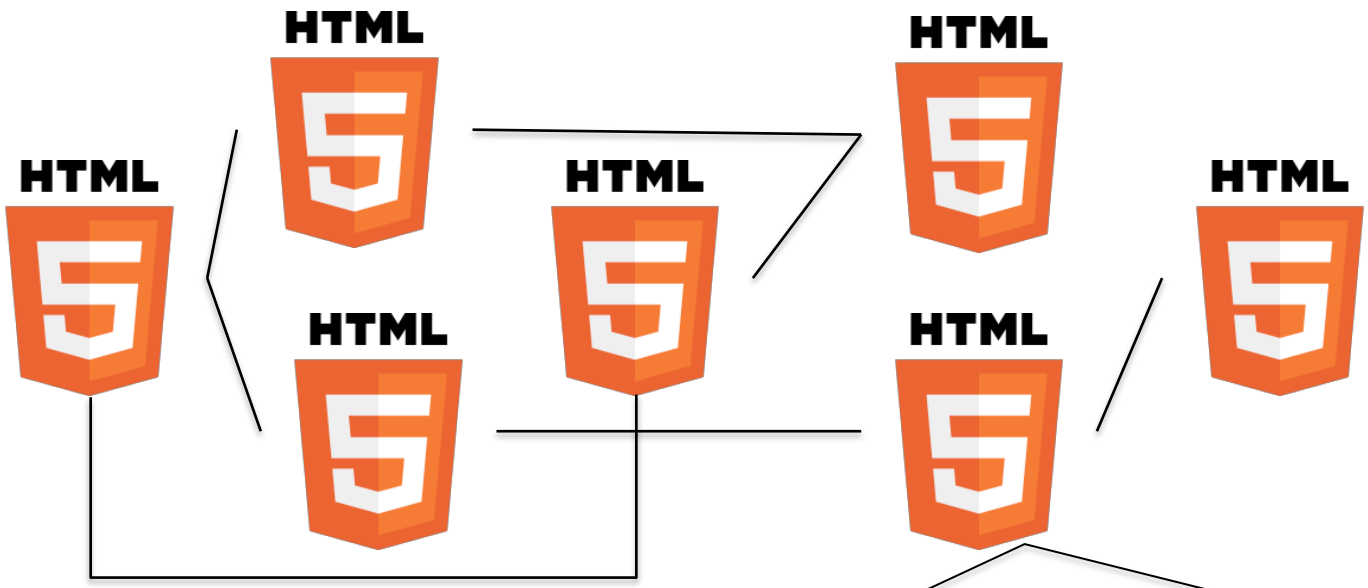
Semantic Web and Language Technologies

Felix Sasaki
DFKI / W3C Fellow
felix.sasaki@dfki.de

Overview

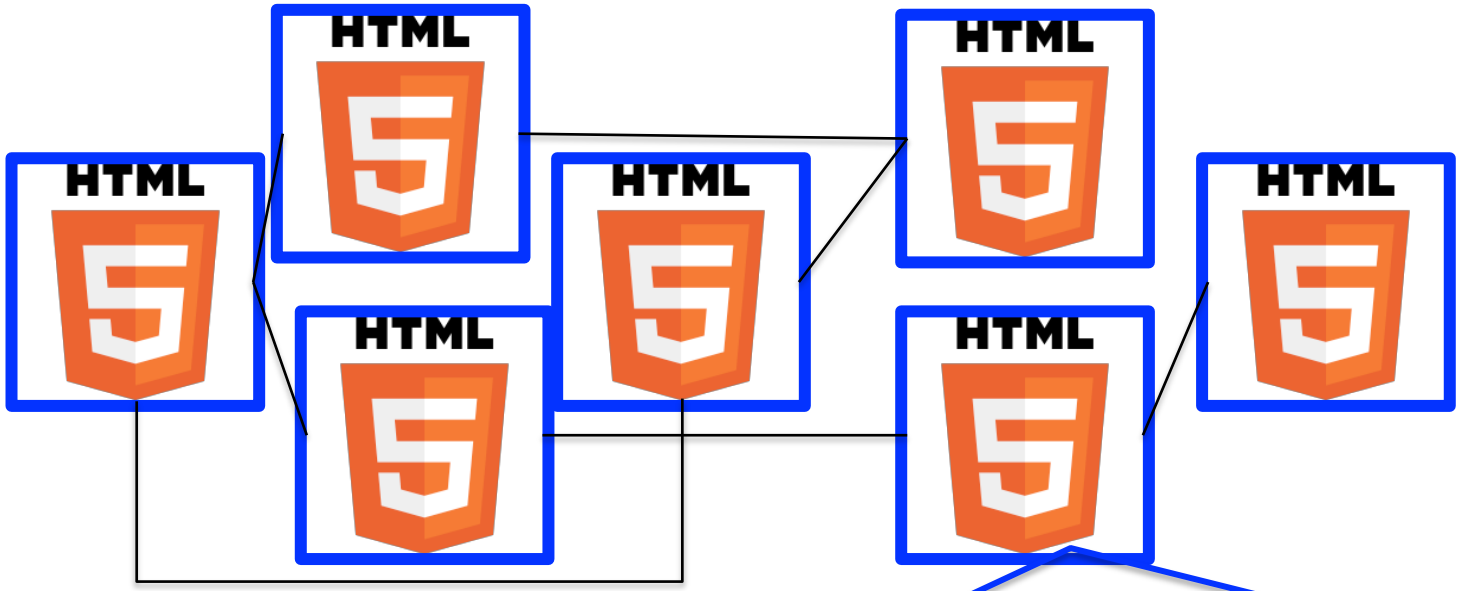
- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

Building blocks of the Web



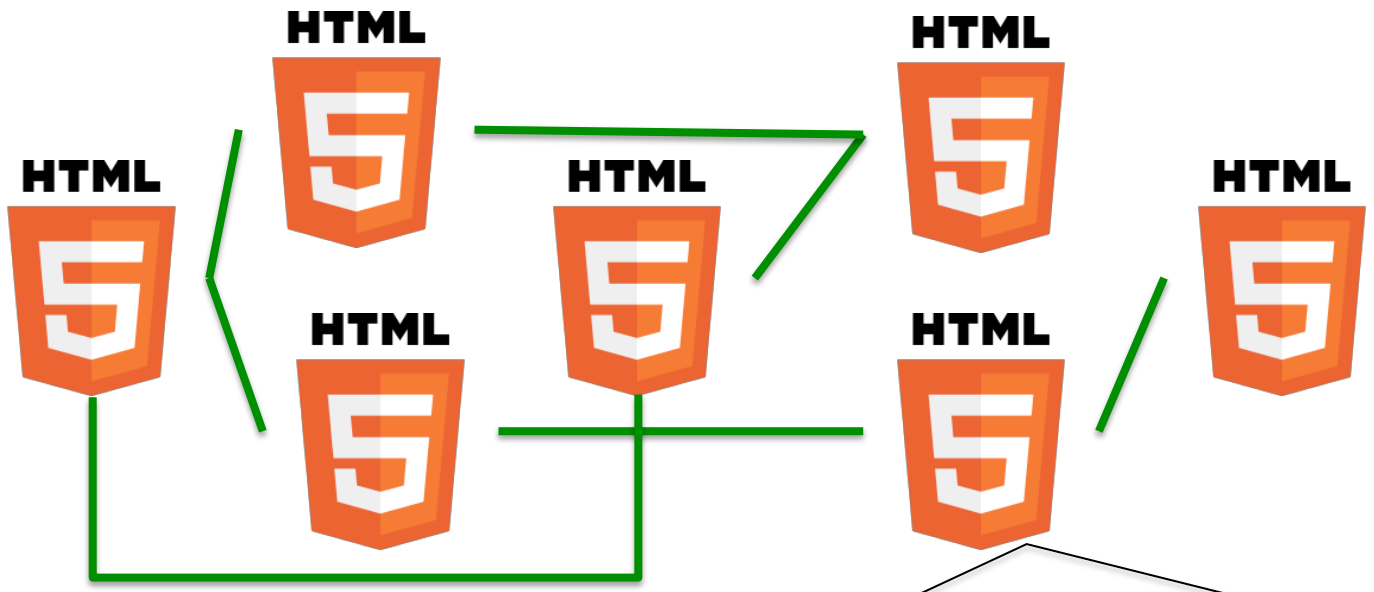

```
<p>All content on this site is licensed under
<a
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

Content



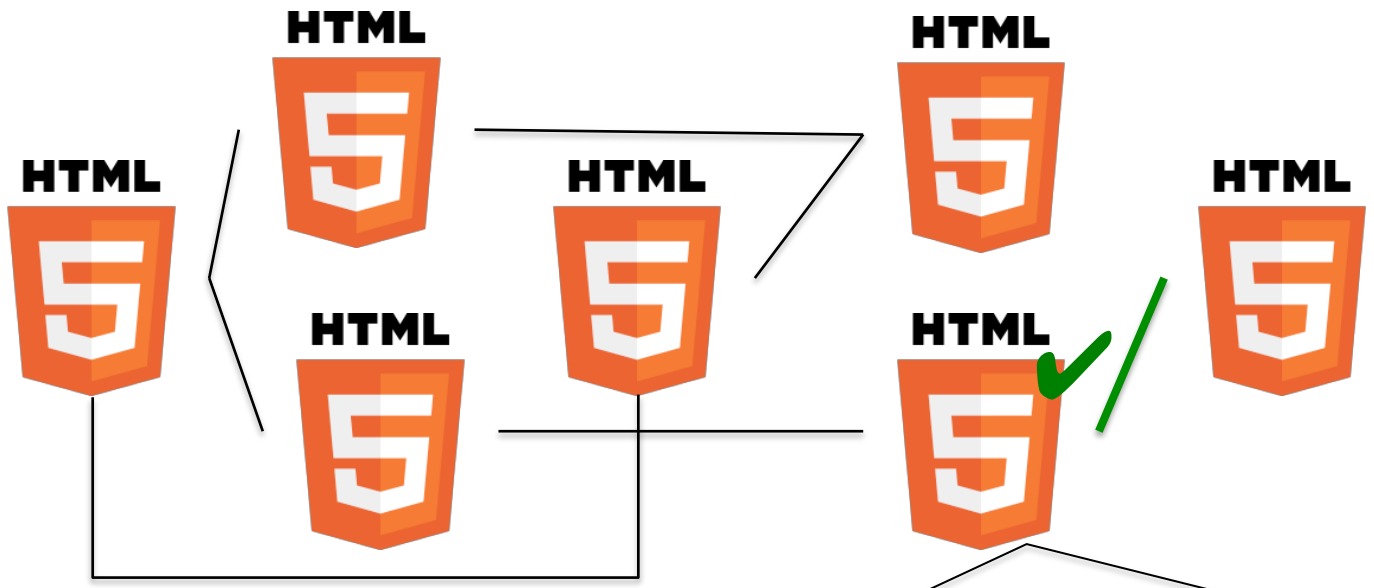

```
<p>All content on this site is licensed under
<a
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

Links (or “identifiers”)



