synset IDs

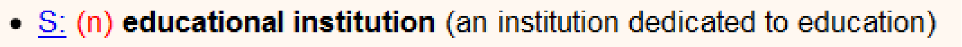
Strengths

- Mapping with 6 languages
- Lexical GAPs can be defined

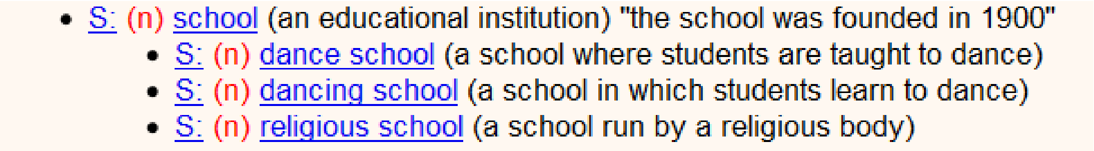
Weaknesses

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

Problems with WordNet-like resources

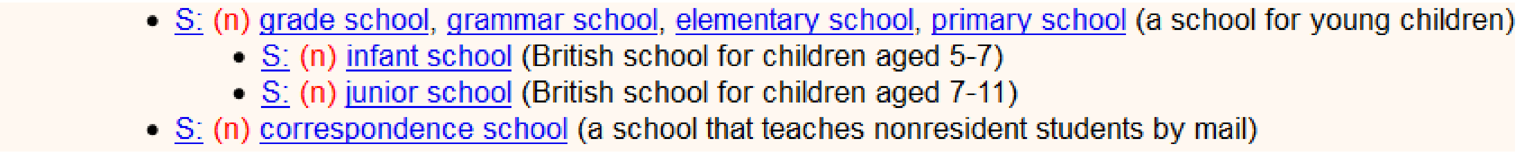


- S: (n) **educational institution** (an institution dedicated to education)



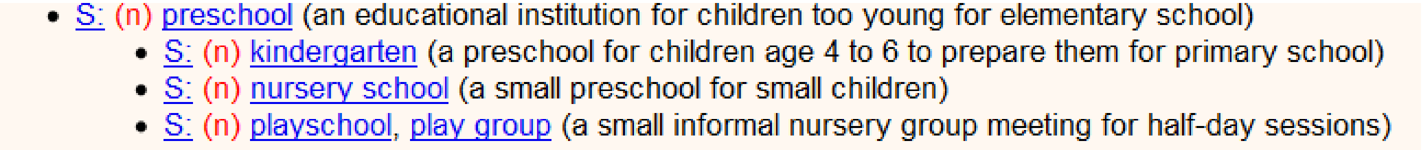
  - S: (n) school (an educational institution) "the school was founded in 1900"
    - S: (n) dance school (a school where students are taught to dance)
    - S: (n) dancing school (a school in which students learn to dance)
    - S: (n) religious school (a school run by a religious body)

Nodes in similar position do not share same ontological properties



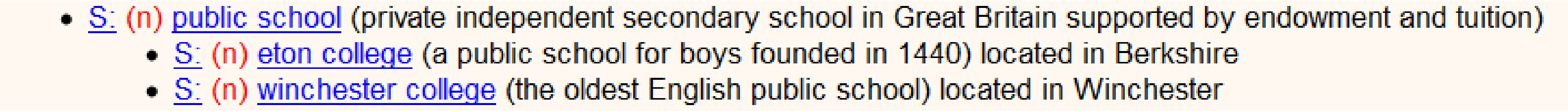
- S: (n) grade school, grammar school, elementary school, primary school (a school for young children)
  - S: (n) infant school (British school for children aged 5-7)
  - S: (n) junior school (British school for children aged 7-11)
- S: (n) correspondence school (a school that teaches nonresident students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

