Building and using ontologies

Elena Simperl, University of Southampton, UK

With contributions from "Linked Data: Survey of Adoption", Tutorial at the 3rd Asian Semantic Web School ASWS 2011, Incheon, South Korea, July 2011 by Aidan Hogan, DERI, IE

Ontologies in Computer Science

EUCLID
EdUcational Curriculum for the usage of LInked Data

- An ontology defines a **domain of interest**
  - … in terms of the **things** you talk about in the domain, their **attributes,** as well as **relationships** between them
- Ontologies are used to
  - Share a **common understanding** about a domain among people and machines
  - Enable **reuse** of domain knowledge

ontology ***vocabulary***

microformat ***conceptual graph***

topic map *thesaurus*

schema

classification object model

semantic network

glossary **taxonomy**

class relationship *entity* type

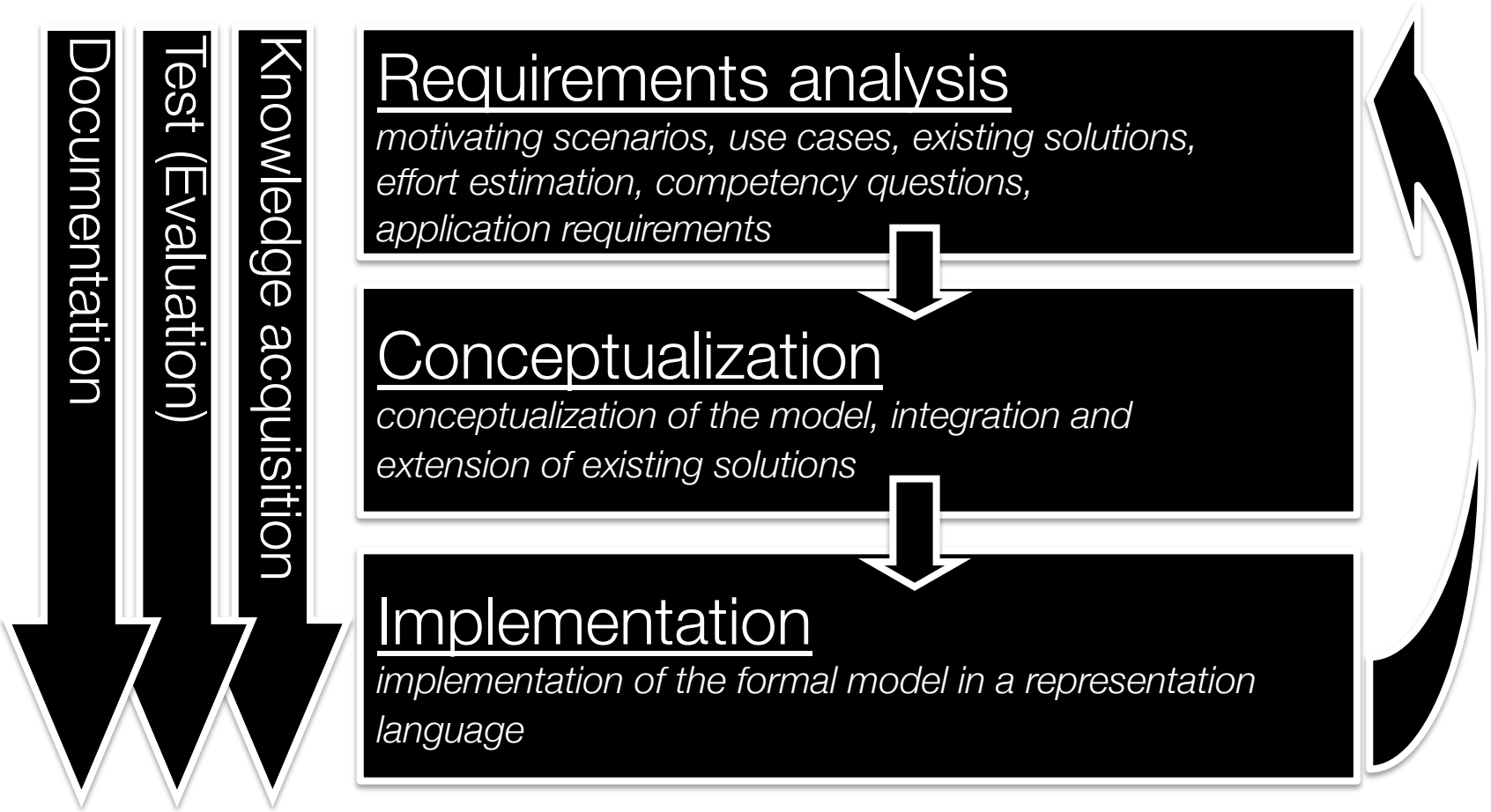
slot ***particular*** object ***instance***

universal **attribute** association

individual *property* **role**

Classical ontology engineering process

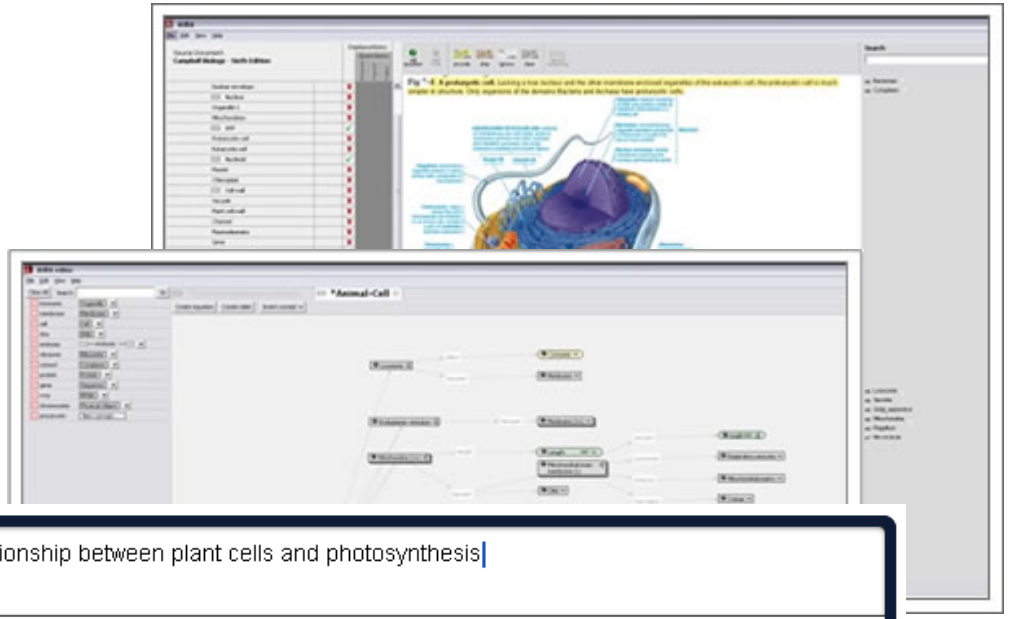
EUCLID
EdUcational Curriculum for the usage of LInked Data

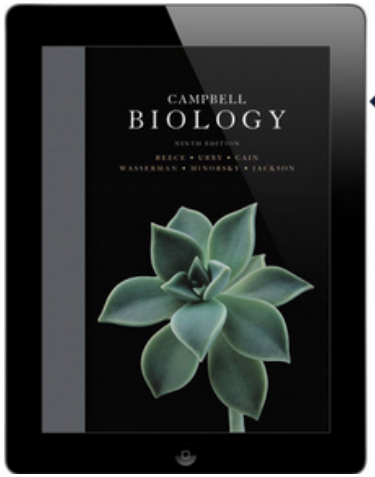


Example: Project Halo

EUCLID
EdUcational Curriculum for the usage of LInked Data

- Knowledge acquisition from text(books)
- Professional and crowdsourced annotation
- Question analysis and answering through a combination of NLP and reasoning techniques





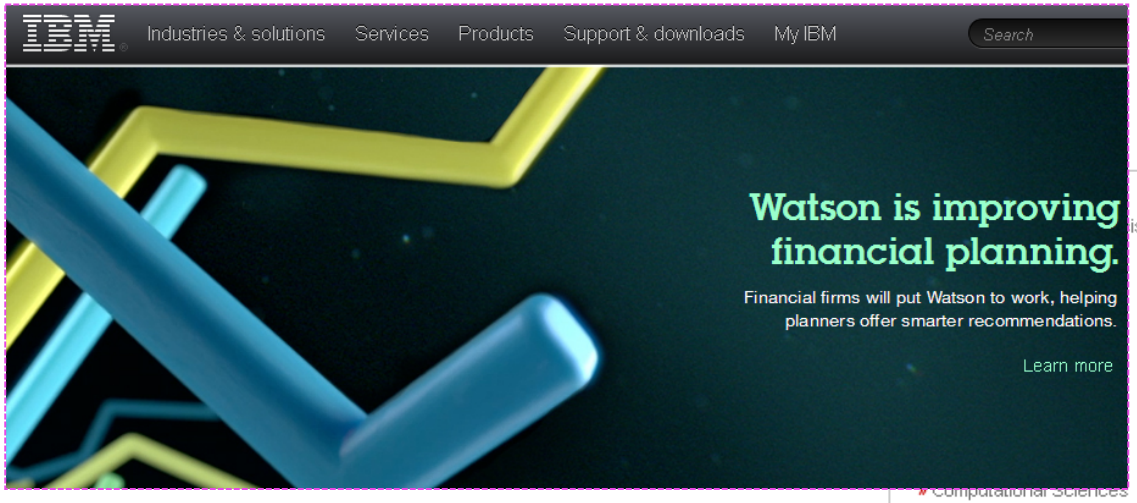
What is the relationship between plant cells and photosynthesis