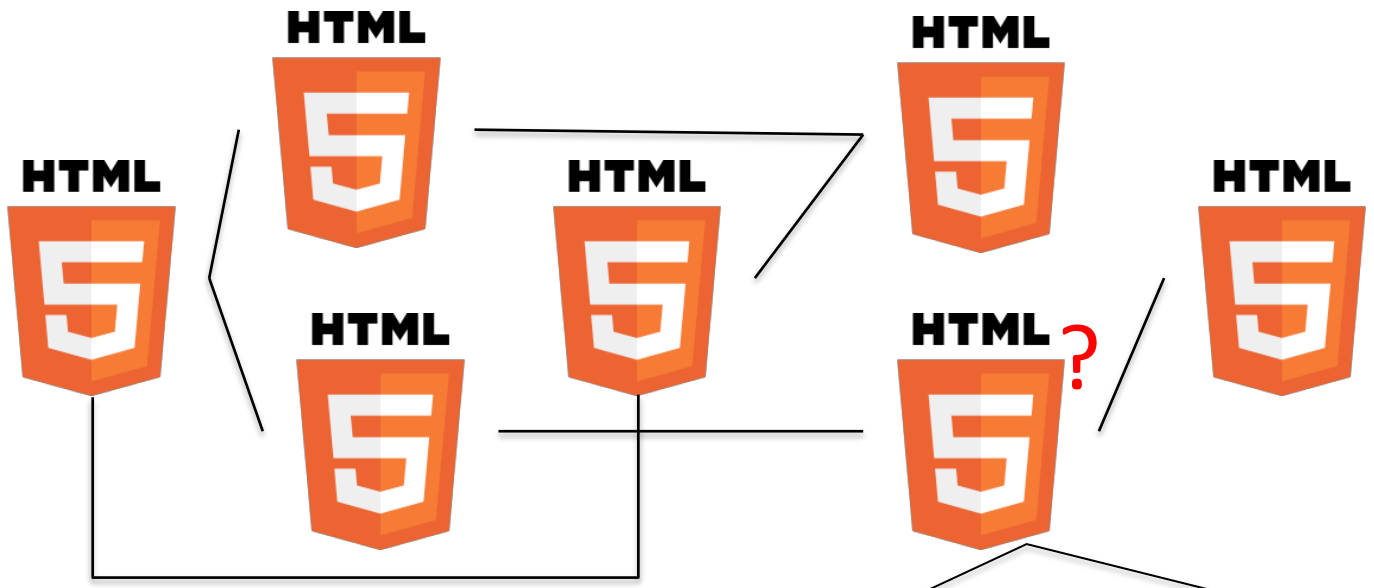

```
<p>All content on this site is licensed under
<a
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

Easy: “Find all content that links to http://creativecommons.org/licenses/by/3.0/“



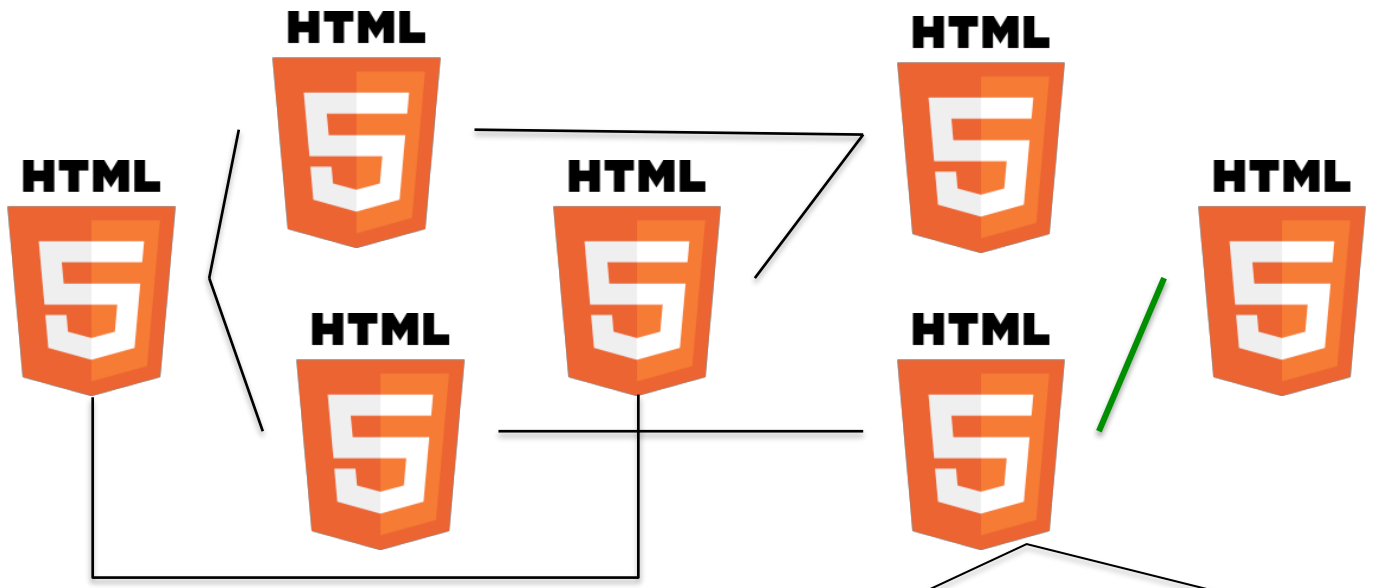

```
<p>All content on this site is licensed under
<a
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

Still difficult: “Find all content that links to a creative commons license“




```
<p>All content on this site is licensed under
<a
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

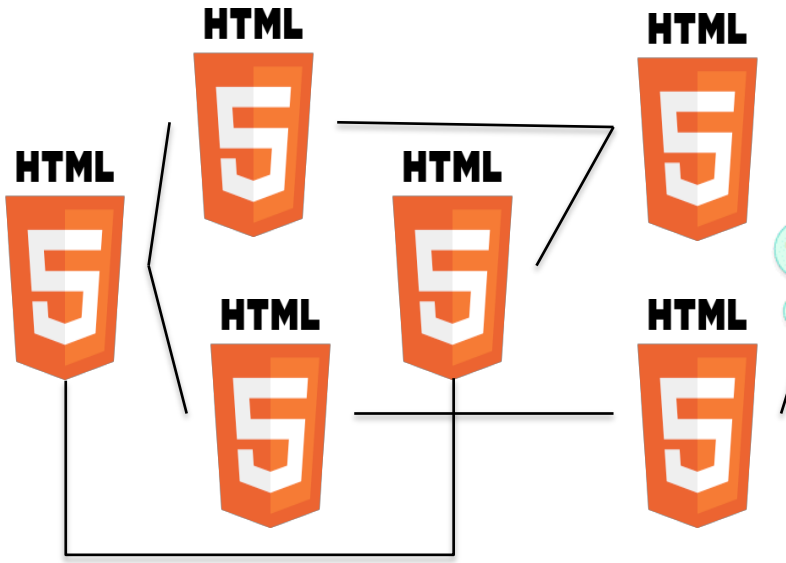
Semantic Web to the rescue = Providing machine readable information on the Web



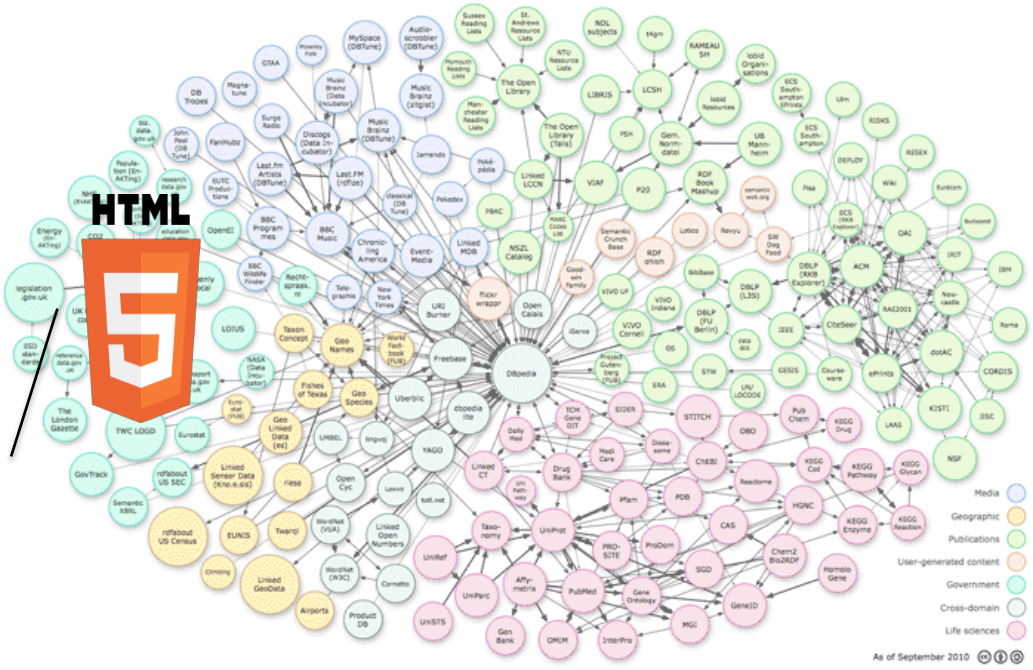