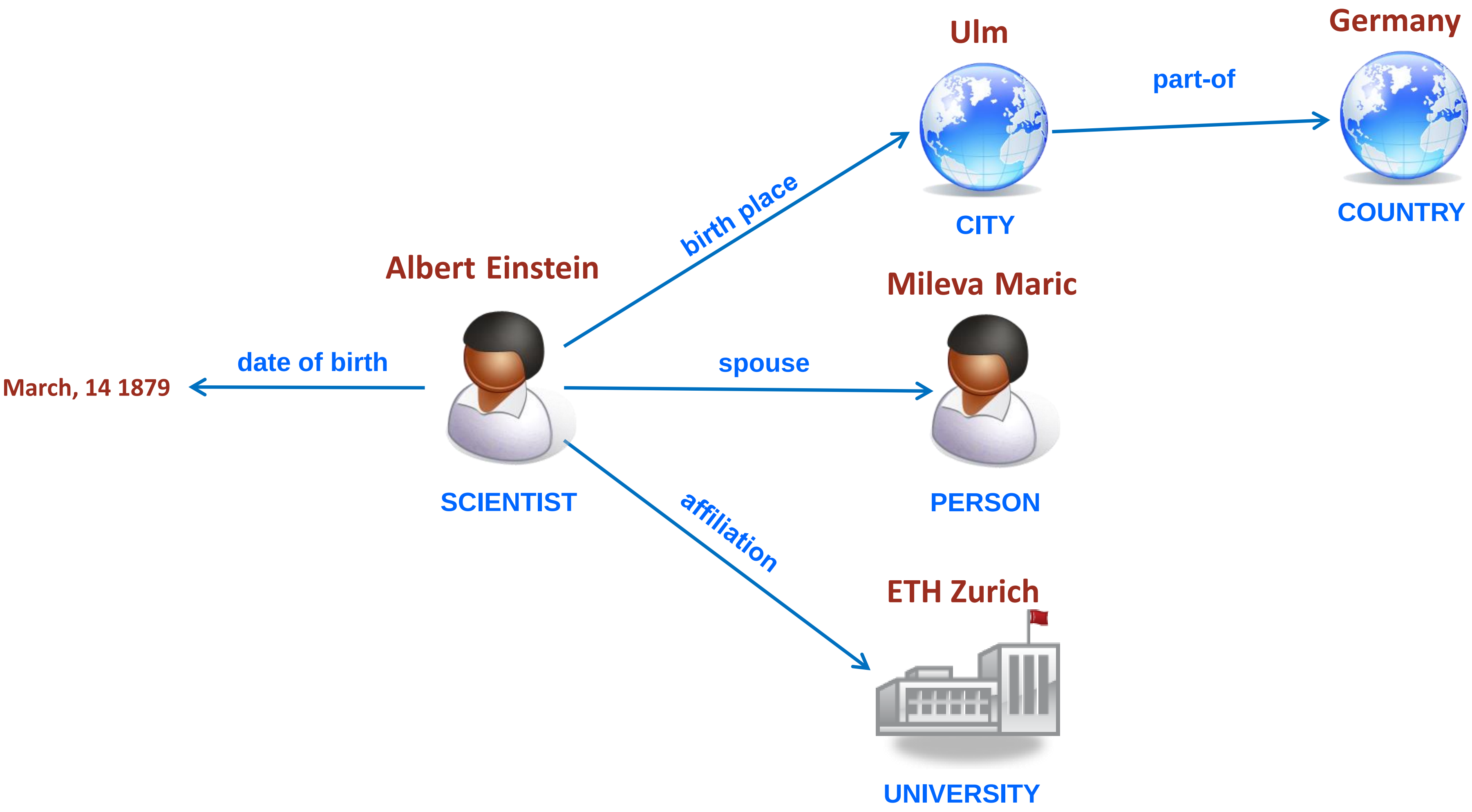


- S: (n) public school (private independent secondary school in Great Britain supported by endowment and tuition)
  - S: (n) eton college (a public school for boys founded in 1440) located in Berkshire
  - S: (n) winchester college (the oldest English public school) located in Winchester

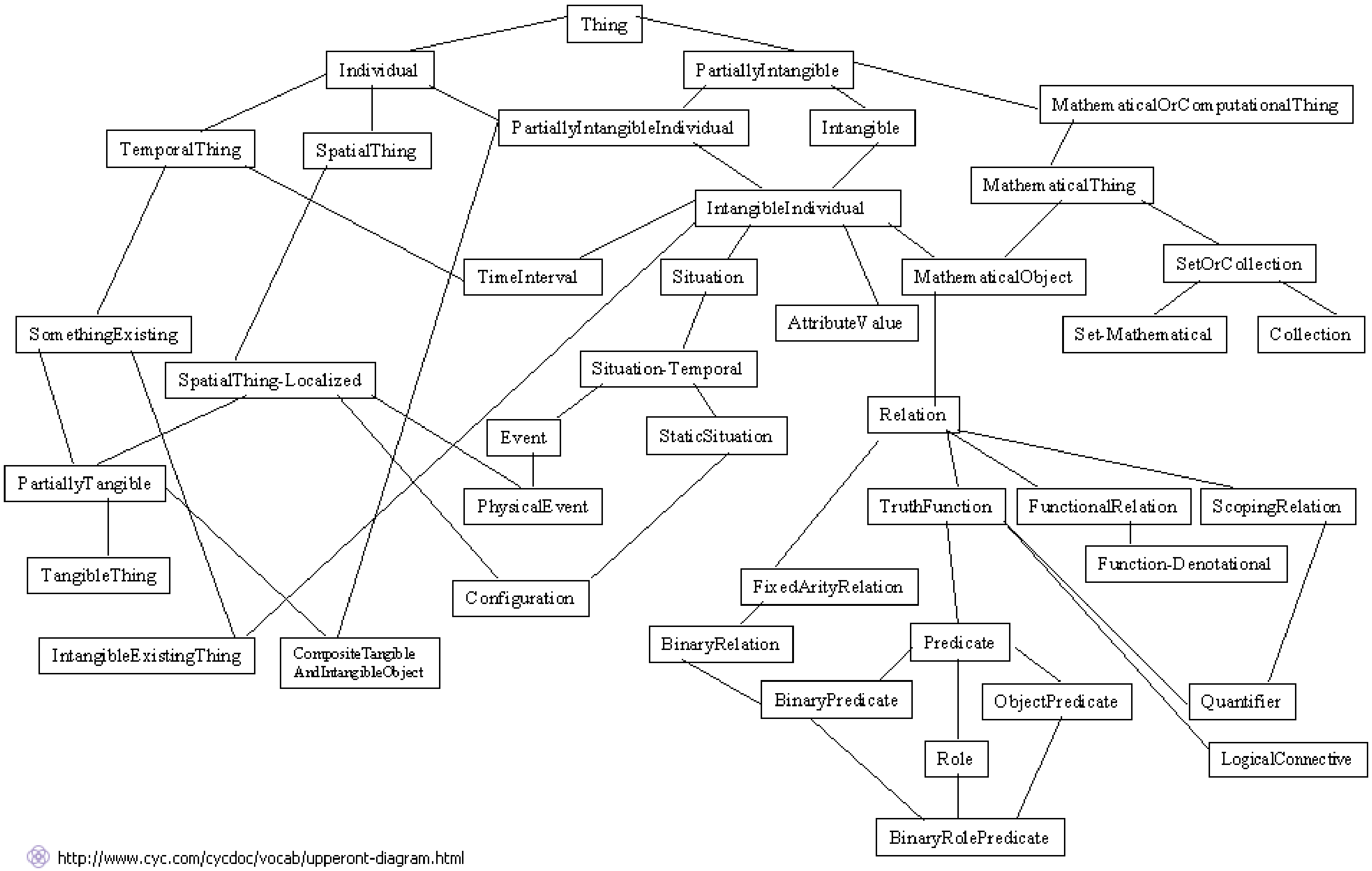
Some concepts are actually individuals

Knowledge resources

Example of content



CYC ontology (1984)