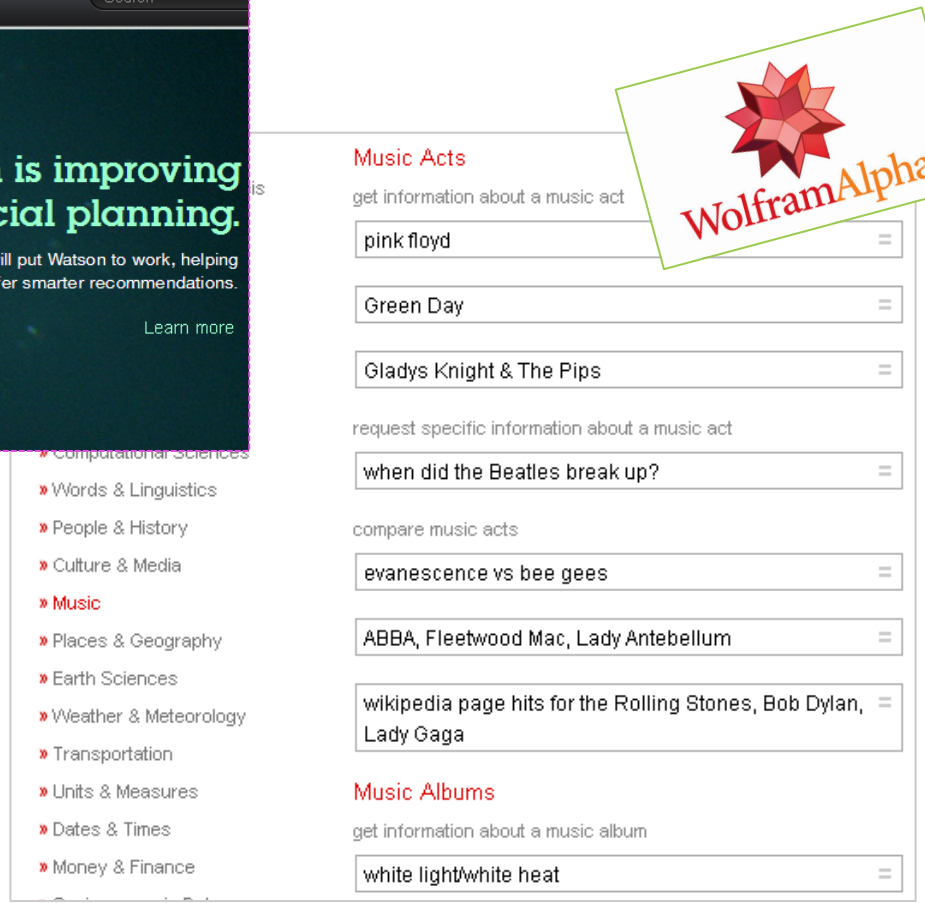
define	Define **cellular respiration**
structure	What is the structure of a **chloroplast**?
function	What is the function of a **plasma membrane** in a **eukaryotic cell**?
compare	What are the differences between **chloroplasts** and **mitochondria**?
relate	If the **chloroplasts** were removed from a **plant**, what events would be affected?
search	Search book for **photosynthesis**

Images from http://www.projecthalo.com and http://www.inquireproject.com/

More examples

EUCLID
EdUcational Curriculum for the usage of LInked Data





Images from http://www.ibm.com/watson, http://www.wolframalpha.com/examples/Music.html, http://www.apple.com

Semantic technologies are not THE solution to creating intelligent applications, but only one (essential) component

The Linked Data movement has promoted one approach to create and publish semantic data

- They created momentum for the Semantic Web, as well as several useful data sets

Rich knowledge representations do have their value, but are costly to achieve

System	Scope	Input	Result	Core Technology
Inquire intelligent textbook	single textbook	simple English queries	formatted data and relevant textbook content	symbolic AI
Wolfram Alpha computational knowledge engine	curated data from "primary sources"	word phrases with mathematical operators	formatted data	Mathematica
Siri virtual personal assistant	emails, calendar, weather, maps, movies, etc.	voice commands	performs tasks	service integration via speech dialog
Google Search index of world's information	open domain text on the web	keywords and search queries	web documents	statistical AI, PageRank

Table from http://www.inquireproject.com/

Our scenario

EUCLID
EdUcational Curriculum for the usage of LInked Data

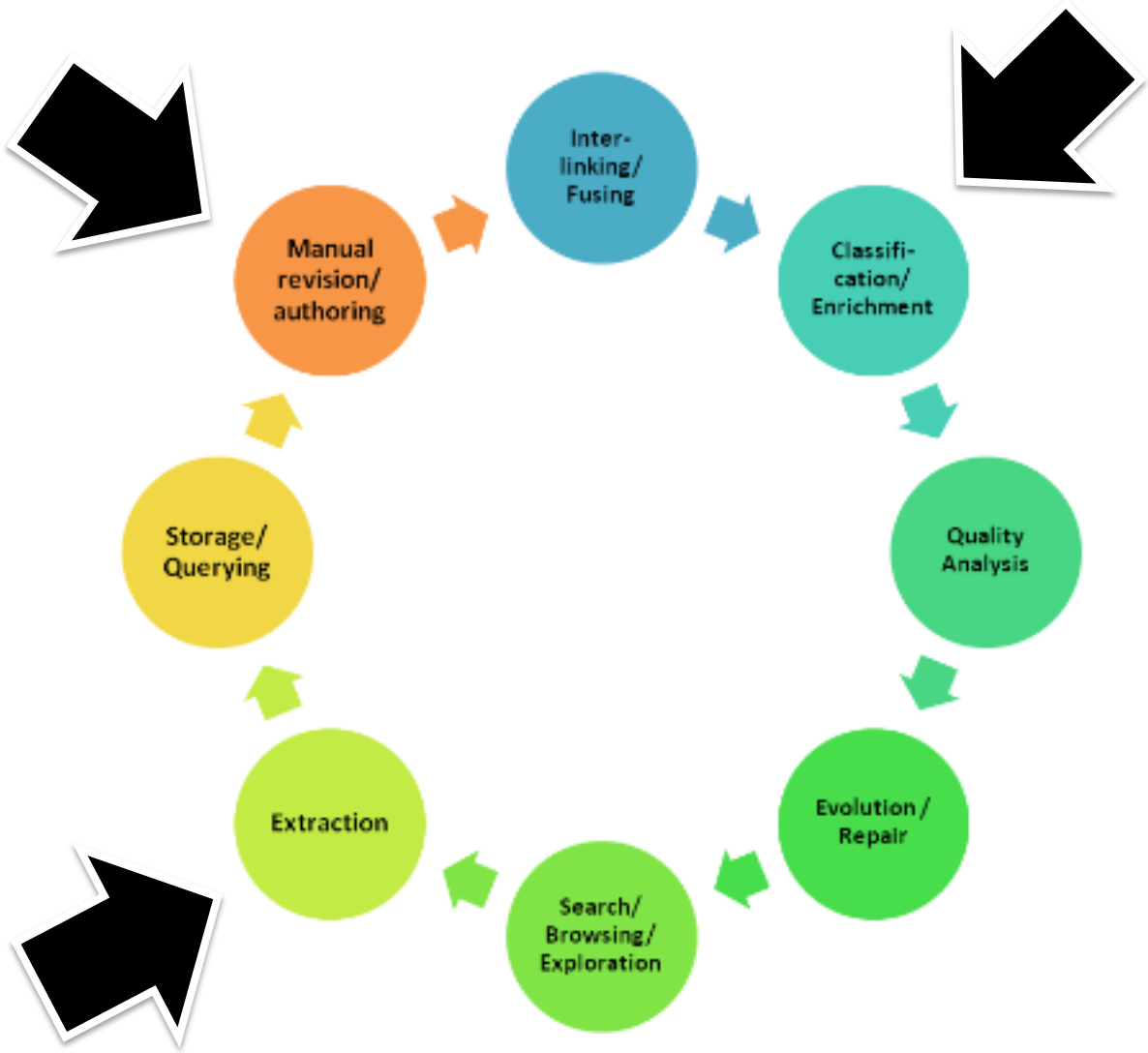
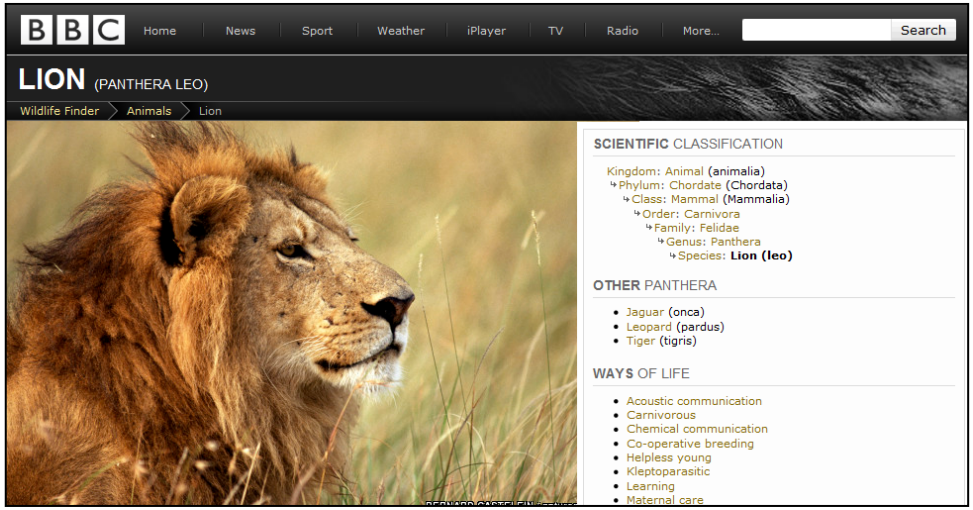


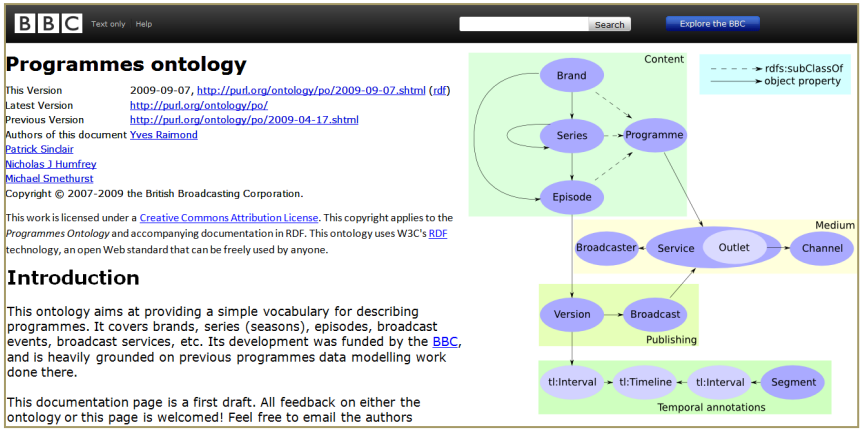
Image from http://wiki.lod2.eu/

Example: BBC

„Design for a world where Google is your homepage, Wikipedia is your CMS, and humans, software developers and machines are your users“

- Various micro-sites built and maintained manually
- No integration across sites in terms of content and metadata
- Use cases
  - Find and explore content on specific (and related) topics
  - Maintain and re-organize sites
  - Leverage external resources
- Ontology: One page per thing, reusing DBpedia and MusicBrainz IDs, different labels





http://www.slideshare.net/reduxd/beyond-the-polar-bear

Core ontology engineering activities in our scenario

EUCLID
EdUcational Curriculum for the usage of LInked Data

- Find ontologies
- Select ontologies
- Extend ontologies

- Popular activities we do **not** consider
  - Requirements analysis
  - Knowledge representation
  - Ontology learning
  - Ontology alignment
  - ...
- See previous summer schools e.g., http://videolectures.net/eswc2012_summer_school/