```
<p>All content on this site is licensed under
<a property="http://creativecommons.org/ns#license"
 href="http://creativecommons.org/licenses/by/3.0/">
   a Creative Commons License</a>. </p>
```

Semantic Web = Providing machine readable information on the Web

Web of documents



Web of data



Overview

- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

Tasks …

Write Semantic Web data

Create Semantic Web vocabularies

Representation and data creation

Query

… and technologies: past and presence

Write Semantic Web data: RDF (Resource Description Framework)

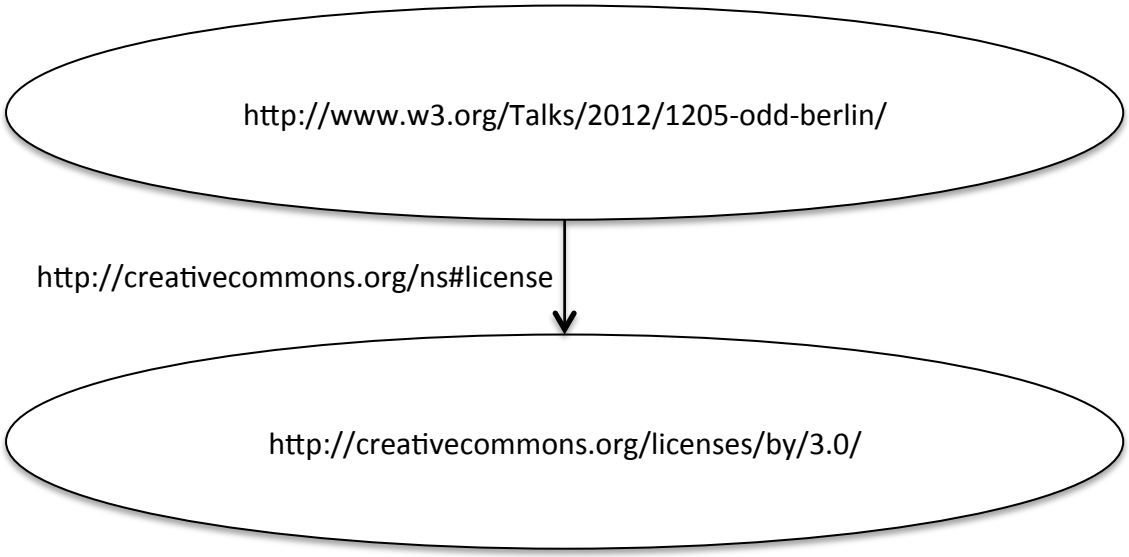
Create Semantic Web vocabularies: RDFS, SKOS, OWL (for complex ontologies)

Representation and data creation: Turtle, RDFa, R2RML, …

Query: SPARQL

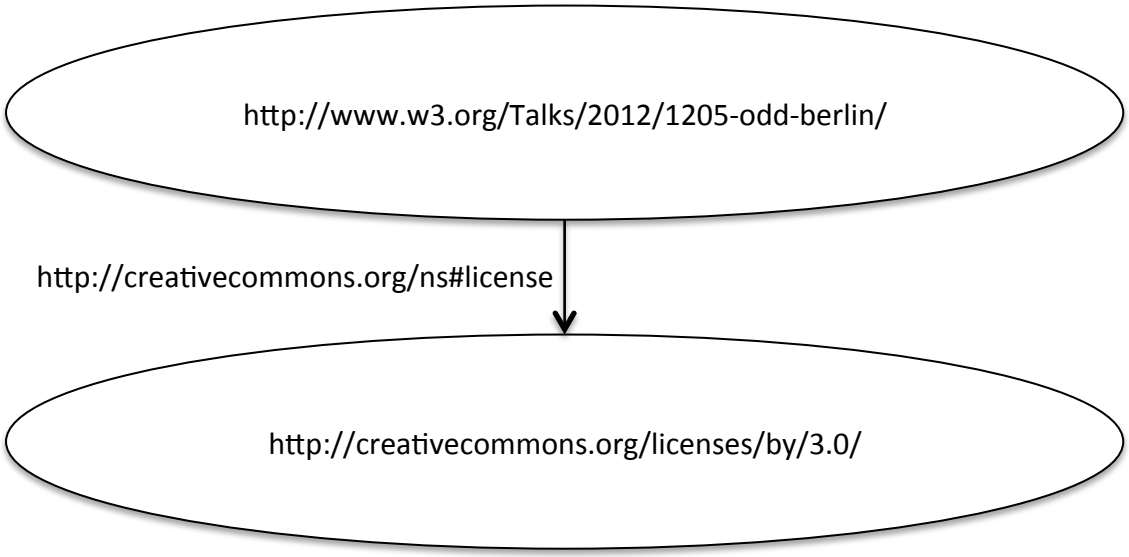
Write Semantic Web Data: RDF “statements”

Reference to license CC BY with an RDF statement (visualization):



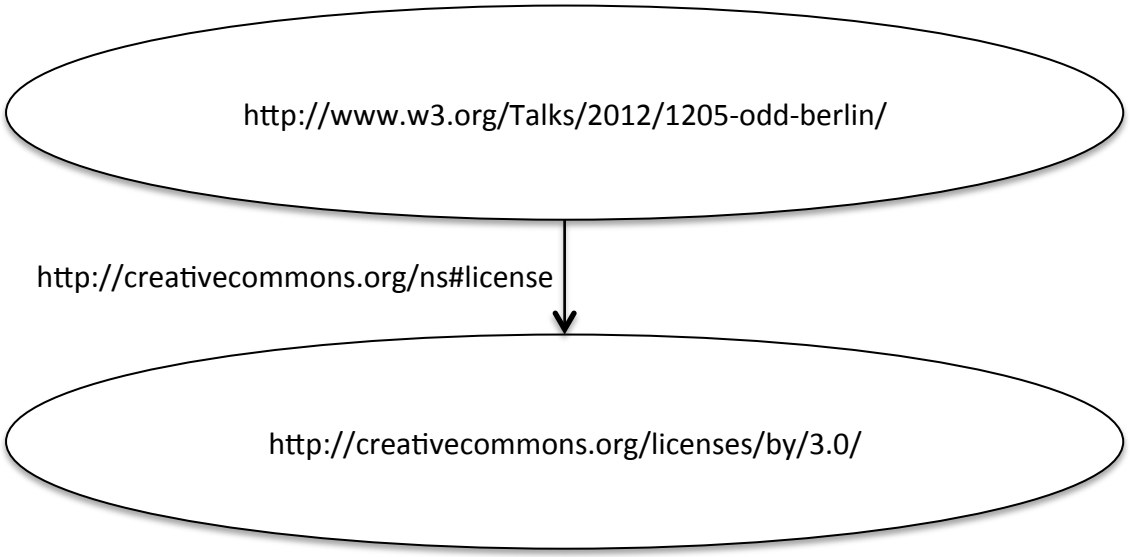
Turtle syntax

Reference to license CC BY with an RDF statement (visualization) + Turtle syntax:

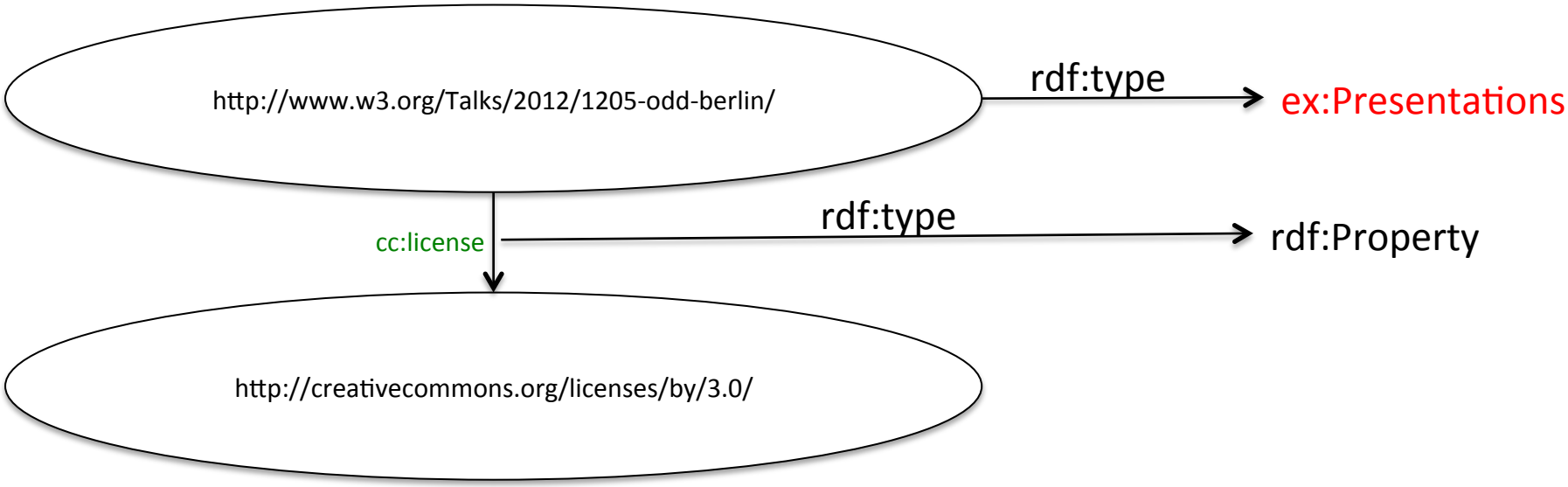



```
@prefix cc: <http://creativecommons.org/ns#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
<http://www.w3.org/Talks/2012/1205-odd-berlin/>
  cc:license
  <http://creativecommons.org/licenses/by/3.0/>.
```

RDF statements …



… can be based on vocabularies



RDF Schema

- For defining Classes (example “Presentations”) and properties (like “cc:license”)

OWL (Web Ontology Language)

- For defining further constraints for vocabularies

SKOS

- For describing e.g. thesauri, taxonomies, classification schemes

Ways to create and store Semantic Web data

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:cc="http://creativecommons.org/ns#">
 <rdf:Description
 rdf:about="http://www.w3.org/Talks/2012/1205-odd-berlin/">
```

RDF/XML

```
@prefix cc: <http://creativecommons.org/ns#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
<http://www.w3.org/Talks/2012/1205-odd-berlin/>
  cc:license
  <http://creativecommons.org/licenses/by/3.0/>.
```

Turtle

```
<p>All content on this site is licensed under
<a property="http://creativecommons.org/ns#license"
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

RDFa (embedding in HTML)

- RDFa Lite 1.1 http://www.w3.org/TR/rdfa-lite/: one-to-one mapping to microdata
- Microdata to RDF http://www.w3.org/TR/microdata-rdf/: extracting RDF from microdata
- HTML Data Guide http://www.w3.org/TR/html-data-guide/: guidance about RDFa vs. microdata vs. microformats
- R2ML http://www.w3.org/TR/r2rml/: Mapping relational data bases to RDF

Query - SPARQL

- Query language for RDF
- Patterns in link (=graph) structures
- E.g. “find all presentations with CC BY license”

Query will return

http://www.w3.org/Talks/2012/1205-odd-berlin/

```
PREFIX cc: <http://creativecommons.org/ns#>
SELECT ?presentation WHERE {
?presentation cc:license <http://creativecommons.org/licenses/by/3.0/>.
}
```

SPARQL Query with dbpedia data set: “People who where born in Berlin before 1900”: http://tinyurl.com/born-in-berlin

SPARQL Explorer for http://dbpedia.org/sparql

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dbo: <http://dbpedia.org/ontology/>

SELECT ?name ?birth ?death ?person WHERE {
     ?person dbo:birthPlace :Berlin .
     ?person dbo:birthDate ?birth .
     ?person foaf:name ?name .
     ?person dbo:deathDate ?death .
     FILTER (?birth < "1900-01-01"^^xsd:date) .
}
ORDER BY ?name
```

Results: Browse Go! Reset

SPARQL results:

name	birth	death	person
""Helene" Ellen Franz"@en	"1839-05-30"^^xsd:date	"1923-03-24"^^xsd:date	:Ellen_Franz
"()"@en	"1811-10-29"^^xsd:date	"1873-06-06"^^xsd:date	:Prince_Adalbert_of_Prussia_(1811%E2%80%931873)
"(Carl Heinrich) Eduard Knoblauch Knoblauch"@en	"1801-09-25"^^xsd:date	"1865-05-29"^^xsd:date	:Eduard_Knoblauch
"Achim von Arnim"@en	"1781-01-26"^^xsd:date	"1831-01-21"^^xsd:date	:Ludwig_Achim_von_Arnim
"Adalbert Of Prussia"@en	"1811-10-29"^^xsd:date	"1873-06-06"^^xsd:date	:Prince_Adalbert_of_Prussia_(1811%E2%80%931873)
"Adam Heinrich Müller"@en	"1779-06-30"^^xsd:date	"1829-01-17"^^xsd:date	:Adam_M%C3%BCller
"Adam Müller"@en	"1779-06-30"^^xsd:date	"1829-01-17"^^xsd:date	:Adam_M%C3%BCller
"Adolf Christen"@en	"1811-08-07"^^xsd:date	"1883-07-13"^^xsd:date	:Adolf_Christen
"Adolf Heinrich von Arnim-Boitzenburg"@en	"1803-04-10"^^xsd:date	"1868-01-08"^^xsd:date	:Adolf_Heinrich_von_Arnim-Boitzenburg
"Adolf Otto Reinhold Windaus"@en	"1876-12-25"^^xsd:date	"1959-06-09"^^xsd:date	:Adolf_Otto_Reinhold_Windaus

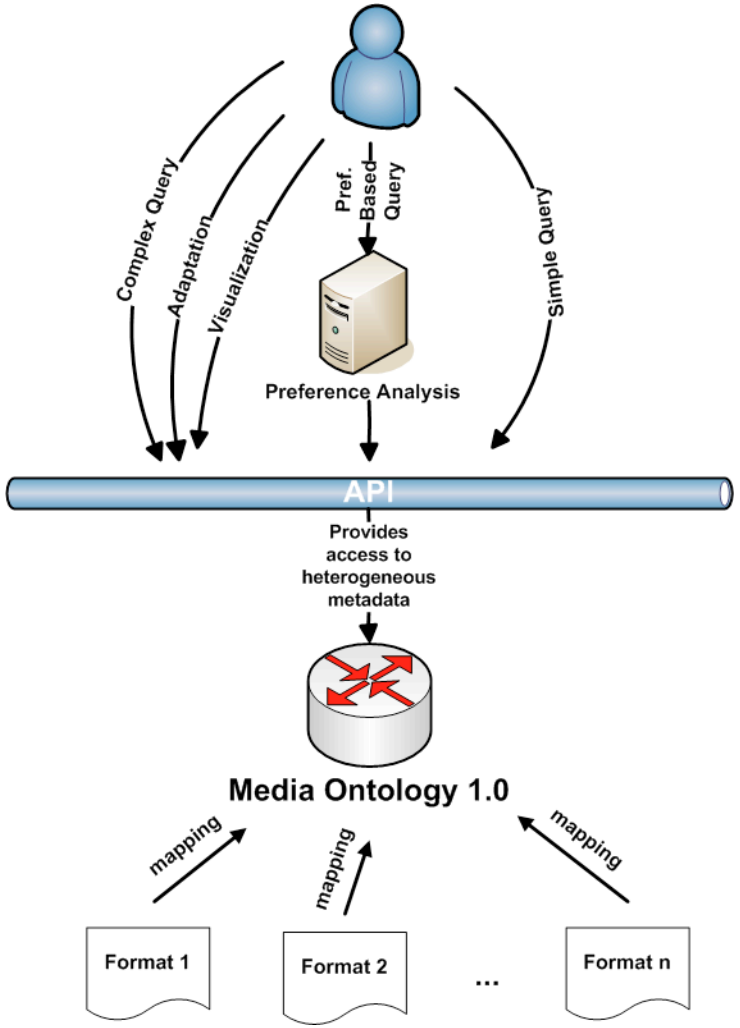
SOME ADDITIONAL BUILDING BLOCKS

Media Fragments URI 1.0 (basic)

- See http://www.w3.org/TR/media-frags/
- Fragments for identifying spatial and temporal dimensions of various medias
- Using these other HTTP