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

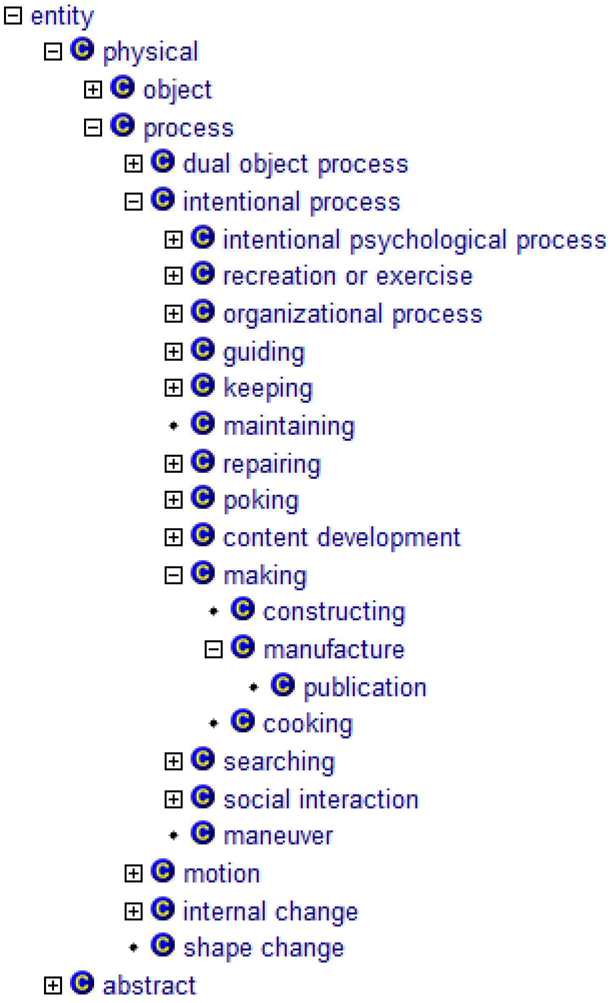
Triples such as:

#$isa
#$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)

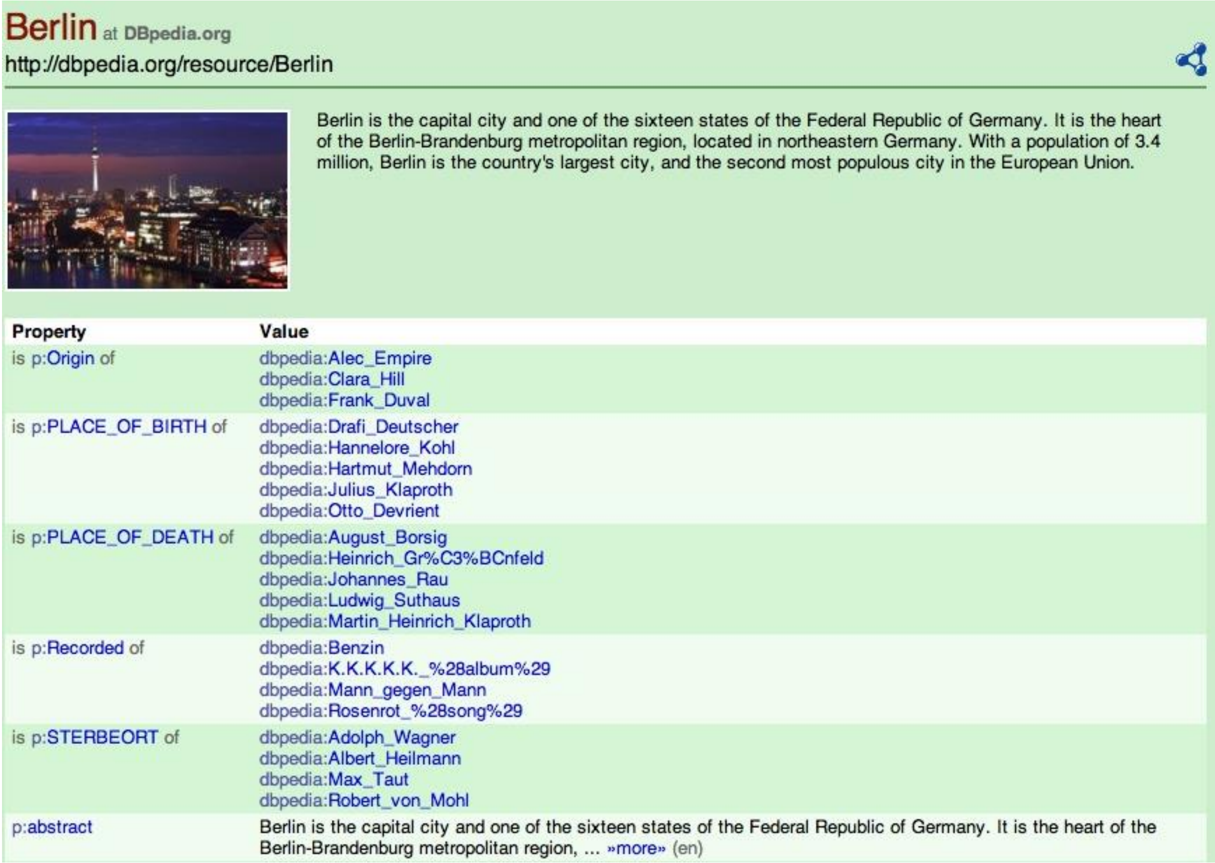


- ⊟ entity
  - ⊟ physical
    - ⊞ object
    - ⊟ process
      - ⊞ dual object process
      - ⊟ intentional process
        - ⊞ intentional psychological process
        - ⊞ recreation or exercise
        - ⊞ organizational process
        - ⊞ guiding
        - ⊞ keeping
        - maintaining
        - ⊞ 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)