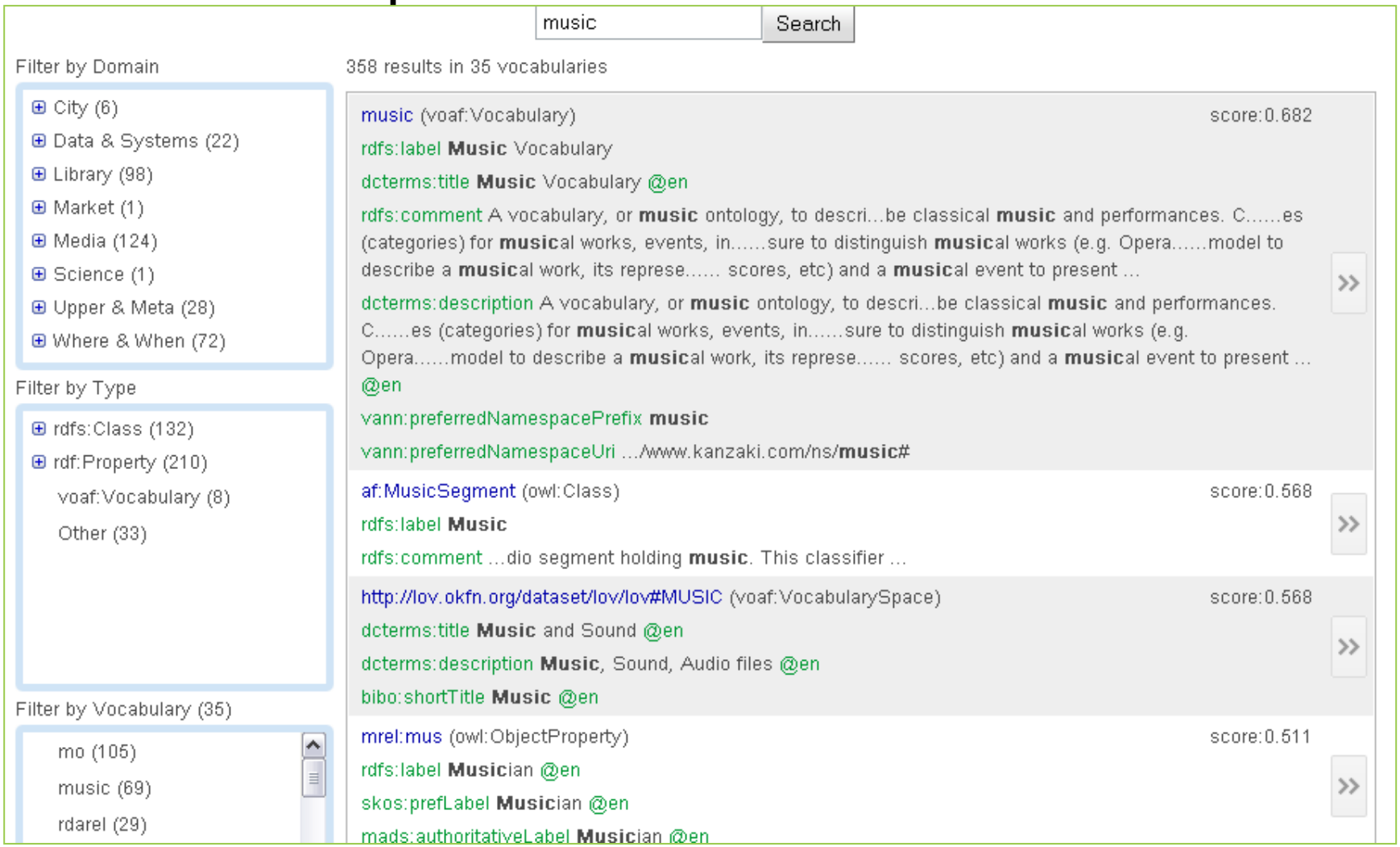
Finding existing ontologies

EUCLID
EdUcational Curriculum for the usage of LInked Data

- Linked Open Vocabularies: vocabularies used in the LOD cloud
  - http://lov.okfn.org
- Protégé Ontologies: several hundreds of ontologies, cross-domain
  - http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library#OWL_ontologies
- Open Ontology Repository: life sciences and other domains
  - http://ontolog.cim3.net/cgi-bin/wiki.pl?OpenOntologyRepository
- Tones: 218 ontologies, life sciences and core ontologies.
  - http://owl.cs.manchester.ac.uk/repository/browser
- Watson: several tens of thousands of documents, cross-domain
  - http://watson.kmi.open.ac.uk/Overview.html
- OBO Foundation Ontologies: hundreds of life sciences ontologies, including mappings
  - http://www.obofoundry.org/
- VoCamps
  - http://vocamp.org/wiki/Main_Page

Linked Open Vocabularies

EUCLID
EdUcational Curriculum for the usage of LInked Data



See http://lov.okfn.org

Linked Open Vocabularies (2)

EUCLID
EdUcational Curriculum for the usage of LInked Data

FOAF - **Friend of a Friend vocabulary**

Metadata:

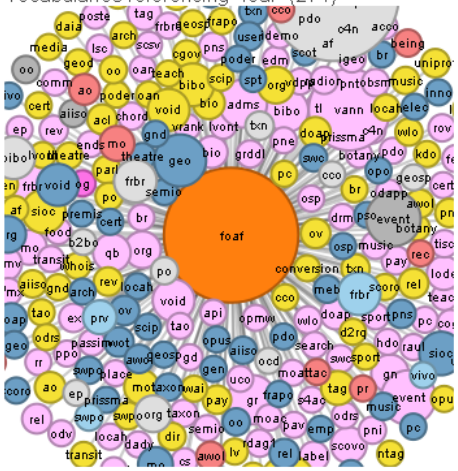
Property	Value
is part of vocabulary space	All > City > People
Vocabulary URI	http://xmlns.com/foaf/0.1/
Prefix	foaf
Namespace URI	http://xmlns.com/foaf/0.1/
Description	FOAF is a project devoted to linking people and information using the Web. Regardless of whether information is i people's heads, in physical or digital documents, or in the form of factual data, it can be linked. @en
Last modified	2010-08-09
Creator	Dan Brickley, Libby Miller
Publisher	Dan Brickley
Class number	13
Property number	62
Homepage	http://www.foaf-project.org/
See also	http://stats.lod2.eu/vocabularies/90
Represented by	format-foaf
Has review	(2013-06-04) Bernard Vatant: From the specification mid-2000. There is now a stable core of classes and to their documentation to track implementation feedb

Vocabulary used in **182** datasets (information coming from LOD Stats project)

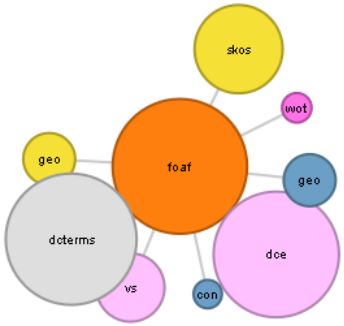
Occurrences	Dataset
20757535	Bibliographica.org BNB export
14673353	B3Kat - Library Union Catalogues of Bavaria, Berlin and Brandenburg
4687220	Ookaboo RDF dump
3586838	Europeana Linked Open Data
2882067	Rådata nå!
See the full list...	

Vocabulary links:

Vocabularies referencing "foaf" (271)

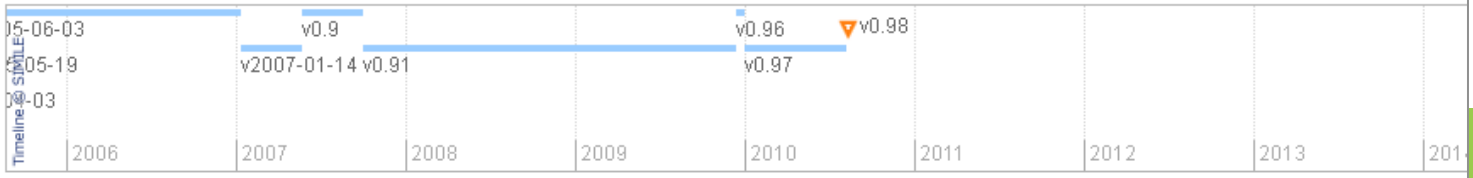


Vocabularies referenced by "foaf" (8)



Link types
- Imports
- Similar to
- Used by
- Relies on
  - Metadata vocabulary
  - Extends
  - Specializes
  - Generalizes
  - Has equivalences with
  - Has disjunctions with

Vocabulary history:



LODStats

EUCLID
EdUcational Curriculum for the usage of LInked Data

499 vocabularies

music Search

URI	Overall	Datasets
http://www.w3.org/1999/02/22-rdf-syntax-ns	278,646,233	671
http://www.w3.org/2000/01/rdf-schema	58,329,356	533
http://www.w3.org/2002/07/owl	12,934,565	277
http://purl.org/dc/terms/	127,660,537	225
http://www.w3.org/2004/02/skos/core	45,773,412	174
http://purl.org/dc/terms	60,043,900	153
http://xmlns.com/foaf/0.1/	22,121,715	150
http://purl.org/dc/elements/1.1/	9,342,467	141
http://xmlns.com/foaf/0.1	29,489,527	106
http://purl.org/dc/elements/1.1	20,261,128	90
http://rdfs.org/ns/void	12,664	77
http://www.w3.org/2003/01/geo/wgs84_pos	997,603	61
http://www.aktors.org/ontology/portal	33,286,647	54

See http://stats.lod2.eu/vocabularies

Dublin Core

EUCLID
EdUcational Curriculum for the usage of LInked Data

Properties in the */terms/* namespace	abstract, accessRights, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, available, bibliographicCitation, conformsTo, contributor, coverage, created, creator, date, dateAccepted, dateCopyrighted, dateSubmitted, description, educationLevel, extent, format, hasFormat, hasPart, hasVersion, identifier, instructionalMethod, isFormatOf, isPartOf, isReferencedBy, isReplacedBy, isRequiredBy, issued, isVersionOf, language, license, mediator, medium, modified, provenance, publisher, references, relation, replaces, requires, rights, rightsHolder, source, spatial, subject, tableOfContents, temporal, title, type, valid
Properties in the legacy */elements/1.1/* namespace	contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type
Vocabulary Encoding Schemes	DCMIType, DDC, IMT, LCC, LCSH, MESH, NLM, TGN, UDC
Syntax Encoding Schemes	Box, ISO3166, ISO639-2, ISO639-3, Period, Point, RFC1766, RFC3066, RFC4646, RFC5646, URI, W3CDTF
Classes	Agent, AgentClass, BibliographicResource, FileFormat, Frequency, Jurisdiction, LicenseDocument, LinguisticSystem, Location, LocationPeriodOrJurisdiction, MediaType, MediaTypeOrExtent, MethodOfAccrual, MethodOfInstruction, PeriodOfTime, PhysicalMedium, PhysicalResource, Policy, ProvenanceStatement, RightsStatement, SizeOrDuration, Standard