Base URI: http://www.example.com/example.ogv . Example fragments:
#t=10,20
#track=audio&t=10,20
#id=Cap%C3%ADtulo%202
xywh=160,120,320,240

Ontology for Media Resources 1.0



MAWG	Relation	EXIF 2.2
Descriptive Properties (Core Set)		
Identification		
identifier	exact	ImageUniqueID
title	more specific	ImageDescription, INAM
language	N/A	
locator	N/A	
Creation		
contributor	exact	IART, IENG, ISRC, ITCH
creator	more specific	IART, ISRC

Overview

- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

Vocabularies

- A growing set
- Various choices, e.g. syntax
- Application scenarios
  - eGovernment
  - Search Engine Optimization
  - …
  - LT

Not W3C, but related: Schema.org

- See http://www.schema.org/
- Collection of schemas
  - Creative works, person, place, product, ..
- Markup approach(es)
  - Microdata, RDFa Lite 1.1
- Recognized by major search engine providers

“Two syntax” approach

RDFa Lite 1.1

```
<p vocab="http://schema.org/" typeof="Person">
  My name is <span property="name">Felix Sasaki</span>.
</p>
```

Microdata

```
<p itemscope itemtype="http://schema.org/Person">
 My name is <span itemprop="name">Felix Sasaki</span>.
</p>
```

Schema.org: *not* W3C work, but also discussed in W3C Web Schemas task force

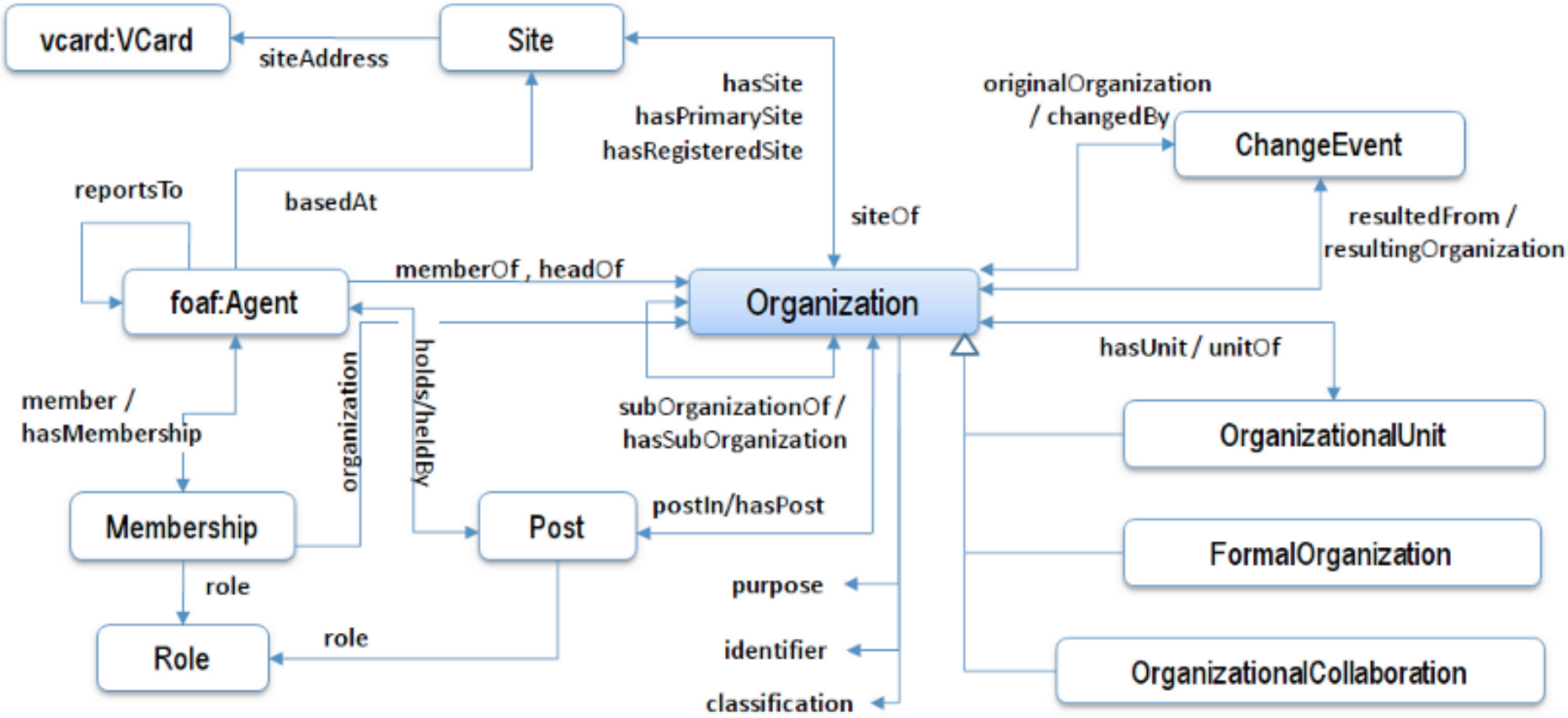
- See http://www.w3.org/wiki/WebSchemas
- Discussing extension proposals and mappings
- Example extension (proposals)
  - rNews (collab. with IPTC)
  - Health and Medical (collab. with W3C health care and life sciences Interest Group)
  - GoodRelations (collab. with GoodRelations)
- Public discussion list at http://lists.w3.org/Archives/Public/public-vocabs/

eGov @ W3C

- See http://www.w3.org/egov/
- Work on actual vocabularies, e.g. in the area of eGovernment
  - "An Organization Ontology"
    - http://www.w3.org/TR/vocab-org/
  - "RDF Data Cube Vocabulary"
    - http://www.w3.org/TR/vocab-data-cube/
  - "Data Catalog Vocabulary (DCAT)"
    - http://www.w3.org/TR/vocab-dcat/
- Discussing how to actually use open data
  - "Using Open Data" workshop June 2012 http://www.w3.org/2012/06/pmod/
- Actual standardization work (Working Group) and discussion form (Interest Group – please join ☺)