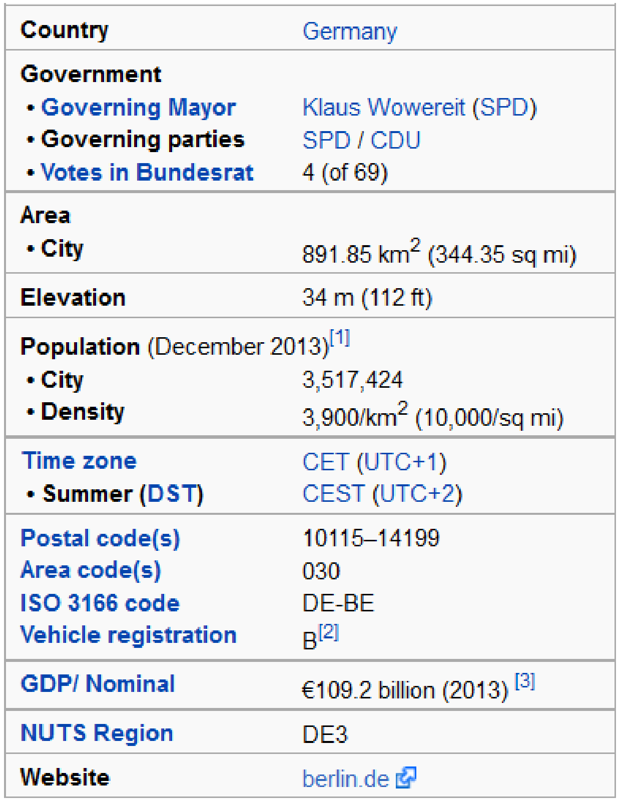
Berlin at DBpedia.org

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)

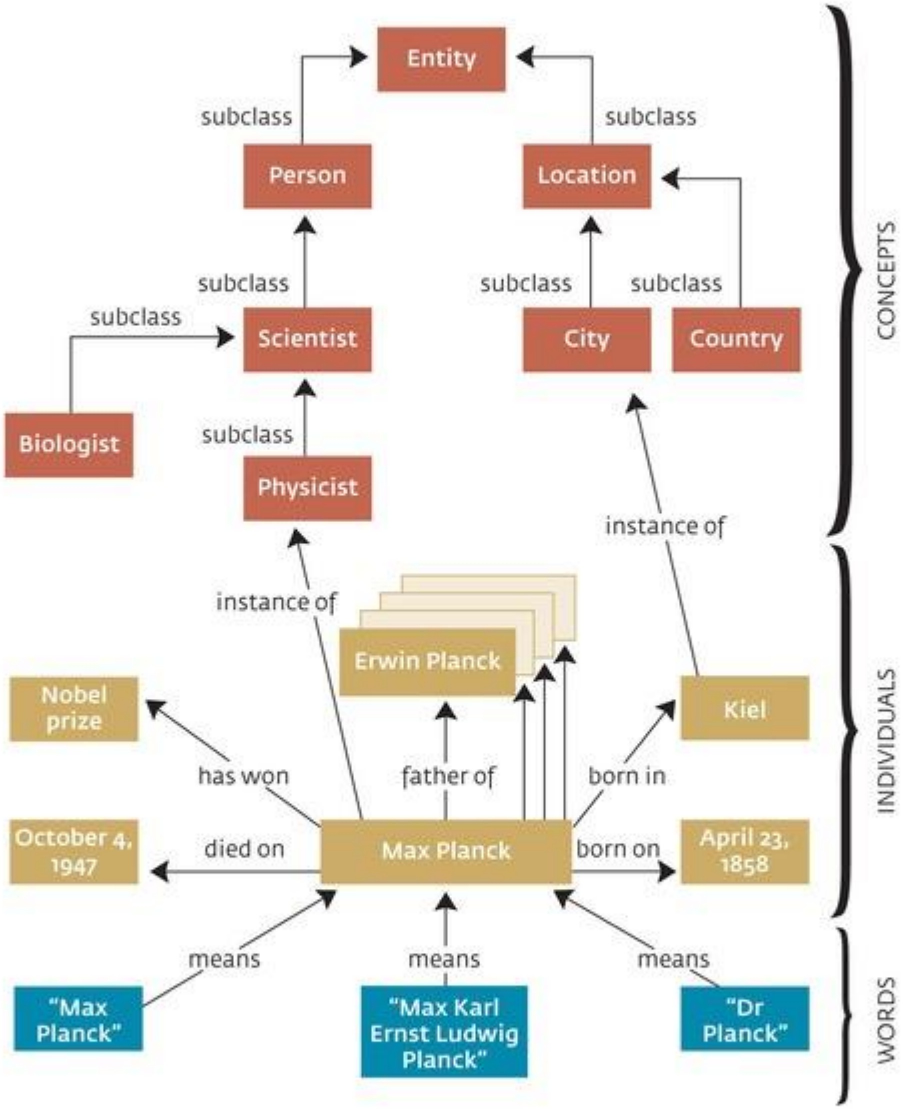
Wikipedia

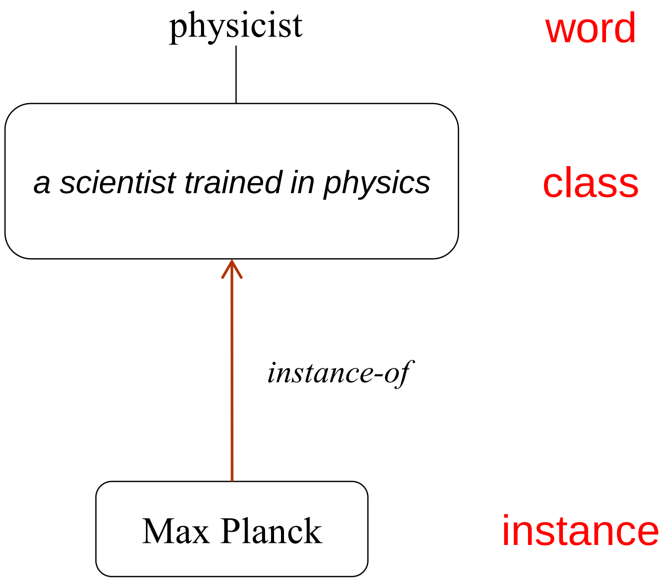


Country	Germany
Government	
• **Governing Mayor**	Klaus Wowereit (SPD)
• **Governing parties**	SPD / CDU
• **Votes in Bundesrat**	4 (of 69)
Area	
• **City**	891.85 km^2 (344.35 sq mi)
Elevation	34 m (112 ft)
Population (December 2013)[1]	
• **City**	3,517,424
• **Density**	3,900/km^2 (10,000/sq mi)
Time 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)