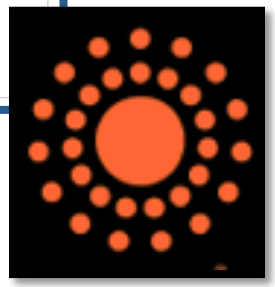
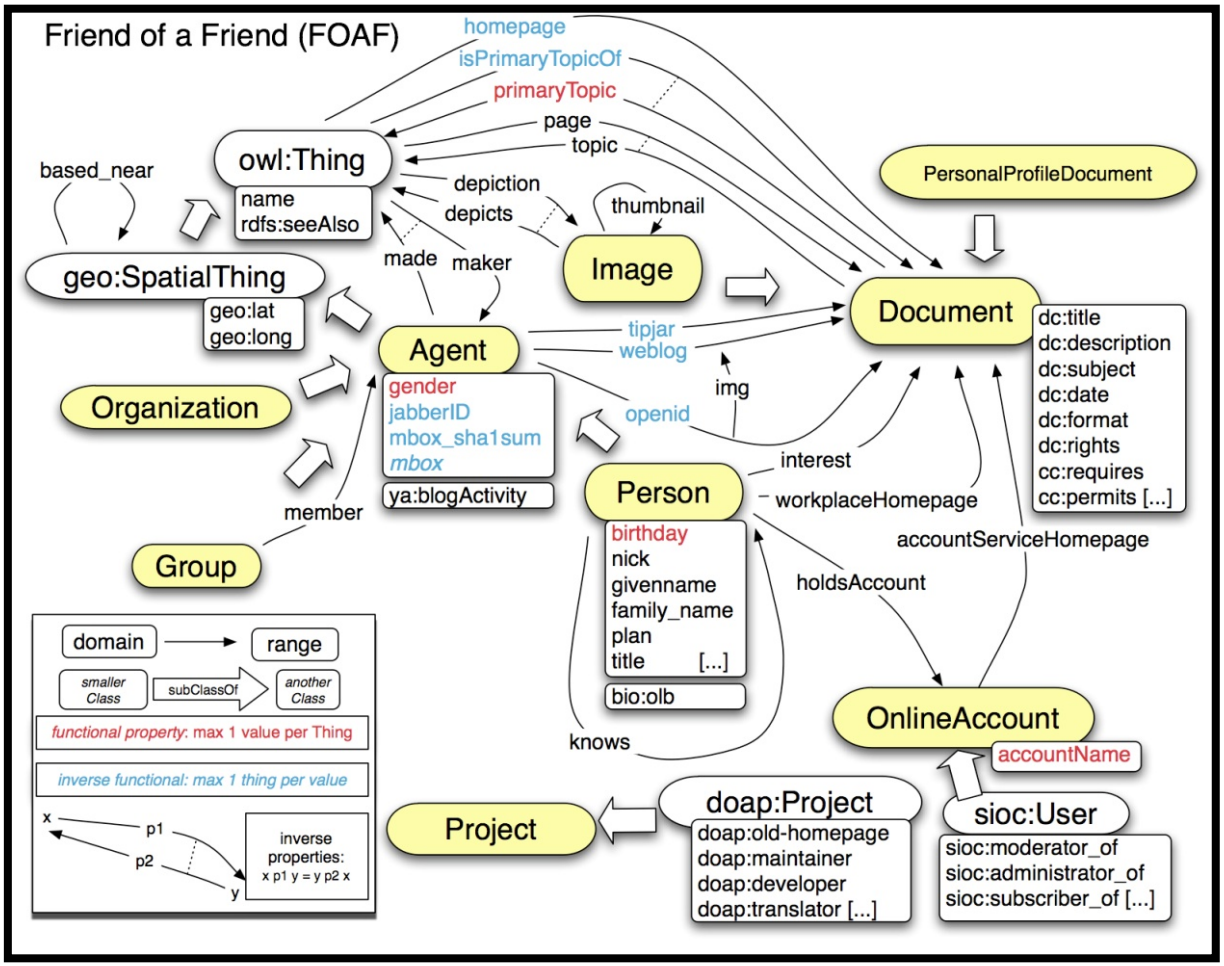


Table from http://dublincore.org/documents/dcmi-terms/

Friend Of A Friend

EUCLID
EdUcational Curriculum for the usage of LInked Data



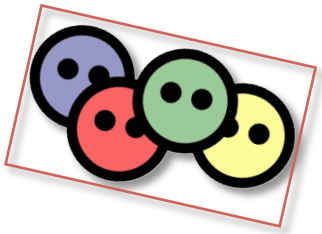


Image from http://www.deri.ie/fileadmin/images/blog/: Breslin

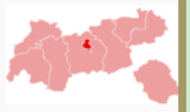
DBpedia

EUCLID
EdUcational Curriculum for the usage of LInked Data

- Classes and properties for Wikipedia export (infoboxes), regularly updated

```
{{Infobox Town AT |
  name = Innsbruck |
  image_coa =  InnsbruckWappen.png |
  image_map = Karte-tirol-I.png |
  state = [[Tyrol]] |
  regbzk = [[Statutory city]] |
  population = 117,342 |
  population_as_of = 2006 |
  pop_dens = 1,119 |
  area = 104.91 |
  elevation = 574 |
  lat_deg = 47 |
  lat_min = 16 |
  lat_hem = N |
  lon_deg = 11 |
  lon_min = 23 |
  lon_hem = E |
  postal_code = 6010-6080 |
  area_code = 0512 |
  licence = I |
  mayor = Hilde Zach |
  website = [http://innsbruck.at] |
}}
```

Innsbruck



Country	Austria
State	Tyrol
Administrative region	Statutory city
Population	117,342 (2006)
Area	104.91 km²
Population density	1,119 /km²
Elevation	574 m
Coordinates	47°16' N 11°23' E
Postal code	6010-6080
Area code	0512
Licence plate code	I
Mayor	Hilde Zach
Website	www.innsbruck.at

About: Innsbruck

An Entity of Type : city, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

Innsbruck is the capital city of the federal state of Tyrol in western Austria. It is located in the Inn Valley at the junction with the Wipptal, which provides access to the Brenner Pass, some 30 kilometers (19 mi) south of Innsbruck.

Property	Value
dbpedia-owl:PopulatedPlace/populationDensity	1119.0
dbpedia-owl:abstract	Innsbruck ist die Landeshauptstadt des Bundeslandes Tiro Transit-Strecke Brenner (Auto- und Eisenbahn) nach Südti (Brücke über den Inn). Innsbruck ist mit 118.082 (Stand 1. und Salzburg die fünftgrößte Stadt Österreichs, im Ballung zu kommen ca. 30.000 Studenten und andere Nebenwol chtigungen von Städtetouristen. nnsbruck is the capital city of the federal state of Tyrol in ion with the Wipptal, which provides access to the

DBpedia

See http://wiki.dbpedia.org/

Freebase

EUCLID
EdUcational Curriculum for the usage of LInked Data

An entity graph of people, places and things, built by a community that loves open data.

Freebase is joining Google! Learn more »

- Featured Data
- Arts & Entertainment
- Products & Services
- Science & Technology
- Society
- Special Interests
- Sports
- System
- Time & Space
- All

	May 31 – Jun 28 – last week	Facts	Topics	Top User
Film, 81 members		38K last week	3M	340K
Location, 30 members		27K	10M	920K
People, 58 members				
Books, 47 members				
TV, 35 members				
Music, 100+ members				
Business, 100+ members				
Government, 29 members				

Type	ID	# of Properties	# of Instances	Description
Musical Director	/theater/musical_director	1	12	A musical director is a person w
Musical Soundtrack	/theater/musical_soundtrack	2	18	A musical soundtrack is a record
Play	/theater/play	10	4,415	A play is a work written to be pe
Theater	/theater/theater	1	1,166	The theater type is for performa
Theater Actor	/theater/theater_actor	1	4,020	People who have performed in a
Theater Character	/theater/theater_character	2	882	This type is for all characters tha
Theater Choreographer	/theater/theater_choreographer	1	20	A theater choreographer is some
Theater Company	/theater/theater_company	1	647	A theater company is a group of
Theater Designer	/theater/theater_designer	1	74	A theater designer is someone w
Theater Designer Role	/theater/theater_designer_role	1	27	This type represents possible po
Theater Director	/theater/theater_director	1	99	A theater director is someone wh
Theater Genre	/theater/theater_genre	1	49	This type is for all genres of play
Theater Producer	/theater/theater_producer	1	39	A theater producer is a person w
Theater Production	/theater/theater_production	15	599	A theater production is a product
Theater Production Staff Role	/theater/theater_production_staff_role	1	18	This type represents all positions
Theater Production Staffmember	/theater/theater_production_staff	1	33	This type is for any person who
Theatrical Composer	/theater/theatrical_composer	1	140	A theatrical composer is someon
Theatrical Lyricist	/theater/theatrical_lyricist	1	144	A theatrical lyricist is someone w

Semantically Interlinked Online Communities

EUCLID
EdUcational Curriculum for the usage of LInked Data

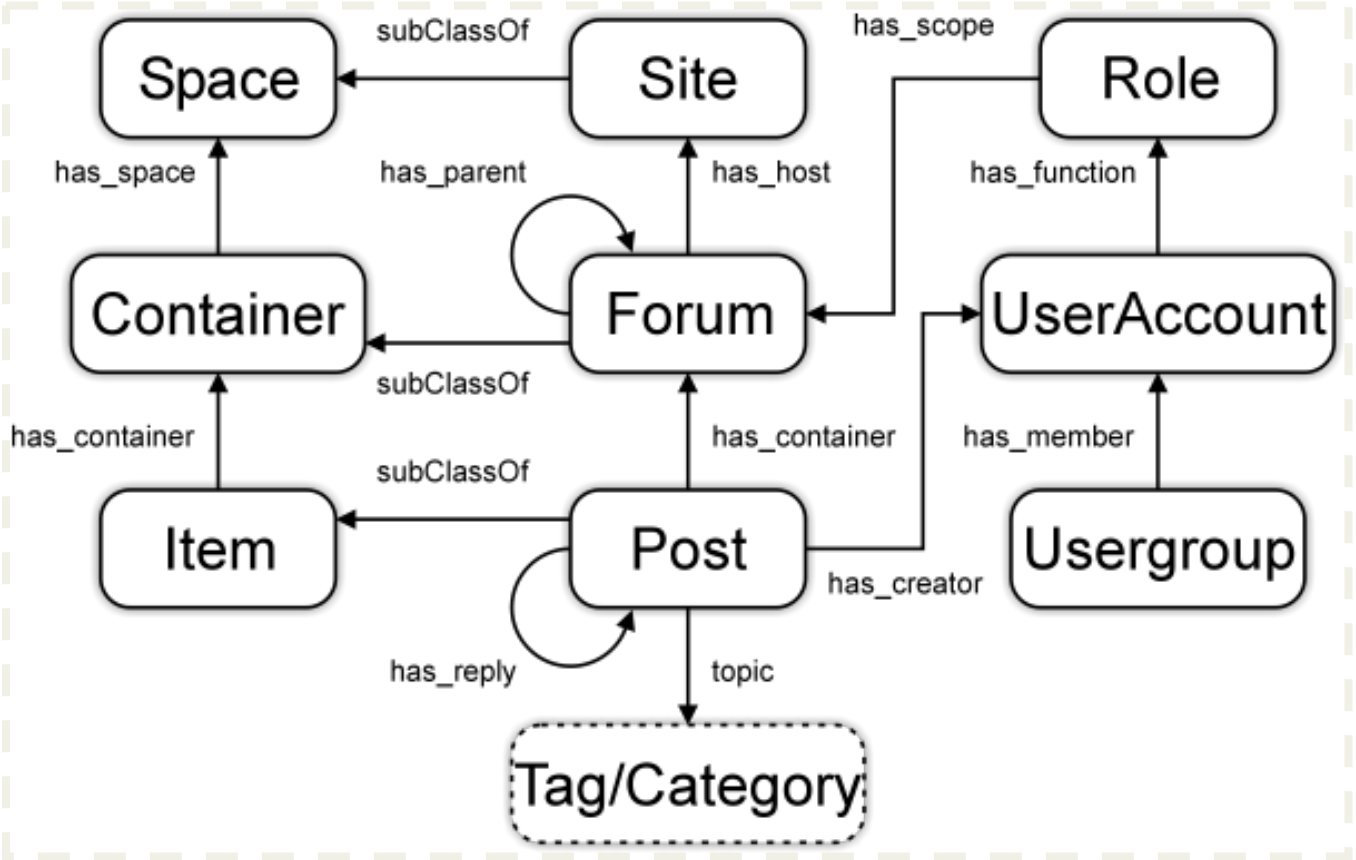


Image from http://rdfs.org/sioc/spec/: Bojārs, Breslin et al.