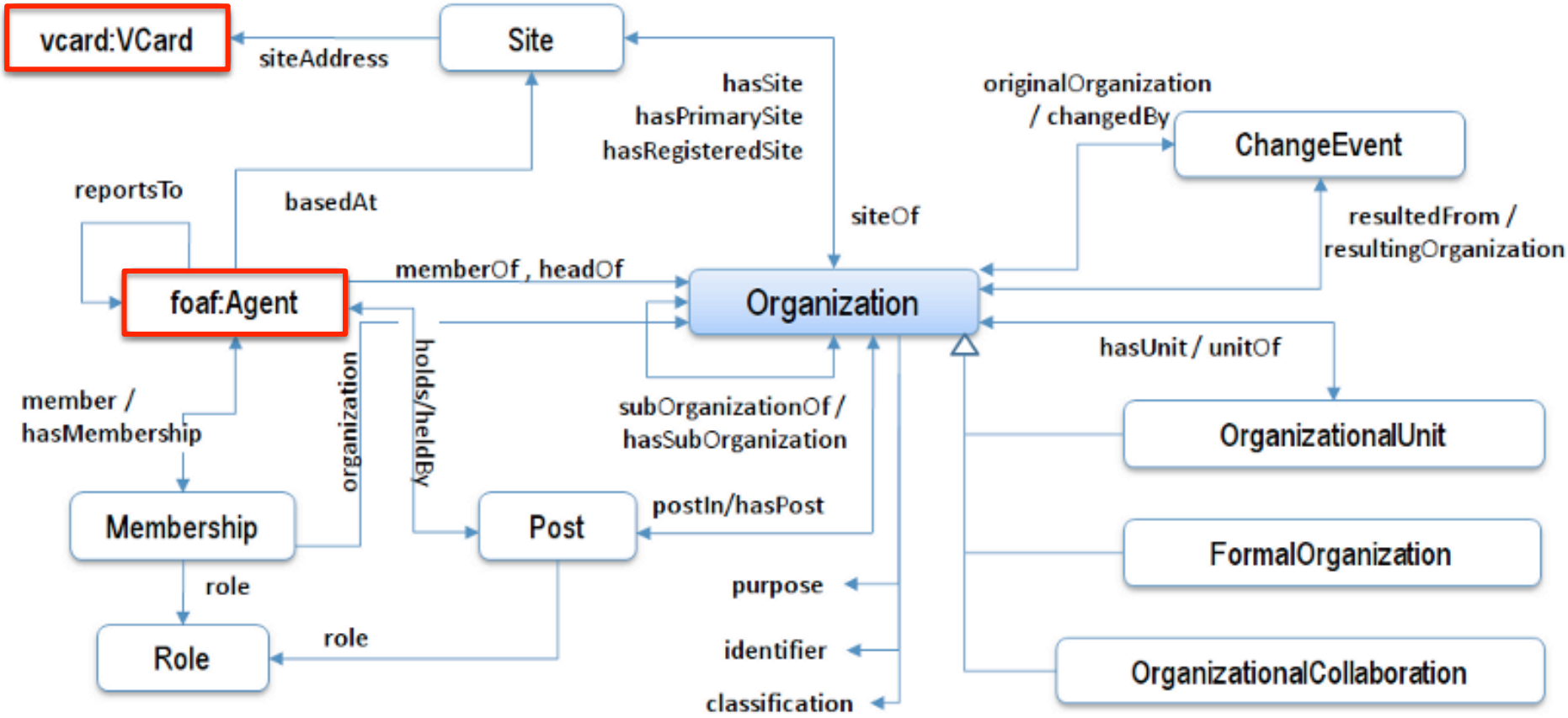
An Organization Ontology

- Organizational information across domains



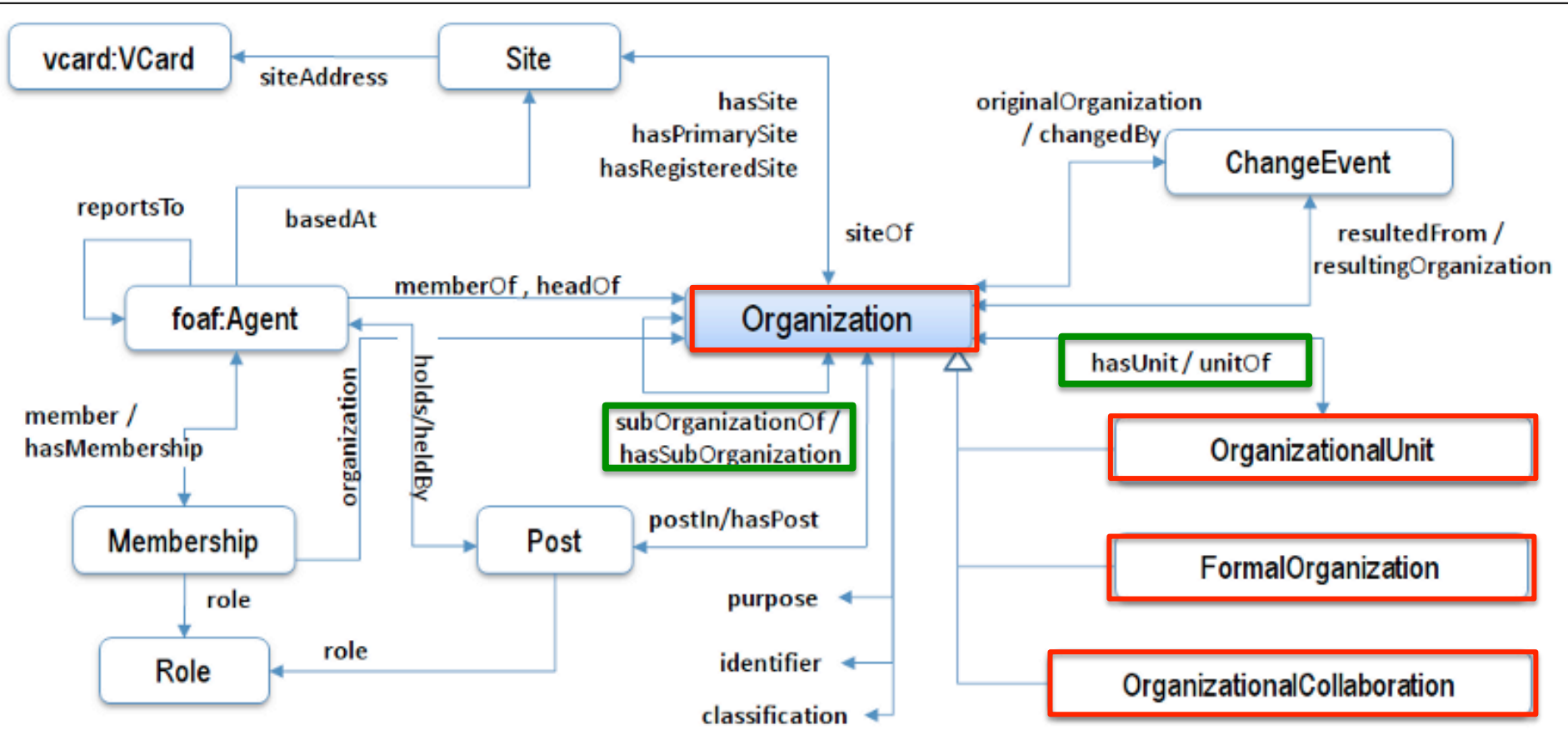
An Organization Ontology

- Re-use of vocabularies



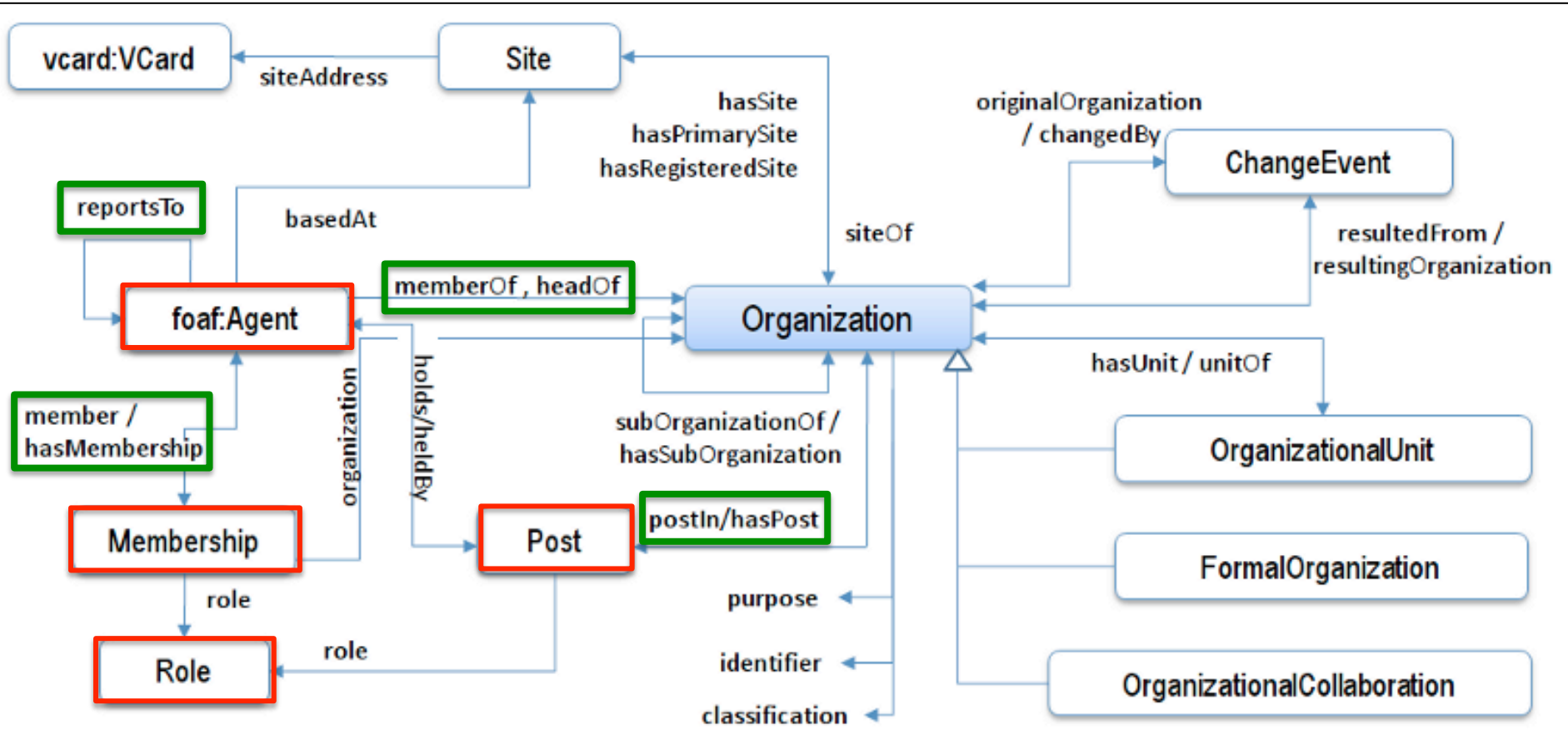
An Organization Ontology

- Organisational structure



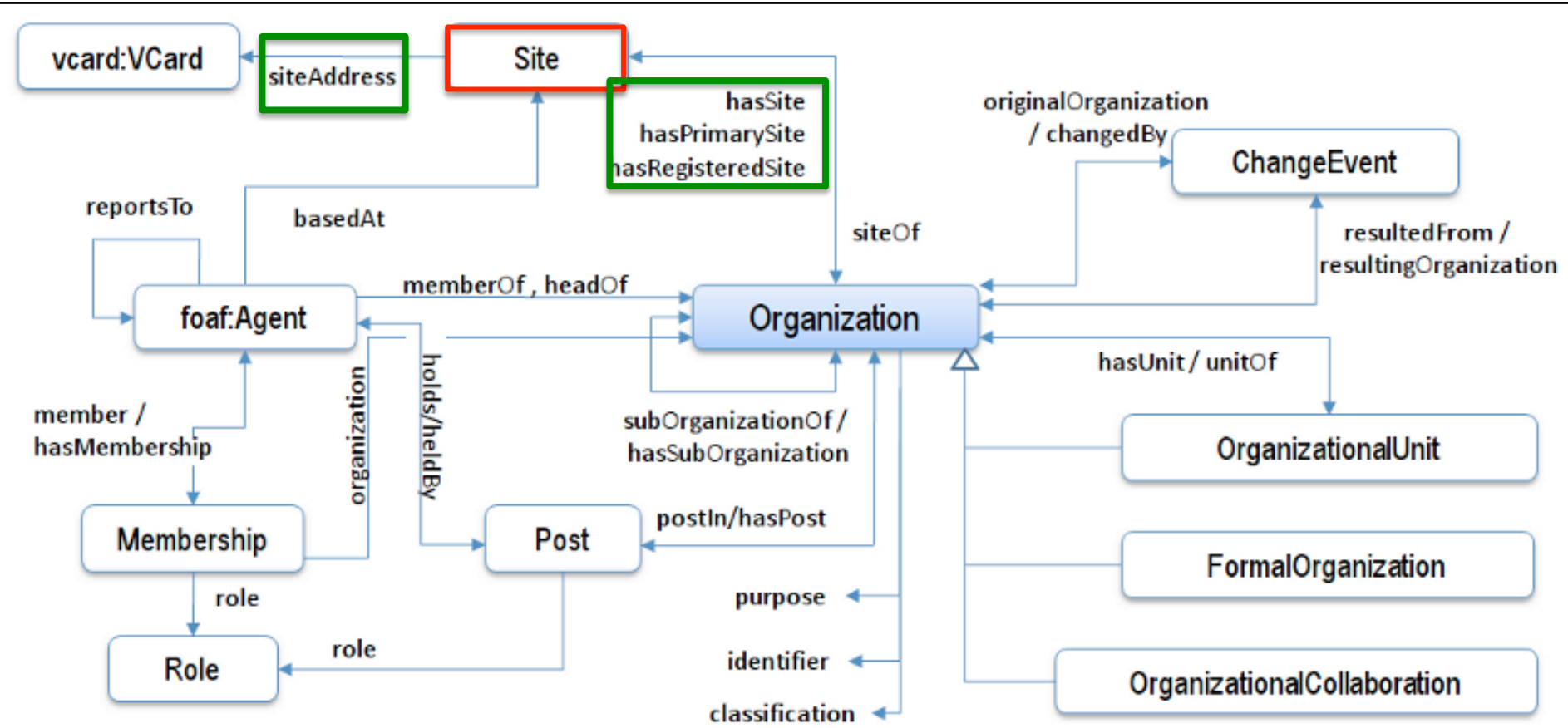
An Organization Ontology

- Membership and reporting structure



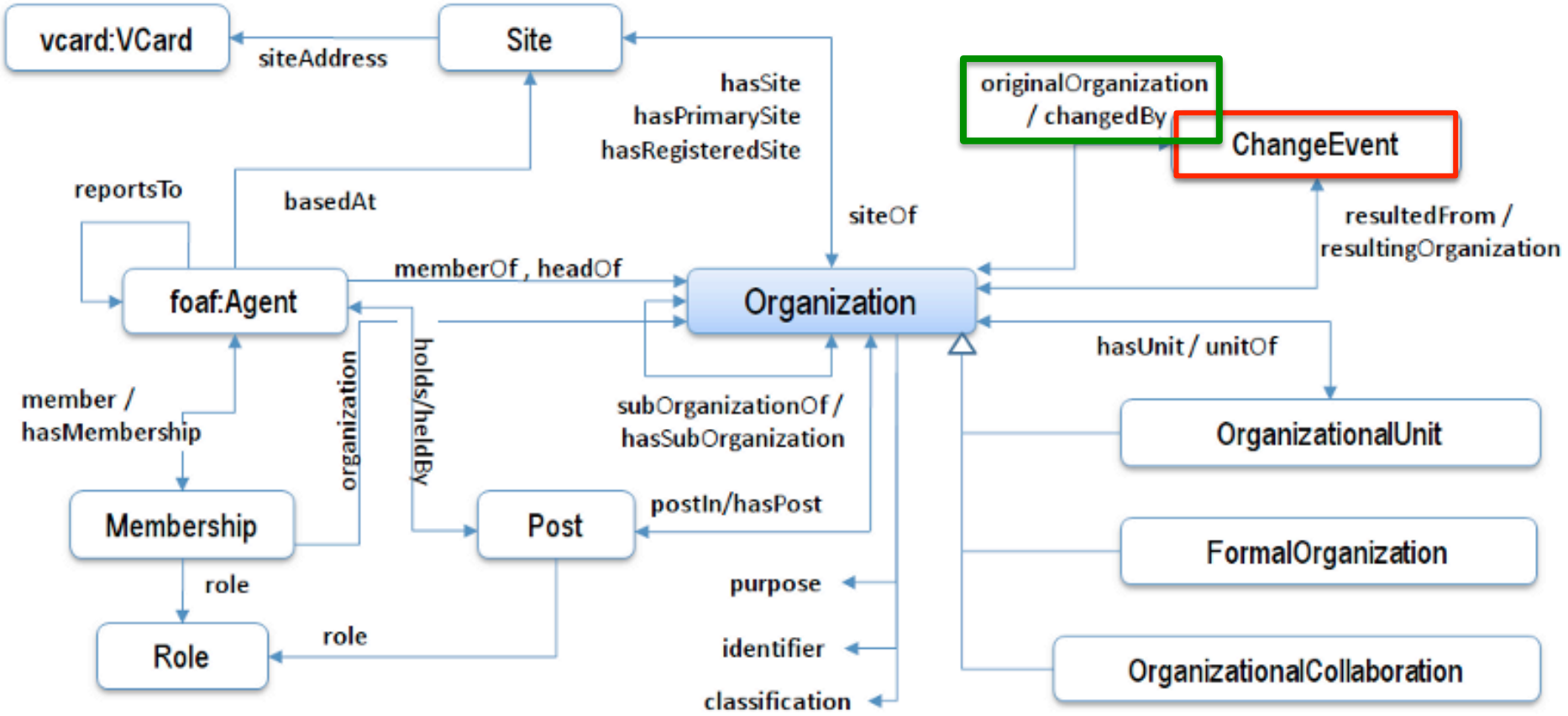
An Organization Ontology

- Location information



An Organization Ontology

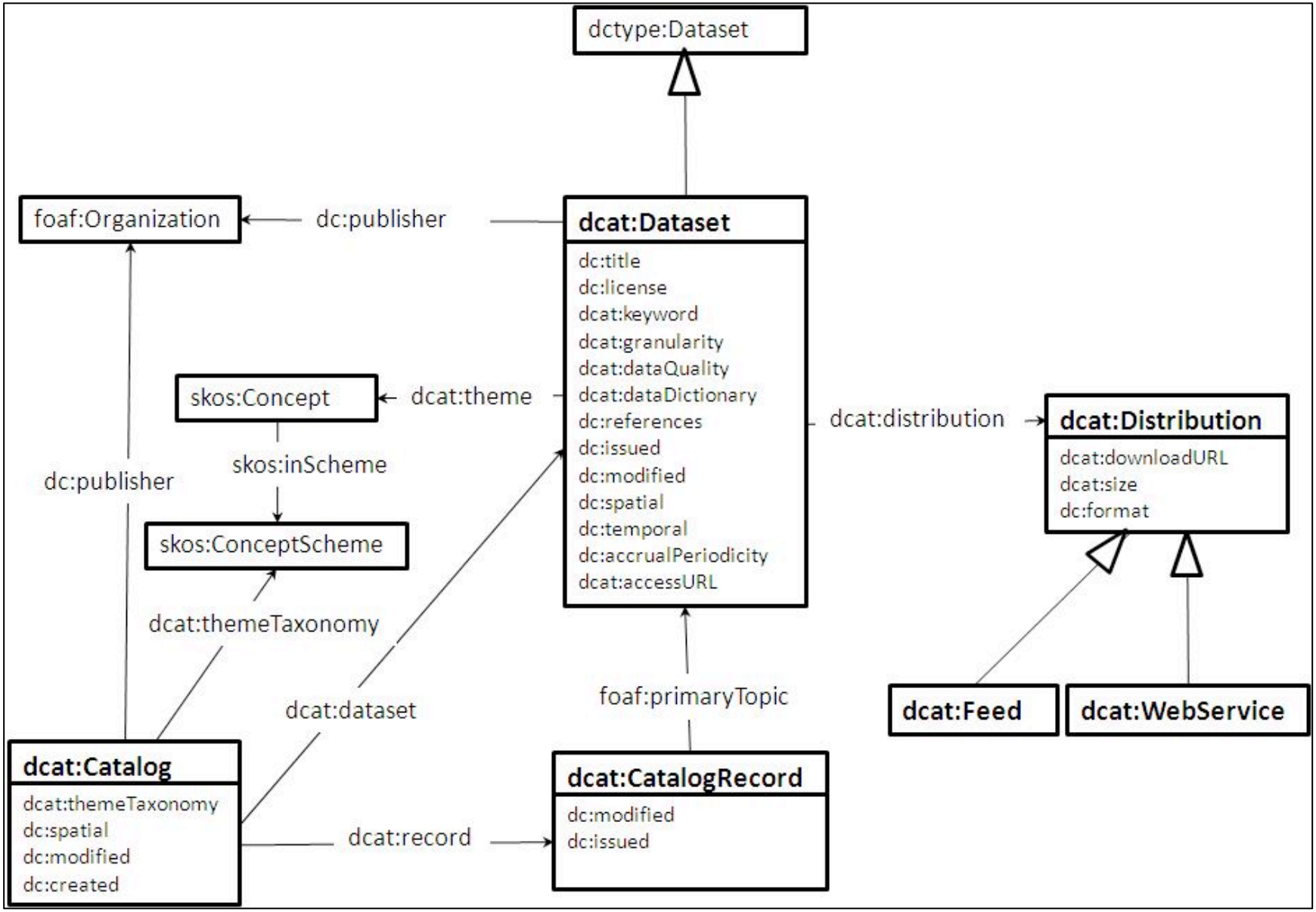
- Organizational History



Data Catalog Vocabulary (DCAT)

- Facilitate interoperability between data catalogs on the Web
- Uses existing vocabularies
  - FOAF
  - Dublin Core
  - SKOS

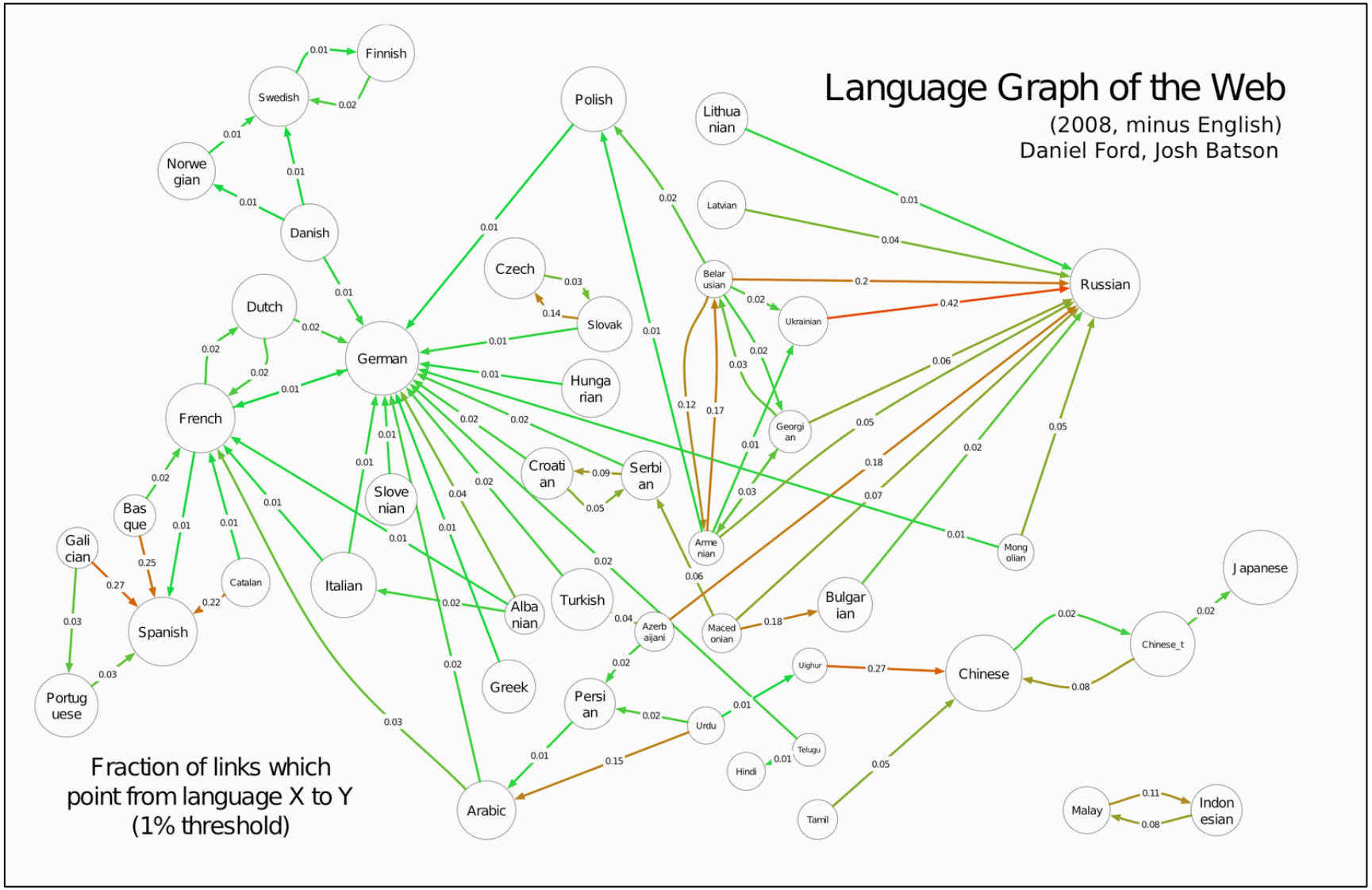
DCAT overview



Overview

- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

Multilinguality?



Source: http://googleresearch.blogspot.de/2011/07/languages-of-world-wide-web.html

Multilinguality and the Semantic Web?

"4.78% human readable descriptions have one language tag"

"Only 0.7% datasets contain several language tags"

"Most commonly language used:

44.72% (en), 5.22% (de), 5.11% (fr), 3.96% (it),..."

Source: Ell et al, 2011 Labels in the Web of Data, ISWC 2011

Internationalization Tag Set 2.0

- See http://www.w3.org/TR/its20/
- Defining metadata for multilingual processing of Web or other content, e.g. via
  - Machine translation
  - Localization workflows
- Where is the metadata needed – *for example*:
  - In Web content, e.g. HTML5
  - In XML-based and localization related formats
  - In the Semantic Web

Metadata example: “Translate” in HTML5 and XLIFF

```
<!DOCTYPE html>
<html> ...
<p>The <span translate=no>World Wide Web Consortium</span> is making the World Web Web worldwide!</p>...</html>
```

```
<xliff ...> ...
 <trans-unit id="1">
  <source xml:lang="en">The <mrk mtype="protected">World Wide Web Consortium</mrk> ...!</source>
  <target> ...
</xliff>
```

Metadata example: “Terminology” in HTML5 and XLIFF

```
<!DOCTYPE html>
<html> ...
<p>We need a new <span its-term=yes>motherboard</span>
...</html>
```

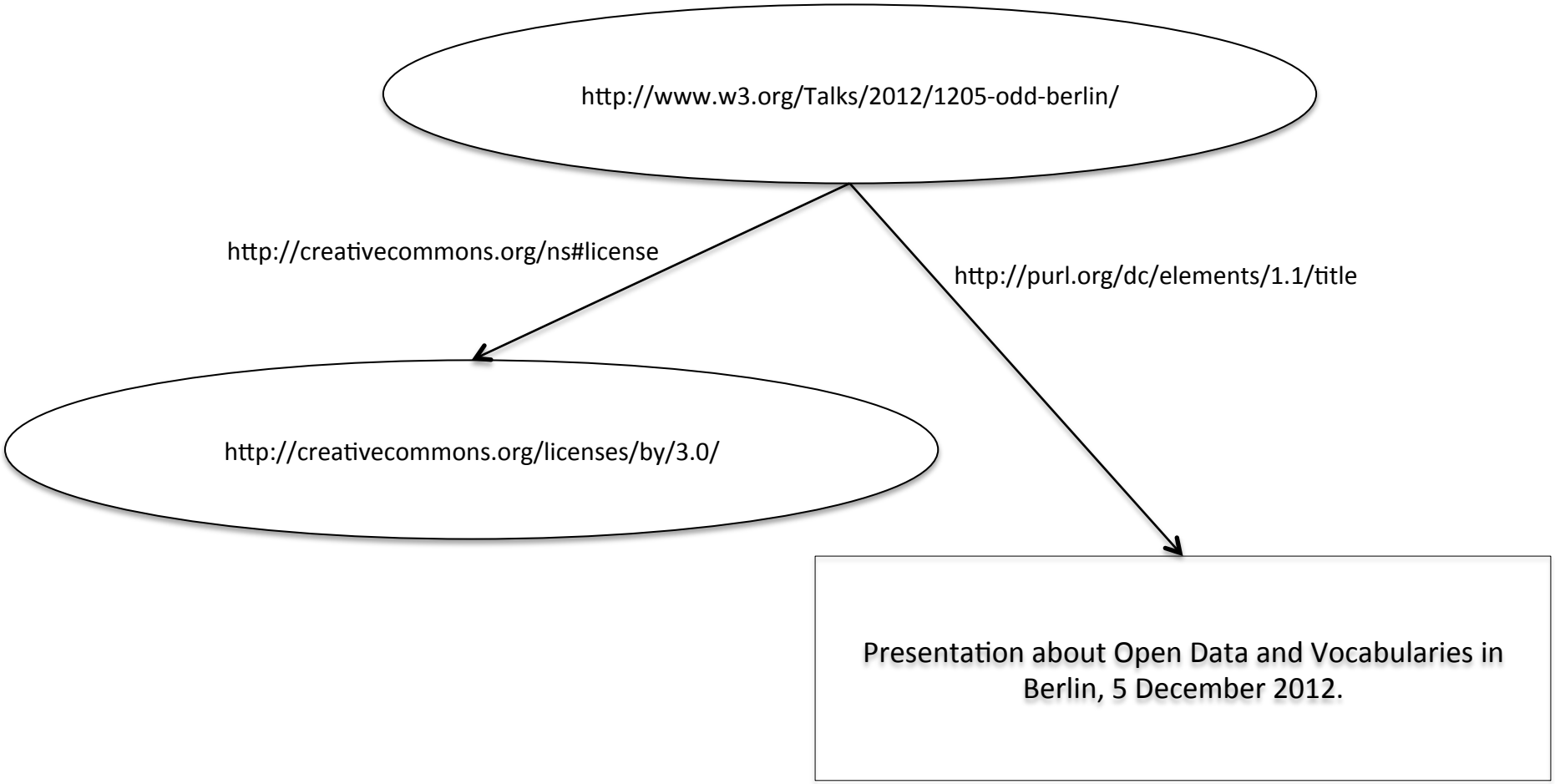
```
<xliff ...> ...
 <trans-unit id="1">
  <source xml:lang="en“>We need a new
   <mrk mtype="term">motherboard</mrk></source>
  <target> ...
</xliff>
```