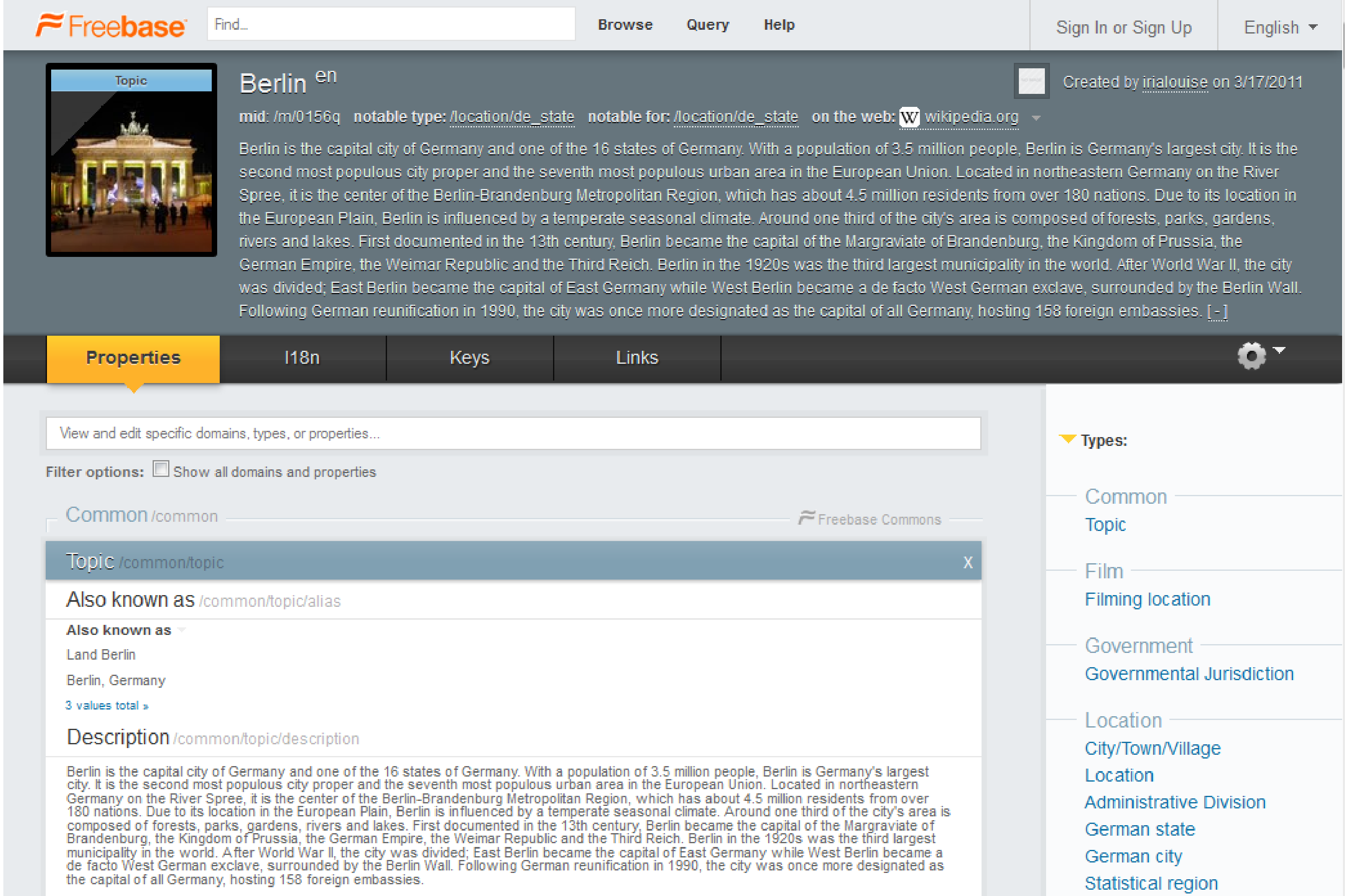




- **Concepts** are taken from noun synsets of WordNet
- **Instances** and their properties are automatically extracted from Wikipedia
- The **linking** of concepts with instances is done via NLP techniques

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

Freebase (2010)



Freebase | Find... | Browse | Query | Help | Sign In or Sign Up | English

Topic

Berlin en

Created by irialouise on 3/17/2011

mid: /m/0156q notable type: /location/de_state notable for: /location/de_state on the web: wikipedia.org

Berlin is the capital city of Germany and one of the 16 states of Germany. With a population of 3.5 million people, Berlin is Germany's largest city. It is the second most populous city proper and the seventh most populous urban area in the European Union. Located in northeastern Germany on the River Spree, it is the center of the Berlin-Brandenburg Metropolitan Region, which has about 4.5 million residents from over 180 nations. Due to its location in the European Plain, Berlin is influenced by a temperate seasonal climate. Around one third of the city's area is composed of forests, parks, gardens, rivers and lakes. First documented in the 13th century, Berlin became the capital of the Margraviate of Brandenburg, the Kingdom of Prussia, the German Empire, the Weimar Republic and the Third Reich. Berlin in the 1920s was the third largest municipality in the world. After World War II, the city was divided; East Berlin became the capital of East Germany while West Berlin became a de facto West German exclave, surrounded by the Berlin Wall. Following German reunification in 1990, the city was once more designated as the capital of all Germany, hosting 158 foreign embassies. [-]

Properties | I18n | Keys | Links

View and edit specific domains, types, or properties...