Simple Knowledge Organization System

EUCLID
EdUcational Curriculum for the usage of LInked Data

W3C Recommendation

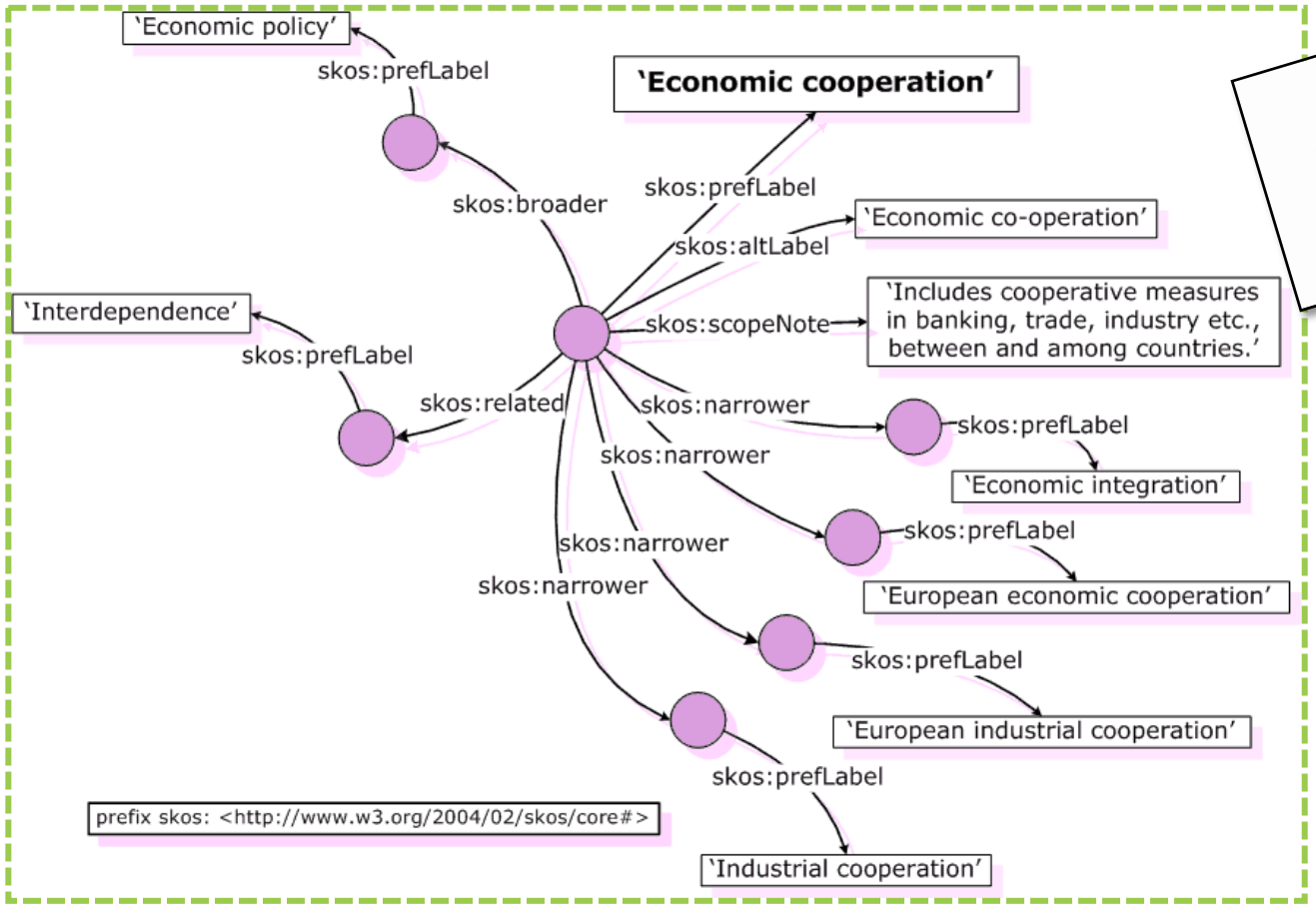


Image from http://www.w3.org/TR/swbp-skos-core-guide: Miles, Brickley

Description Of A Project

EUCLID
EdUcational Curriculum for the usage of LInked Data

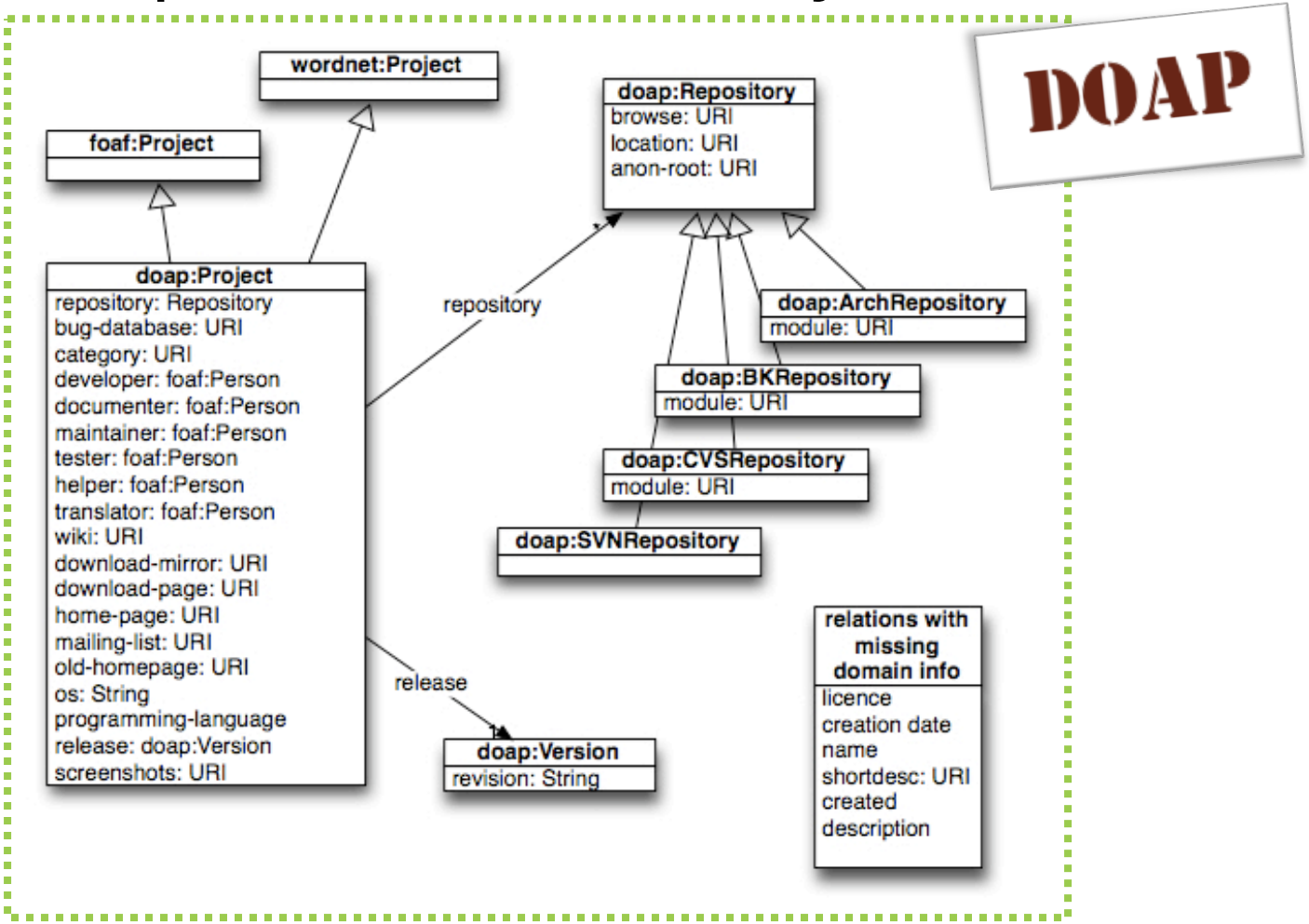
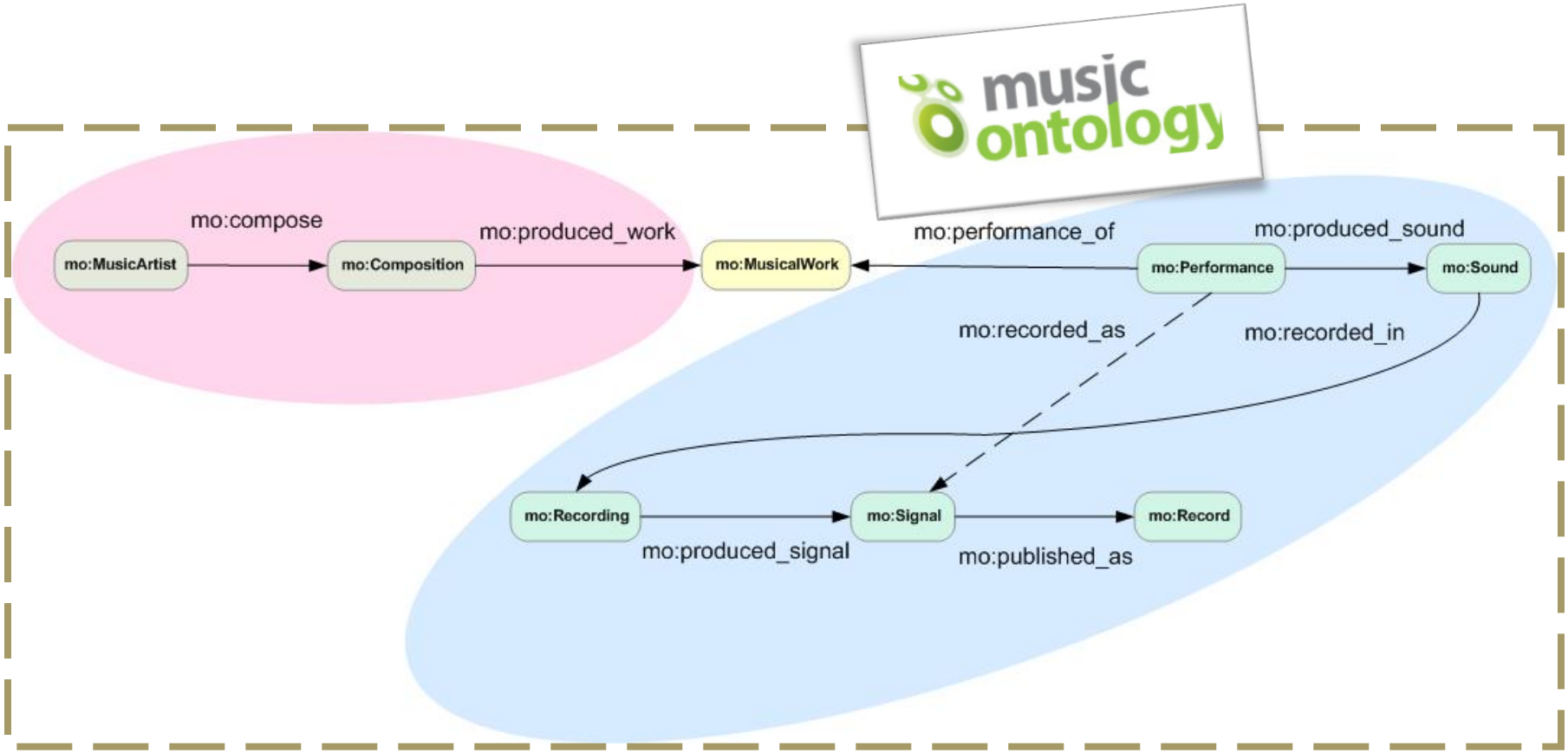