Metadata example: “Disambiguation” in HTML5

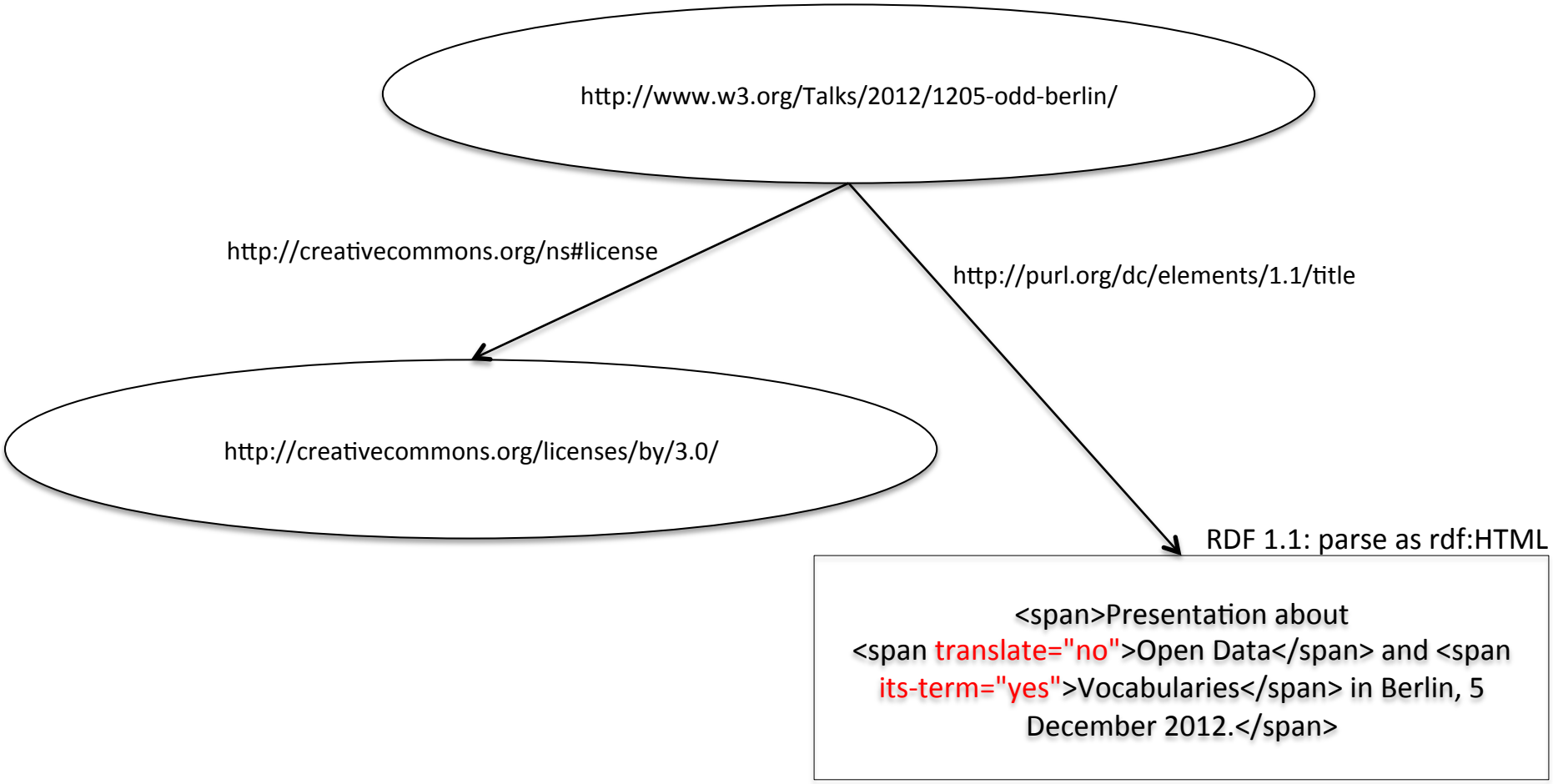
- Identify special conceptual patterns (e.g. named entities, lexical concepts)
- Add further information (e.g. class of entity, unique identifier)

```
<html> .... <p><span ...
    its-disambig-class-ref="http://nerd.eurecom.fr/ontology#Place"
    its-disambig-ident-ref="http://dbpedia.org/resource/Dublin"
    its-disambig-granularity="entity">Dublin</span> is the <span
       its-disambig-source="Wordnet3.0"
       its-disambig-ident="301467919"
       its-disambig-granularity="lexical-concept“ ... >capital</span>
of Ireland.</p> </body></html>
```

ITS 2.0: Metadata for localizing the Semantic Web



ITS 2.0: Metadata for localizing the Semantic Web



Overview

- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

LT using Semantic Web

- Question Answering over the Semantic Web
  - Example: IBM Watson
  - Using e.g. connection between linked data and natural language text in Wikipedia
  - Tasks: relation detection and entity recognition
- Generation of Natural Language from Linked Data
  - SW as an interlingua
  - RDF/OWL as a knowledge representation formalism
  - NLG from distributed ontologies

LT using Semantic Web

- Named Entity Disambiguation by exploiting domain knowledge
  - E.g. Wikipedia as semantic knowledge source
  - Example (demo): dbpedia spotlight http://dbpedia-spotlight.github.com/demo/
  - Potential: schema.org annotations for domain specific LT tasks, e.g.
    - rNews for news domain
    - Health and Medical for life sciences
    - GoodRelations for E-Commerce

LT using Semantic Web

- Reasoning upon facts extracted from text
- Temporal interpretation and reasoning for text analytic tasks
  - "Obama = president" not eternally true
  - Normalization of dates
  - Integration with existing facts

Semantic Web using LT

- Still rather a desire
- (Potential) Tasks
  - (schema.org) annotation quality checking
  - eGov (organization ontology, dcat) linking "ontologies to text"
  - Creating multilingual Semantic Web resources via LT (e.g. MT, see BabelNet http://lcl.uniroma1.it/babelnet/)
  - Creating annotations (e.g. ITS disambiguation) for further NLP processing, see upcoming slides

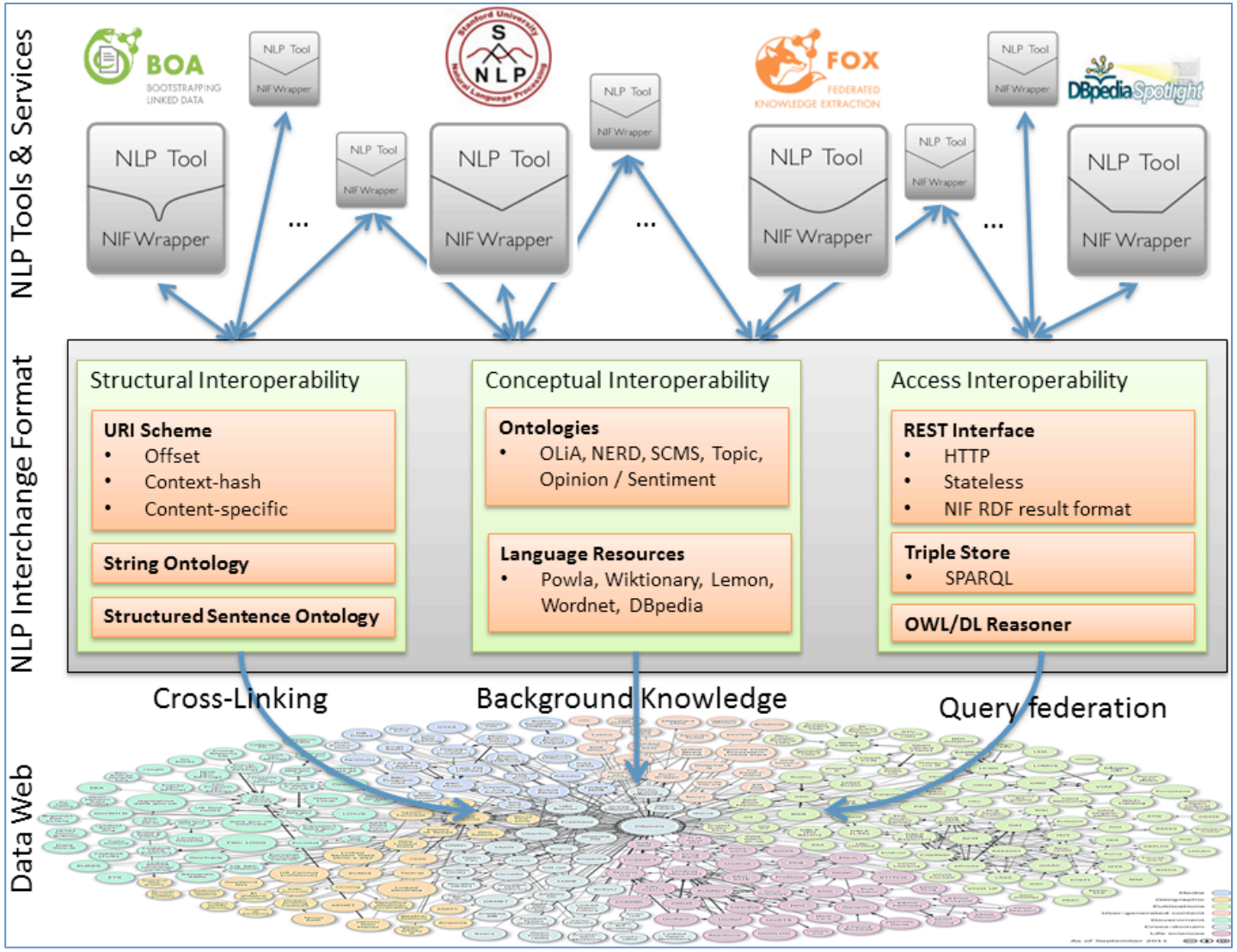
Overview

- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

LT processing

- Organized in workflows, relying on frameworks
  - UIMA
  - GATE
  - …
- Re-use of workflows hindered by data integration challenge
- RDF to the rescue: NLP Interchange Format (NIF) http://nlp2rdf.org/nif-1-0

NIF overview



NIF example: converting HTML5 with ITS 2.0 metadata to NIF

```
<html>...<body><h2 translate="yes">Welcome to <span
  its-disambig-ident-ref="http://dbpedia.org/resource/Dublin"
  translate="no">Dublin</span> in <b translate="no">Ireland</b>!
</h2></body></html>
```

NIF example: converting HTML5 with ITS 2.0 metadata to NIF

```
<html>...<body><h2 translate="yes">Welcome to <span
  its-disambig-ident-ref="http://dbpedia.org/resource/Dublin"
  translate="no">Dublin</span> in <b translate="no">Ireland</b>!
</h2></body></html>
```

```
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
# "Welcome to "
<http://example.com/exampledoc.html#xpath(/html/body[1]/
h2[1]/text()[1])>
   itsrdf:nif <http://example.com/exampledoc.html#offset_0_11> .
```