Filter options: ☐ Show all domains and properties

Common /common — Freebase Commons

Topic /common/topic x

Also known as /common/topic/alias

Also known as

Land Berlin

Berlin, Germany

3 values total »

Description /common/topic/description

Berlin is the capital city of Germany and one of the 16 states of Germany. With a population of 3.5 million people, Berlin is Germany's largest city. It is the second most populous city proper and the seventh most populous urban area in the European Union. Located in northeastern Germany on the River Spree, it is the center of the Berlin-Brandenburg Metropolitan Region, which has about 4.5 million residents from over 180 nations. Due to its location in the European Plain, Berlin is influenced by a temperate seasonal climate. Around one third of the city's area is composed of forests, parks, gardens, rivers and lakes. First documented in the 13th century, Berlin became the capital of the Margraviate of Brandenburg, the Kingdom of Prussia, the German Empire, the Weimar Republic and the Third Reich. Berlin in the 1920s was the third largest municipality in the world. After World War II, the city was divided; East Berlin became the capital of East Germany while West Berlin became a de facto West German exclave, surrounded by the Berlin Wall. Following German reunification in 1990, the city was once more designated as the capital of all Germany, hosting 158 foreign embassies.

Types:

Common
- Topic

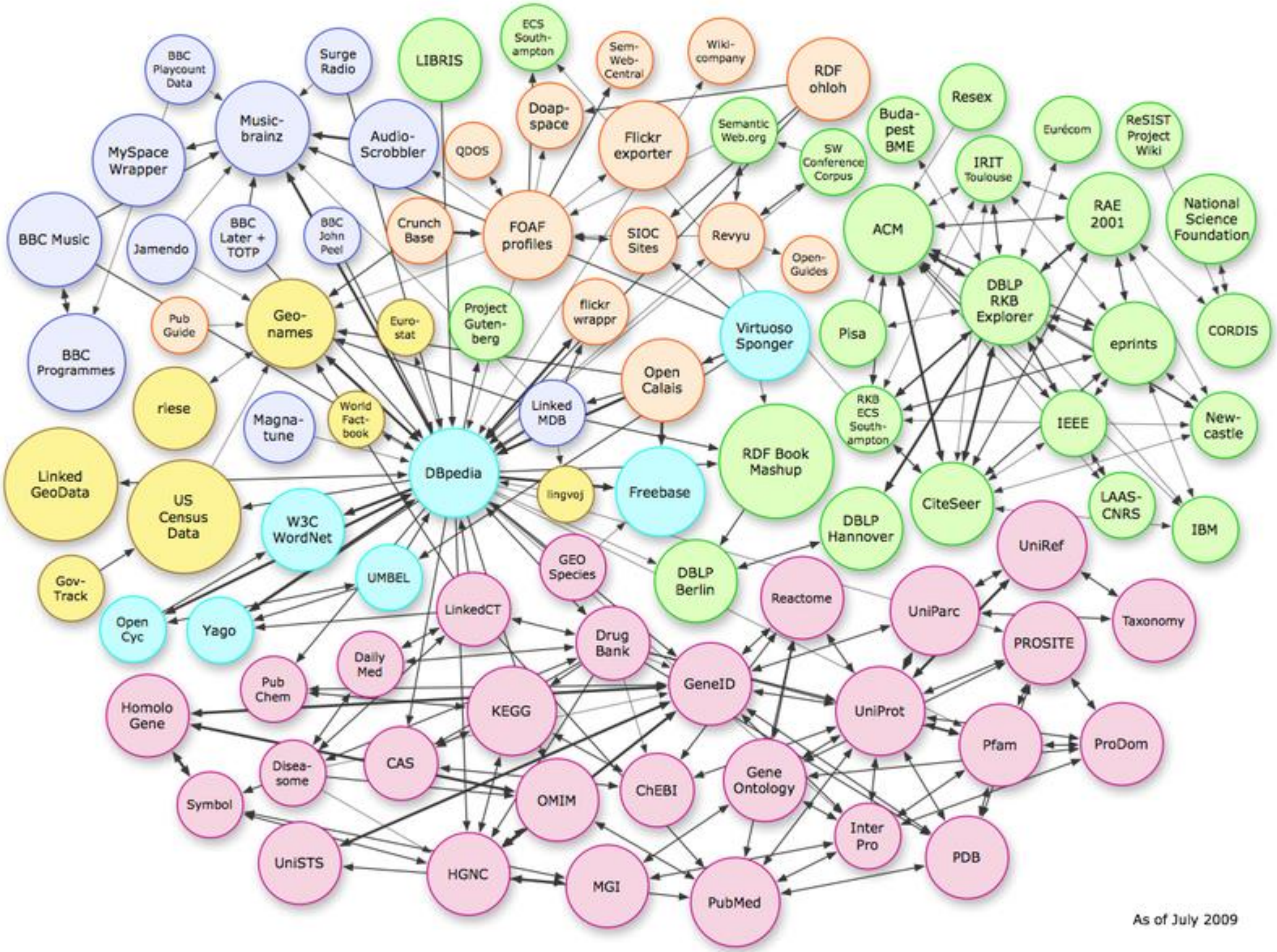
Film
- Filming location

Government
- Governmental Jurisdiction

Location
- City/Town/Village
- Location
- Administrative Division
- German state
- German city
- Statistical region

- 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.