Image from http://code.google.com/p/baetle/wiki/DoapOntology: Breslin

Music Ontology

EUCLID
EdUcational Curriculum for the usage of LInked Data



BBC

Image from http://musicontology.com/:Raimond, Giasson

Life sciences and healthcare

EUCLID
EdUcational Curriculum for the usage of LInked Data

The Open Biological and Biomedical Ontologies

Ontologies | Resources | Participate | About

The OBO Foundry is a collaborative experiment involving developers of science-based ontologies who are establishing a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. The groups developing ontologies who have expressed an interest in this goal are listed below, followed by other relevant efforts in this domain.

In addition to a listing of OBO ontologies, this site also provides a statement of the OBO Foundry principles, discussion fora, technical infrastructure, and other services to facilitate ontology development. We welcome feedback and encourage participation.

Click any column header to sort the table by that column. The s link to the term request trackers for the listed ontologies.

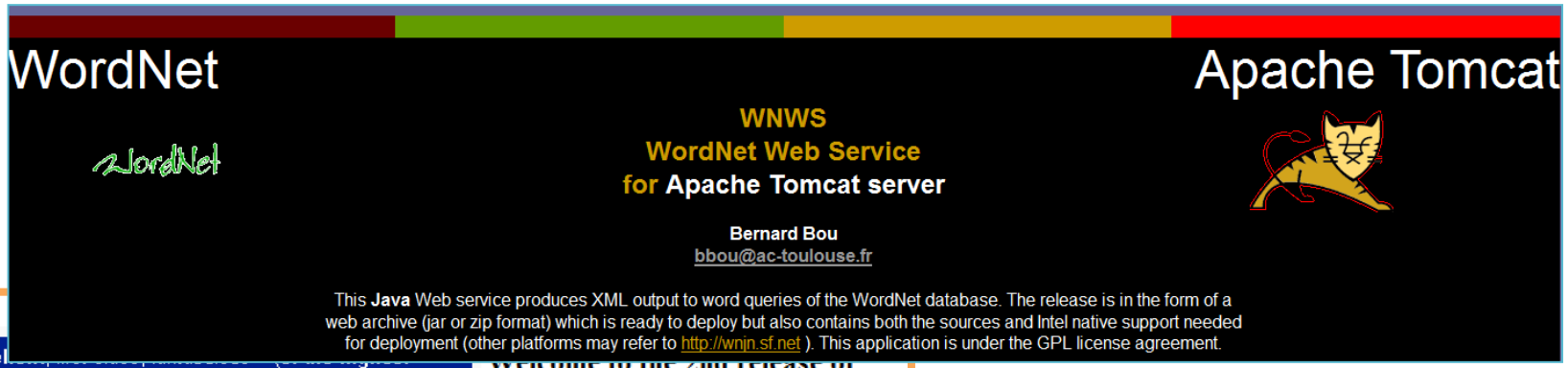
OBO Foundry ontologies

Title	Domain	Prefix	File	Last changed
Biological process	biological process	GO	gene_ontology_edit.obo	2010/07/26
Cellular component	anatomy	GO	gene_ontology_edit.obo	2010/07/26
Chemical entities of biological interest	biochemistry	CHEBI	chebi.obo	2010/07/07
Molecular function	biological function	GO	gene_ontology_edit.obo	2010/07/26
Phenotypic quality	phenotype	PATO	quality.obo	2010/07/25
PRotein Ontology (PRO)	proteins	PRO	pro.obo	2010/07/24
Xenopus anatomy and development	anatomy	XAO	xenopus_anatomy.obo	2009/12/02
Zebrafish anatomy and development	anatomy	ZFA	zebrafish_anatomy.obo	2010/06/14

http://www.w3.org/TR/wordnet-rdf/

WordNet

EUCLID
EdUcational Curriculum for the usage of LInked Data



1. **excel**
quality; "made an excellent speech"; "the school has excellent teachers"; "a first-class mind")

(TOP (S (NP (JJ **excellent**))
(VP (VBZ **is**)
(NP (NP (NN **something**))
(PP (IN **of**)
(NP (DT **the**) (JJS **highest**) (NN **quality**)))))
(. .)))

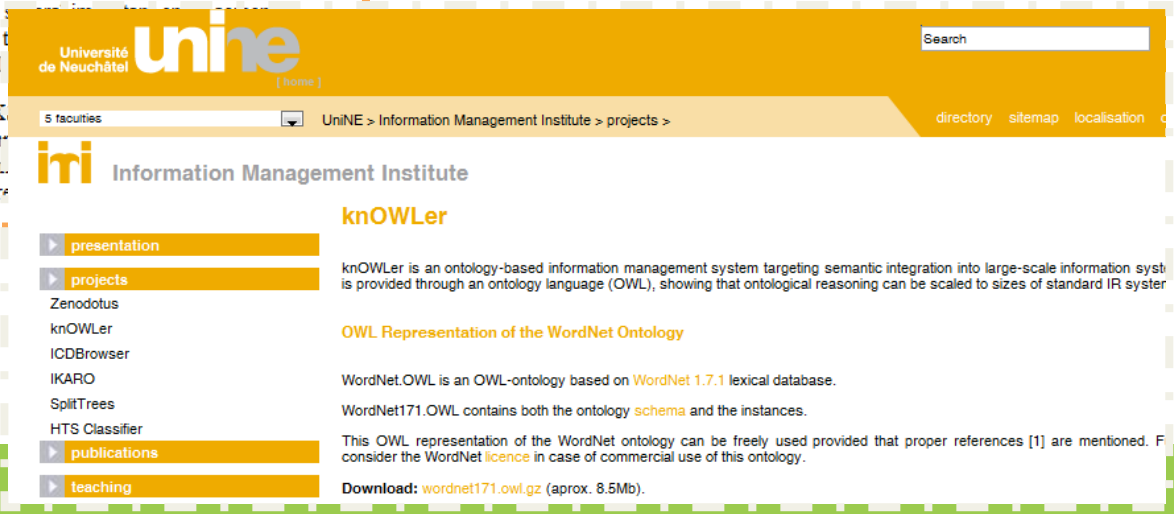
excellent:JJ(x1) -> **of**:IN(x1, x2) **highest**:JJ(x1)
quality:NN(x1)

<wf pos="IN" >**of**</wf>
<wf pos="DT" >**the**</wf>
<wf pos="JJS" lemma="high" quality="silver" wnsn="1"

eXtended WordNet

The goal of this project is to develop a tool that takes as input the current or future versions of WordNet and automatically generates an eXtended WordNet that provides
intended t
of Word

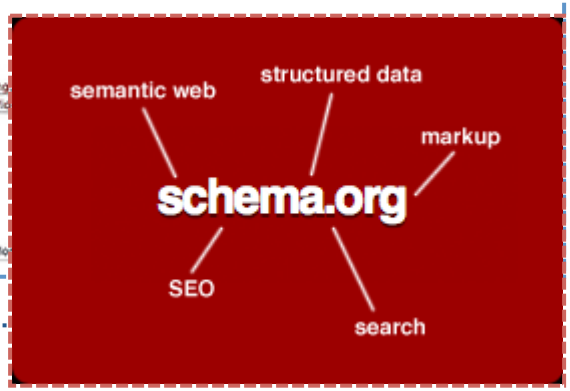
In the eX
glosses a
transforr
words ar

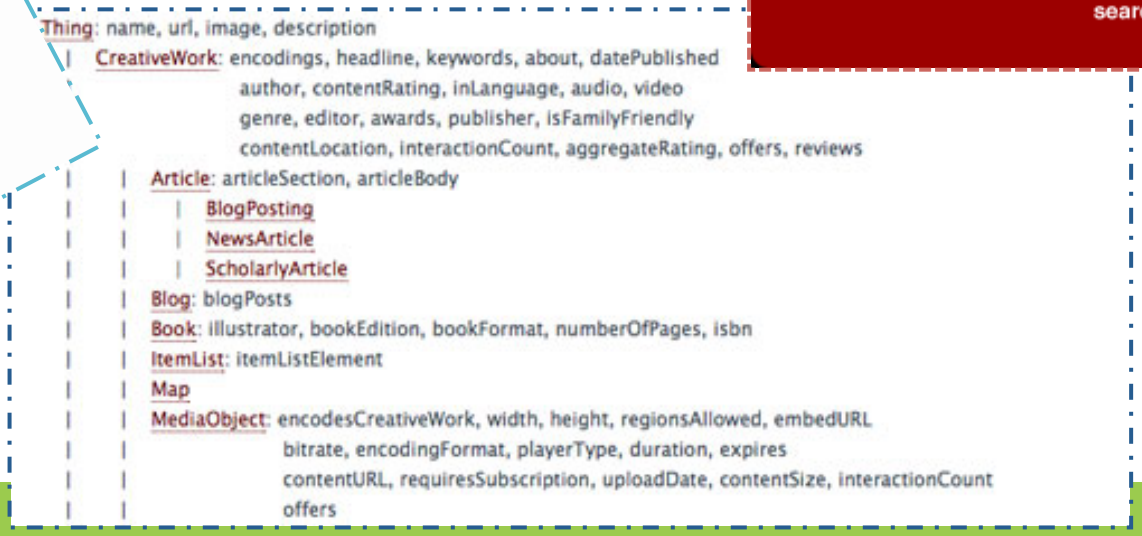


schema.org

EUCLID
EdUcational Curriculum for the usage of LInked Data

- Collection of schemas to mark-up structured content in HTML pages





GoodRelations

EUCLID
EdUcational Curriculum for the usage of LInked Data

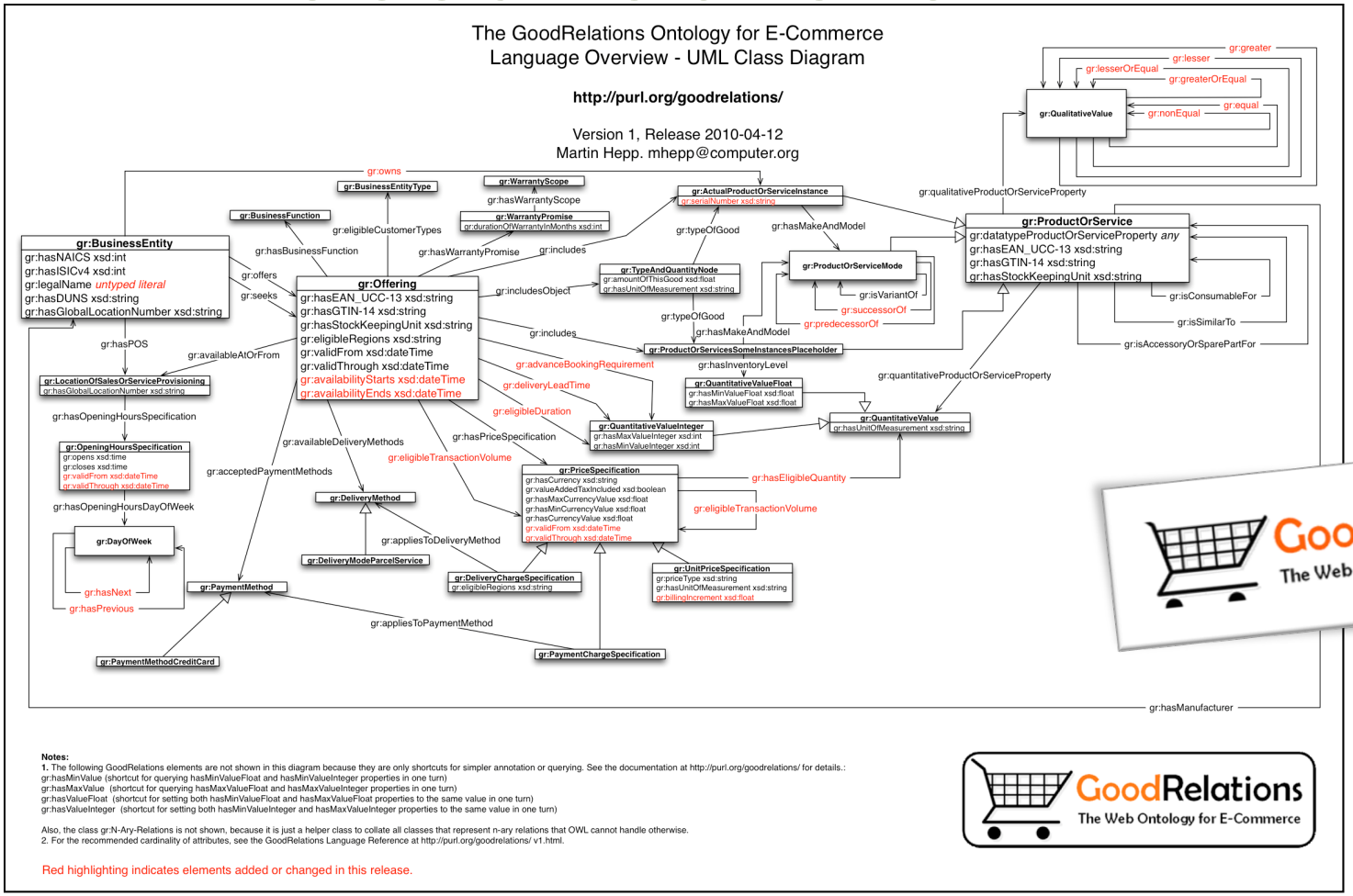
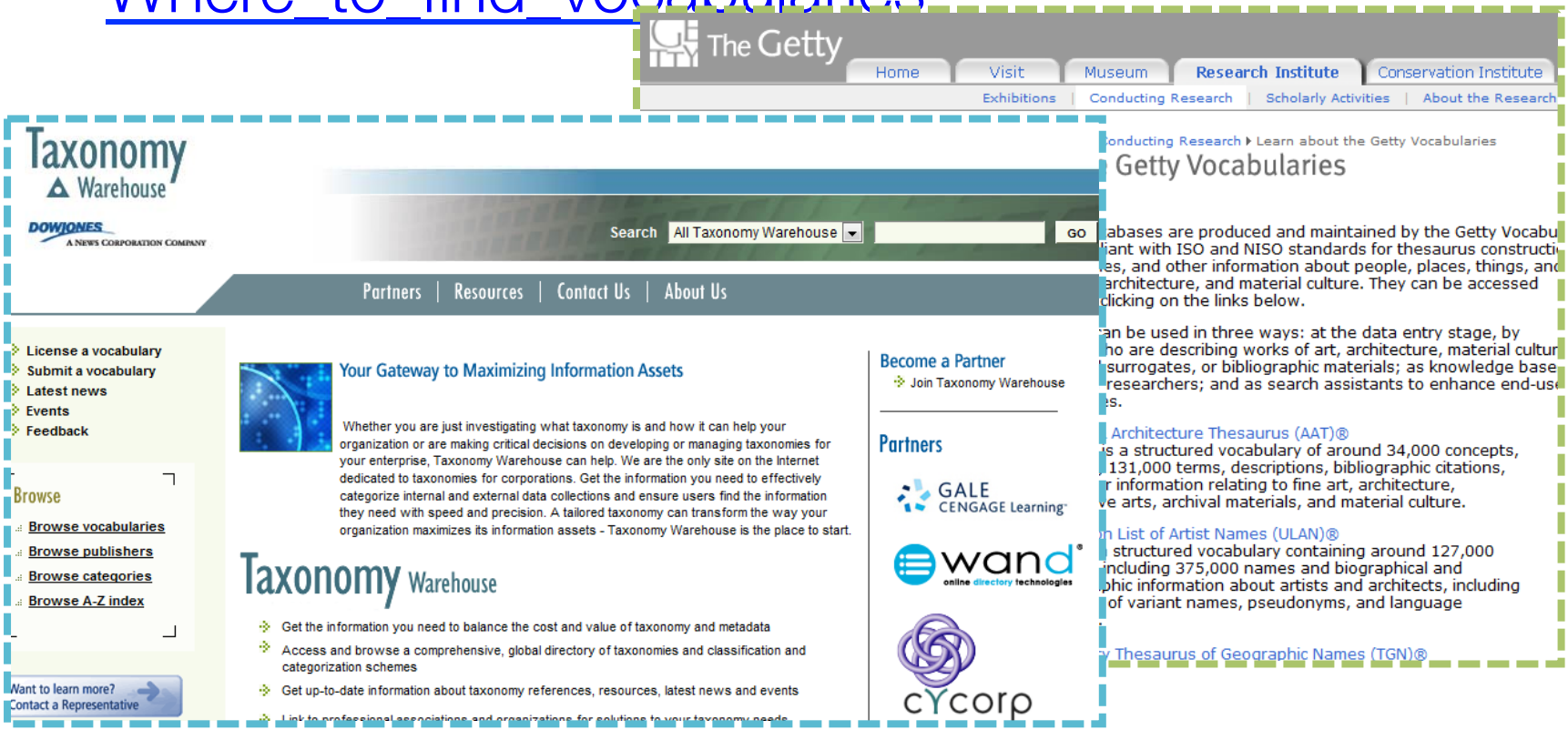


Image from http://www.heppnetz.de/projects/goodrelations/primer/; Hepp

Additional resources

EUCLID
EdUcational Curriculum for the usage of LInked Data

- http://vocamp.org/wiki/Where_to_find_vocabularies



Selecting relevant ontologies

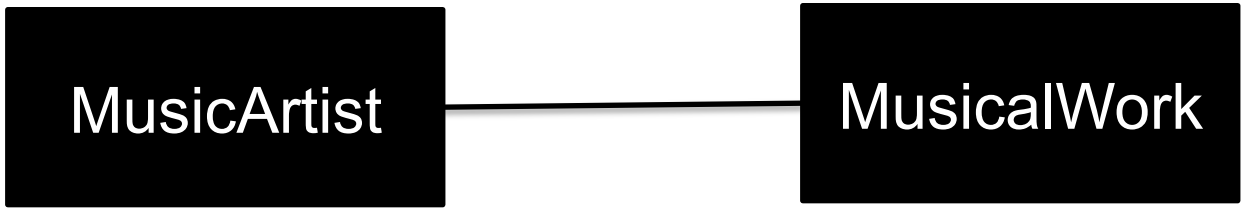
EUCLID
EdUcational Curriculum for the usage of LInked Data

- What will the ontology be used for?
  - Does it need a natural language interface and if yes in which language?
  - Do you have any knowledge representation constraints (language, reasoning)?
  - What level of expressivity is required?
  - What level of granularity is required?
- What will you reuse from it?
  - Vocabulary++
- How will you reuse it?
  - Imports: transitive dependency between ontologies
  - Changes in imported ontologies can result in inconsistencies and changes of meanings and interpretations, as well as computational aspects

Classes

EUCLID
EdUcational Curriculum for the usage of LInked Data

- A class represents a set of instances
- A class should be highly cohesive, precisely nameable, relevant
- A class should have a strong identity



Classes

EUCLID
EdUcational Curriculum for the usage of LInked Data

- Classes represent **concepts** in the domain and **not** the **words** that denote these concepts
  - Synonyms for the same concept do not represent different classes
- Typically nouns and nominal phrases, but not restricted to them
  - **Verbs** can be modeled as classes, if the emphasis is on the process as a whole rather than the actual execution

  - Events: „Being ill“ → „Illness episode“
  - No pronouns

Cohesion and identity

EUCLID
EdUcational Curriculum for the usage of LInked Data

- A class should represent one thing, all of that thing and nothing but that thing
- You can prove cohesion by
  - Giving the class a representative name
  - Noun (+ modifier, sometimes however also captured as attribute value)
    - Examples: Jazz music, classical music
  - Albums is not cohesive
- Avoid ambiguous terms
  - Manager, handler, processor, list, information, item, data…
- Identity ~ individuality: entities change values, but remain members of the same class
  - Examples: Child/Adult: age

Instances

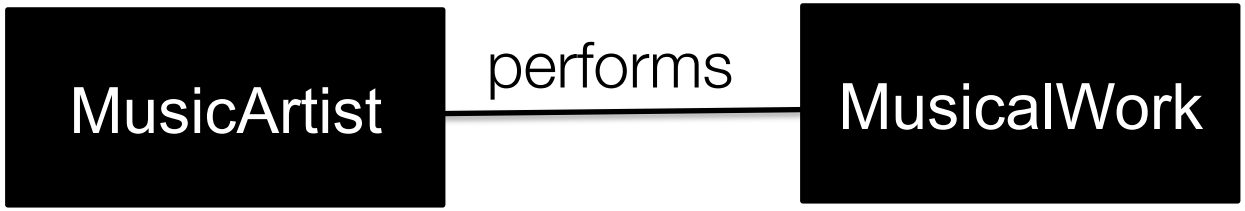
EUCLID
EdUcational Curriculum for the usage of LInked Data