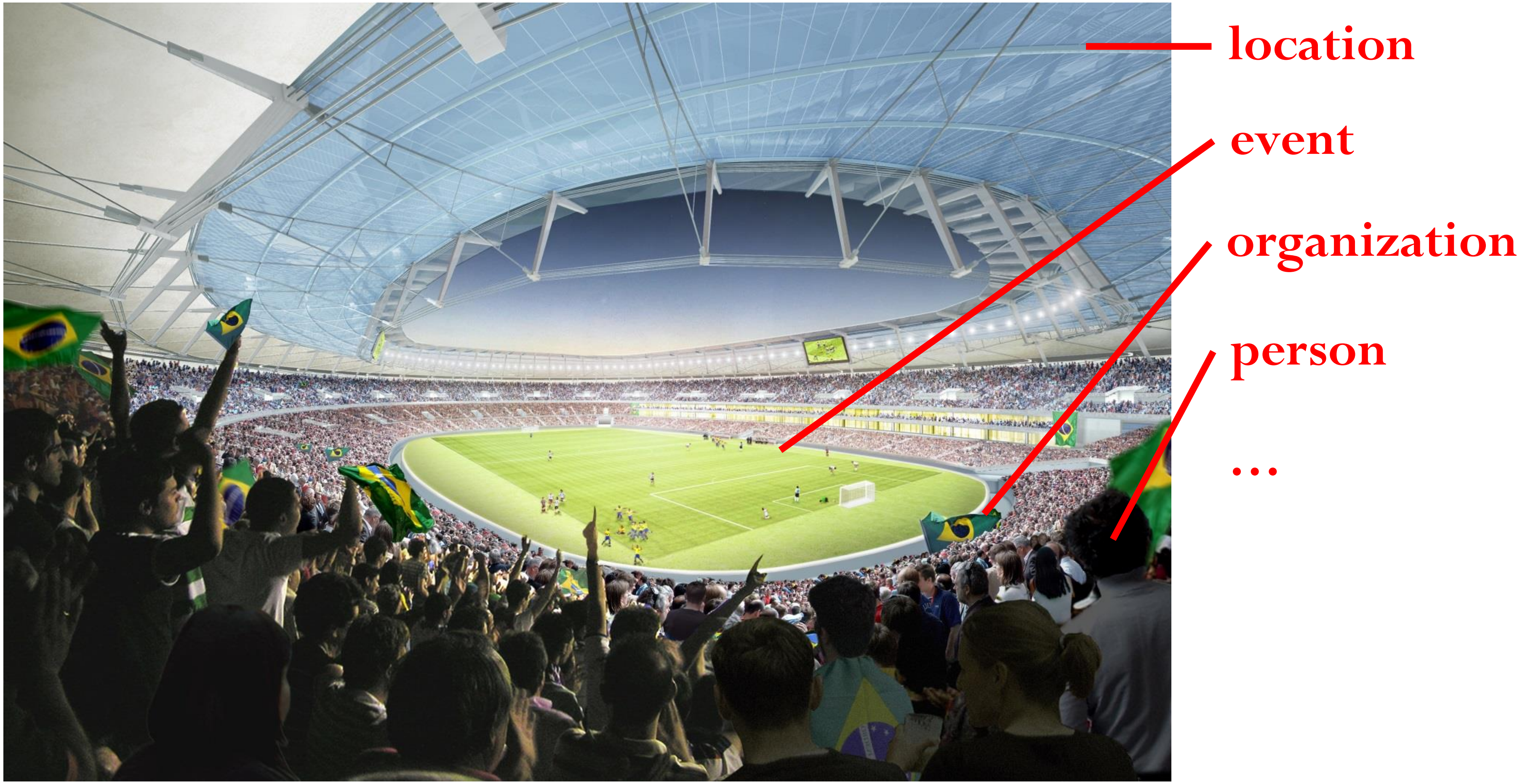
Linked Data Cloud (since 2007)