- Entities of a certain type
  - Abstract entities are allowed
- Issues
  - Distinction between classes and instances
    - Examples: **Jazz as music genre**
  - Choice of the most appropriate class
    - Examples: **Violetta Valery, Stradivarius**
- Instances adopt attributes and properties of their classes
  - Examples: **rectangles and squares**

Characterizing classes

EUCLID
EdUcational Curriculum for the usage of LInked Data

- Two types of principal characteristics
  - Measurable properties: attributes
  - Inter-entity connections: relationships, associations



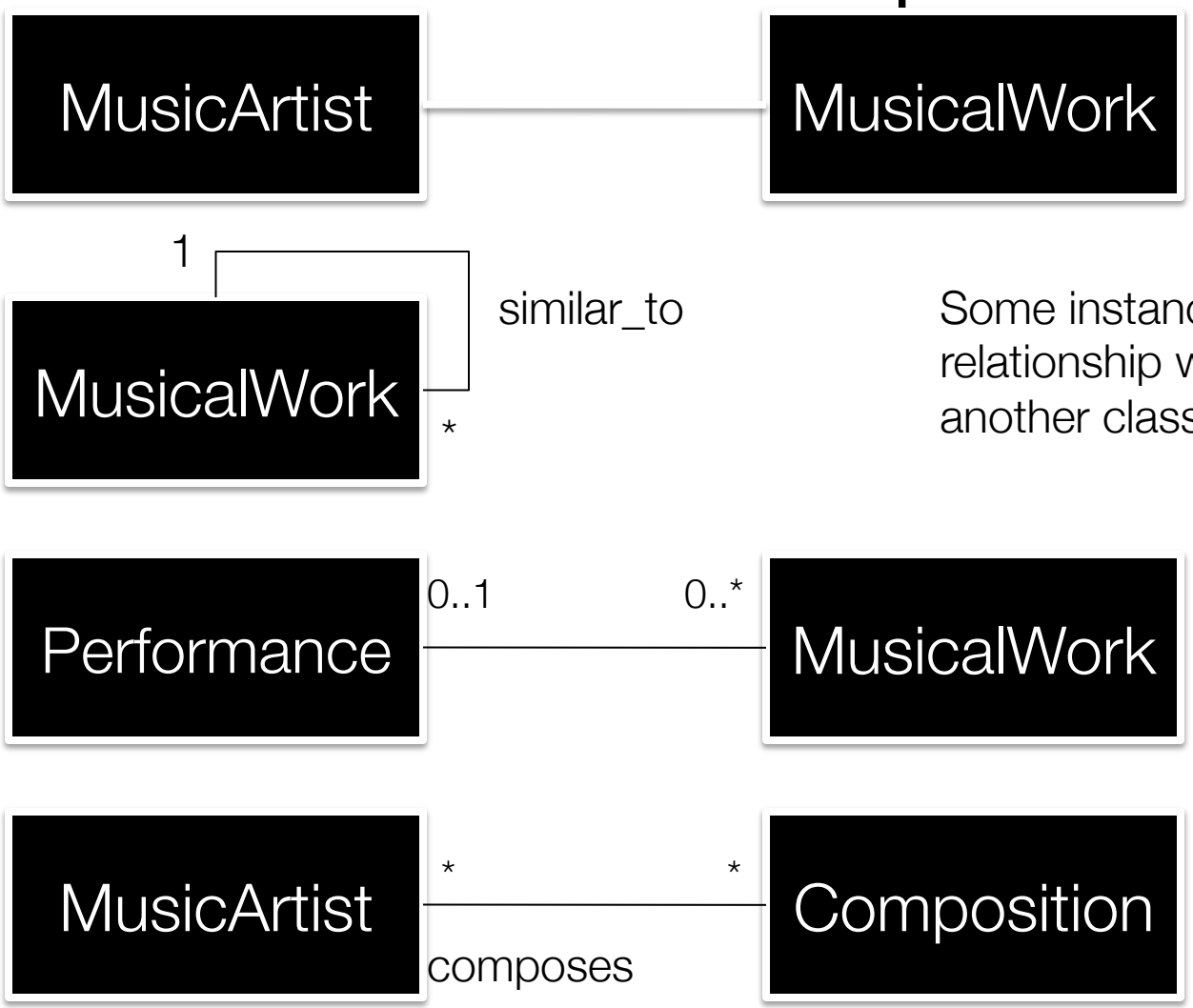
  - Color of an image as attribute vs. class

Attributes

EUCLID
EdUcational Curriculum for the usage of LInked Data

- An attribute is a measurable property of a class
  - Scalar values: choice from a range of possibilities
  - An attribute is NOT a data structure. It is not complicated to measure
  - Attributes do NOT exhibit identity
- Typically nouns in „-ness“
  - Velocity-ness, job-ness, arrested-ness…
- „How much, how many“ test.
  - If you evaluate this, then it is probably an attribute
  - If you enumerate these, it is probably a class

Relationships



Some instances of a class hold a relationship with some instances of another class.

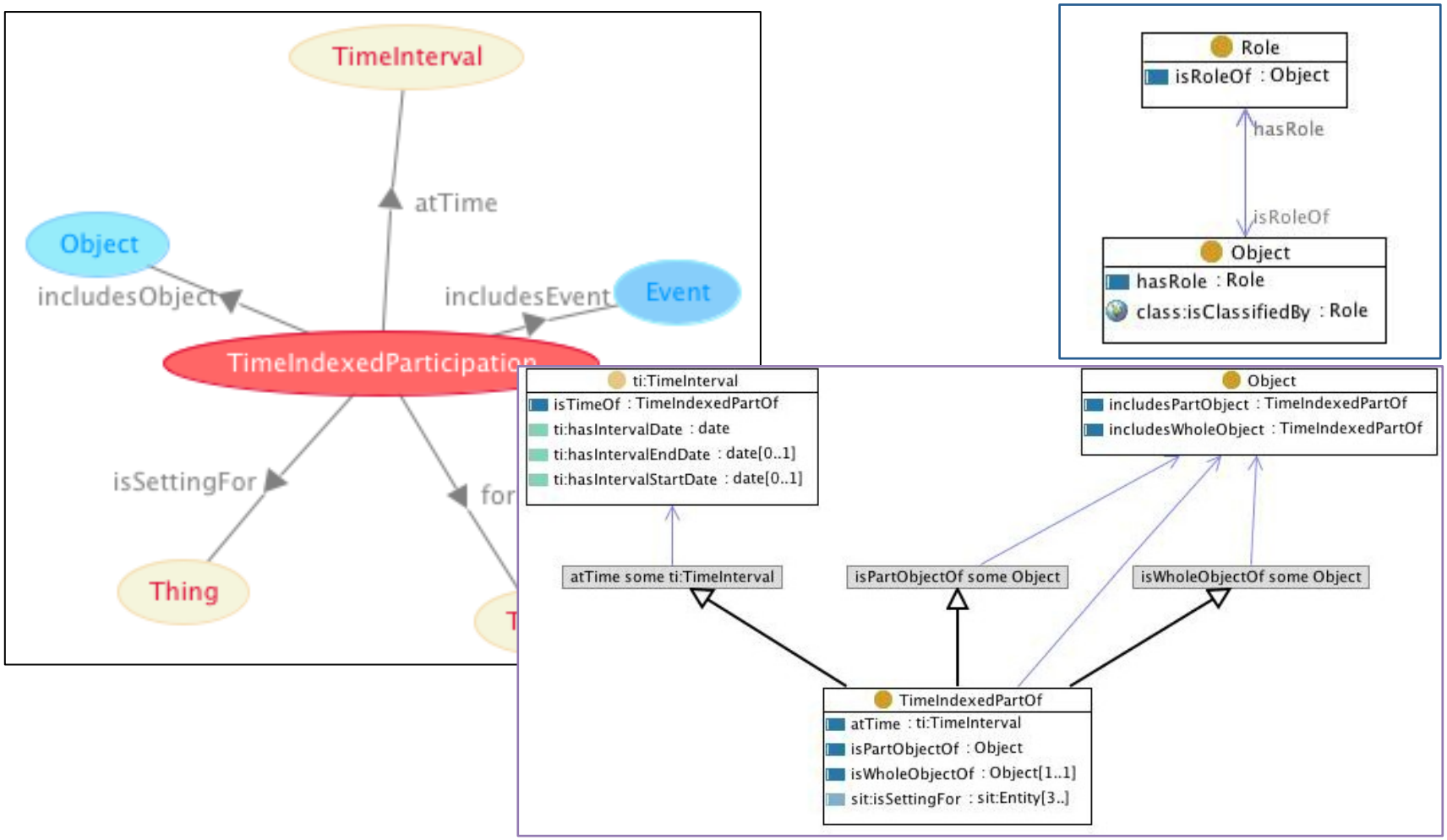
Class hierarchy

EUCLID
EdUcational Curriculum for the usage of LInked Data

- A subclass of a class represents a concept that is a “kind of” the concept that the superclass represents
- It has
  - Additional properties
  - Restrictions different from those of the superclass, or
  - Participates in different relationships than the superclasses
- All the siblings in the hierarchy (except for the ones at the root) must be at the same level of generality
- If a class has only one direct subclass there may be a modeling problem or the ontology is not complete
- If there are more than a dozen subclasses for a given class then additional intermediate categories may be necessary
- Roles are not subclasses
  - Application dependent or subjective
  - Examples: Artist and Person

Ontology Design Paterns

EUCLID
EdUcational Curriculum for the usage of LInked Data



Content from http://ontologydesignpatterns.org/

EUCLID
EdUcational Curriculum for the usage of LInked Data

Assignment

Assignment: Modeling

EUCLID
EdUcational Curriculum for the usage of LInked Data

The current configuration of the “Red Hot Chili Peppers” are: Anthony Kiedis (vocals), Flea (bass, trumpet, keyboards, and vocals), John Frusciante (guitar), and Chad Smith (drums). The line-up has changed a few times during they years, Frusciante replaced Hillel Slovak in 1988, and when Jack Irons left the band he was briefly replaced by D.H. Peligo until the band found Chad Smith. In addition to playing guitars for Red hot Chili Peppers Frusciante also contributed to the band “The Mars Volta” as a vocalist for some time.

From September 2004, the Red Hot Chili Peppers started recording the album “Stadium Arcadium”. The album contains 28 tracks and was released on May 5 2006. It includes a track of the song “Hump de Bump”, which was composed in January 26, 2004. The critic Crian Hiatt defined the album as "the most ambitious work in his twenty-three-year career". On August 11 (2006) the band gave a live performance in Portland, Oregon (US), featuring songs from Stadium Arcadium and other albums.

EUCLID
EdUcational Curriculum for the usage of LInked Data

http://www.euclid-project.eu

eBook

HTML

kindle amazonkindle

ePUB

iBooks

Course

HTML

iTunes U

Other channels

@euclid_project

euclidproject

euclidproject