As of July 2009

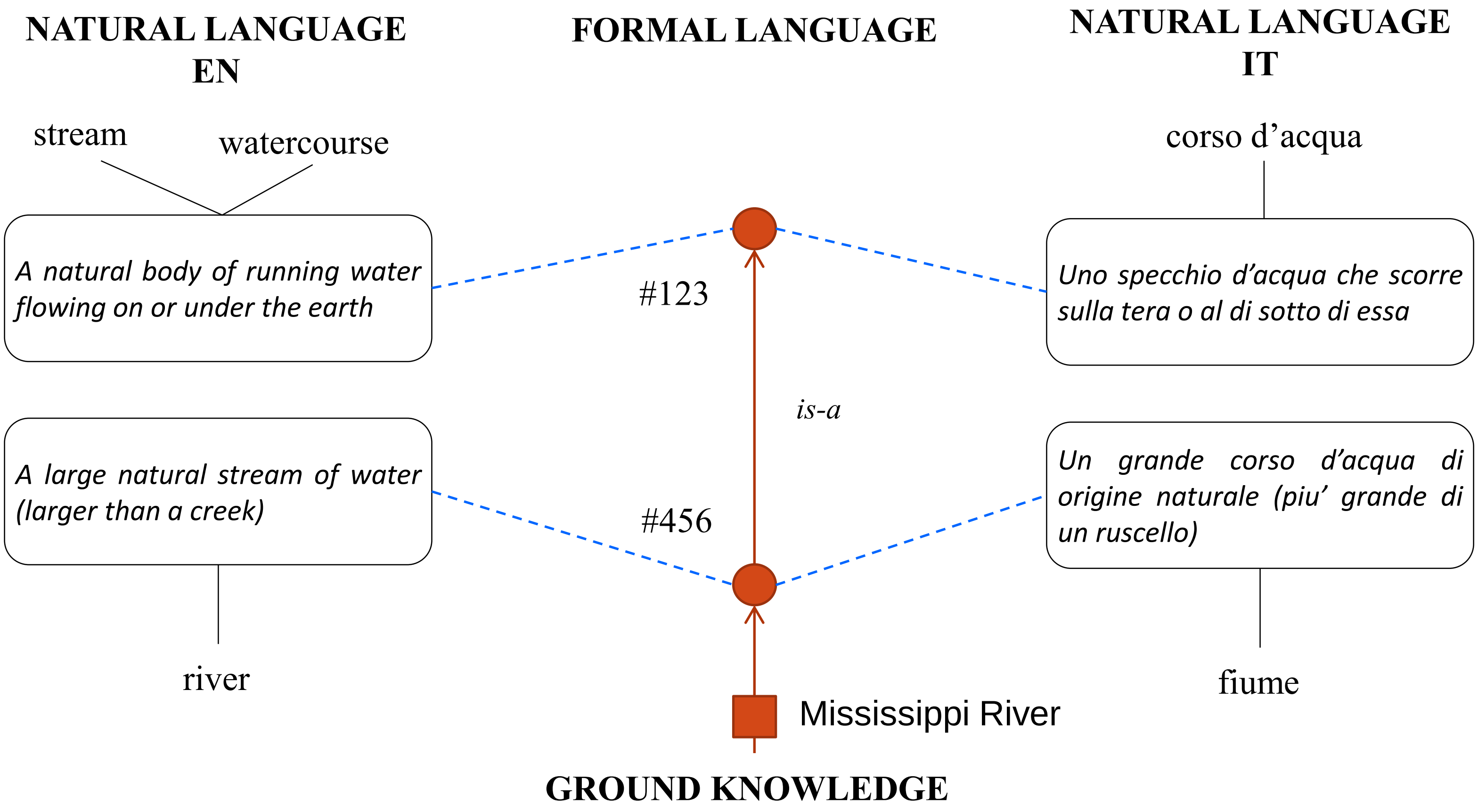
Our approach

The entity-centric view of the world



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

The UKC and Entitypedia (since 2010)



NATURAL LANGUAGE EN

FORMAL LANGUAGE

NATURAL LANGUAGE IT

stream

watercourse

corso d'acqua

A natural body of running water flowing on or under the earth

#123

Uno specchio d'acqua che scorre sulla tera o al di sotto di essa

is-a

A large natural stream of water (larger than a creek)

#456

Un grande corso d'acqua di origine naturale (piu' grande di un ruscello)

river

fiume

Mississippi River

GROUND KNOWLEDGE

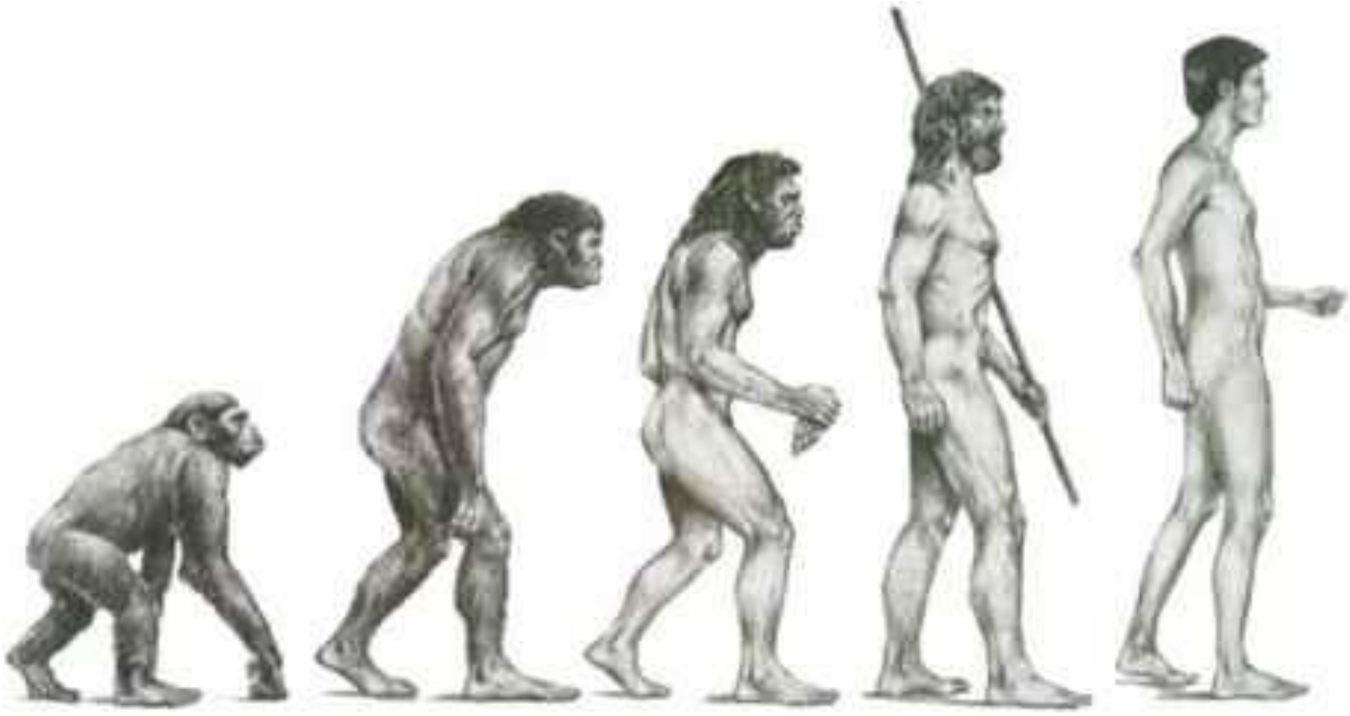
- Manually built via collaborative development [Tawfik et al., 2014], bootstrapped from WordNet, MultiWordNet, GeoNames
- Split natural language, formal language and ground knowledge [Giunchiglia et al., 2012b]
- Domain knowledge is created following the DERA methodology [Giunchiglia et al., 2012a] and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

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

Methodologies for content generation

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?

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.

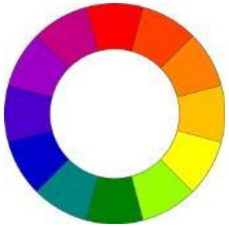
Primitive notions

- **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.



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 (*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*)

Some reference material

[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.

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

[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.

Thank you!
Questions?