NIF example: converting HTML5 with ITS 2.0 metadata to NIF

```
<html>...<body><h2 translate="yes">Welcome to <span
  its-disambig-ident-ref="http://dbpedia.org/resource/Dublin"
  translate="no">Dublin</span> in <b translate="no">Ireland</b>!
</h2></body></html>
```

```
@prefix itsrdf: <http://www.w3.org/2005/11/its/rdf#> .
# "Welcome to “ ...
<http://example.com/exampledoc.html#offset_0_11>
 rdf:type str:String ; rdf:type str:OffsetBasedString ;
 itsrdf:translate "yes";
   str:referenceContext <http://example.com/
exampledoc.html#offset_0_29> .
```

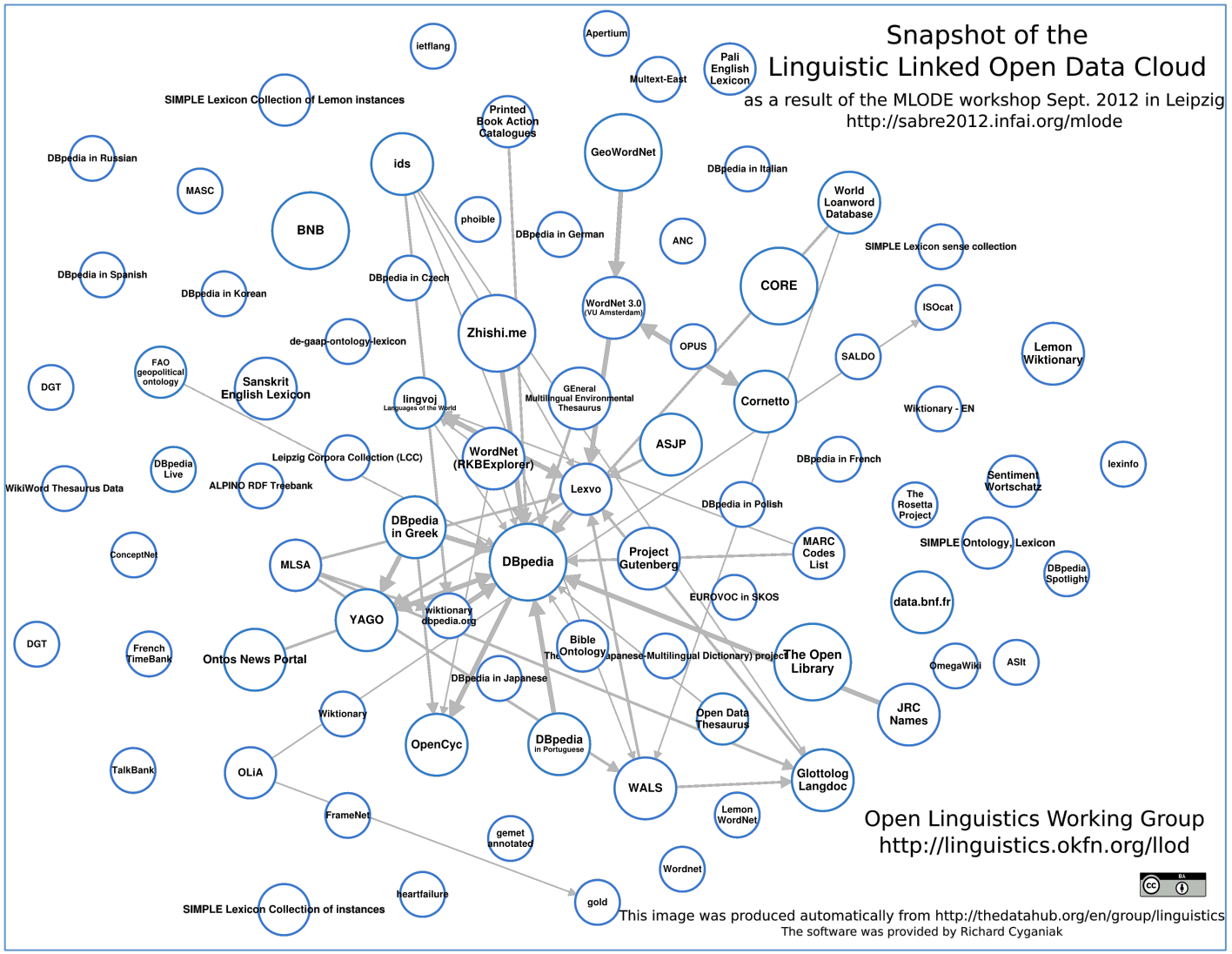
NIF processing: generating ITS 2.0 "disambiguation" metadata

```
# Input: <html>
 <body>
   <h2>Welcome to Dublin in Ireland!</h2>
 </body>
</html>
# ITS2NIF:
<http://example.com/exampledoc.html#xpath(/html/body[1]/
h2[1]/text()[1])>
   itsrdf:nif <http://example.com/exampledoc.html#offset_0_29>
# DBpedia Spotlight returns:
 <http://example.com/exampledoc.html#offset_21_28>
   itsrdf:disambigIdentRef  <http://dbpedia.org/resource/Ireland> .
# final step: NIF2ITS (generation of annotated HTML document)
```

Overview

- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

Linguistic open data cloud



Linguistic open data cloud

- Goal: represent linguistic resources as linked open data
- Potential input: see http://tinyurl.com/llod-input
- Issues
  - Data quality
  - Maintenance / persistence
  - Licensing
  - "What's a linguistic data set?" See http://tinyurl.com/llod-what

Potential for LT: authority files linked across languages via Wikipedia

Johann Wolfgang von Goethe

Goethe ist eine Weiterleitung auf diesen Artikel. Weitere Bedeutungen sind unter Goethe (Begriffsklärung) aufgeführt.

Johann Wolfgang von Goethe (* 28. August 1749 in Frankfurt am Main; † 22. März 1832 in Weimar), geadelt 1782, war ein deutscher Dichter. Er forschte und publizierte außerdem auf verschiedenen naturwissenschaftlichen Gebieten. Ab 1776 bekleidete er am Hof von Weimar unterschiedliche politische und administrative Ämter.

Goethes literarische Produktion umfasst Gedichte, Dramen, erzählende Werke (in Vers und Prosa), autobiografische, ästhetische, kunst- und literaturtheoretische sowie naturwissenschaftliche Schriften. Auch sein umfangreicher Briefwechsel ist von großer literarischer Bedeutung. Goethe war ein Vorreiter und der wichtigste Vertreter des Sturm und Drang. Sein Roman *Die Leiden des jungen Werthers* machte ihn 1774 in ganz Europa berühmt. Später wandte er sich inhaltlich und formal den Idealen der Antike zu und wurde ab den 1790er Jahren, gemeinsam mit Friedrich Schiller und im Austausch mit diesem, zum wichtigsten Vertreter der Weimarer Klassik. Im Alter galt Goethe auch im Ausland als Repräsentant des geistigen Deutschland.

Normdaten (Person): GND: 118540238 | LCCN: n79003362 | NDL: 00441109 | VIAF: 24602065 |

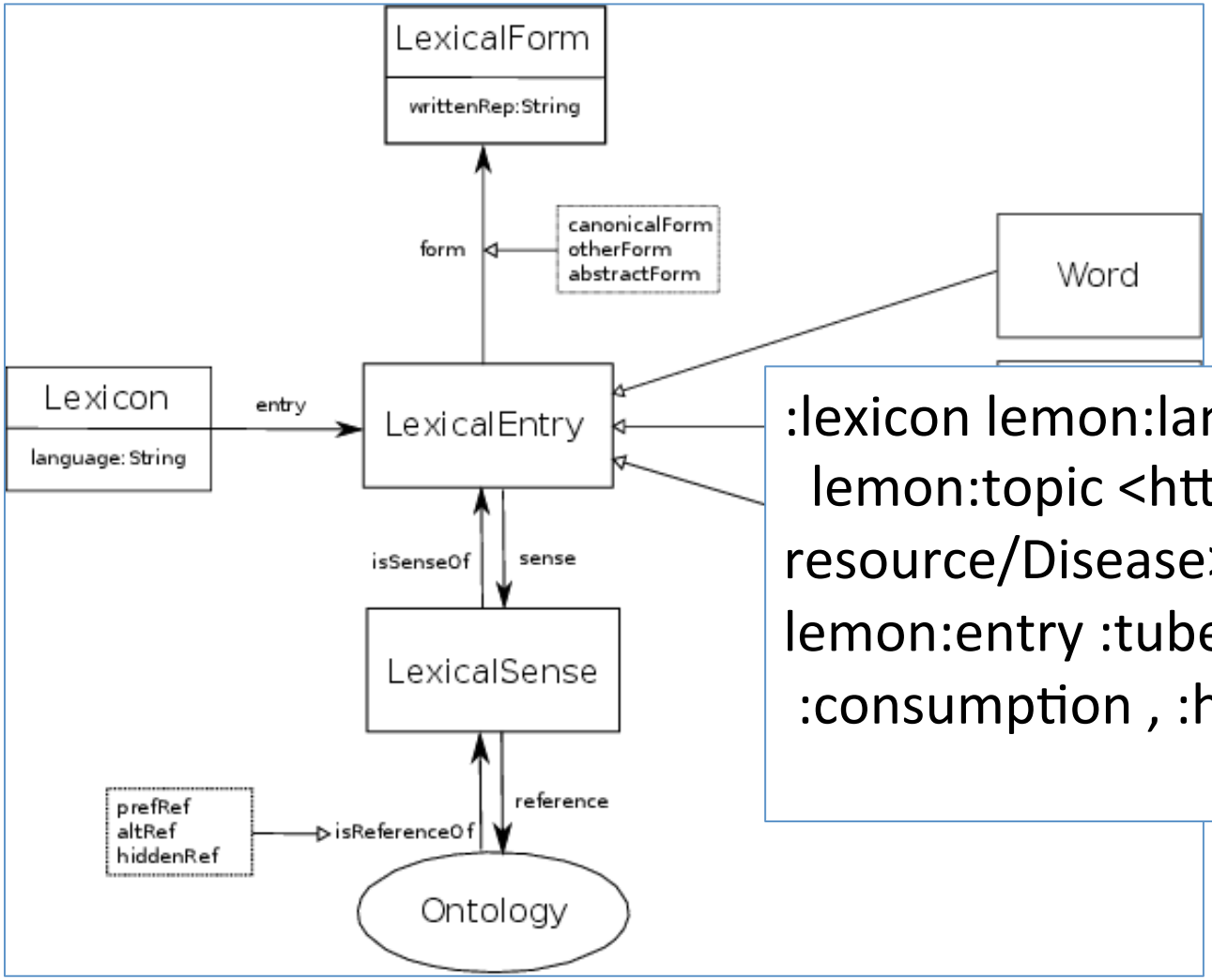
Potential for LT: authority files linked across languages via Wikipedia

Lexicon Model for Ontologies (LEMON)

- Representing lexical information in RDF, relative to ontologies
- Intro at http://www.lemon-model.net/5mins.html

```
:tuberculosis lemon:canonicalForm [
   lemon:writtenRep "tuberculosis"@en ] ;
  lemon:sense [
   lemon:reference <http://dbpedia.org/resource/Tuberculosis> ] .
```

Lexicon Model for Ontologies (LEMON)




```
:lexicon lemon:language "en" ;
  lemon:topic <http://dbpedia.org/
resource/Disease> ;
lemon:entry :tuberculosis ,
 :consumption , :hepatitis .
```

Terminological resources

- Many formats used in localization
  - TBX
  - OLIF
  - …
- TBX/RDF: approach to bring terminological resources to the Semantic Web

http://www.w3.org/International/multilingualweb/dublin/slides/09-melby.pdf

- Still in early stage

META-SHARE

- See http://www.meta-share.org/
- Network of repositories of language data, tools and related web services
- Not (yet) based on Semantic Web representation of resources
- But: holds the "missing block" of llod: curated, high-quality resources with clear licensing information

Overview

- What is Semantic Web?
- Technical building blocks
- Vocabularies
- The multilingual Semantic Web?
- LT using or processing Semantic Web
- Semantic Web and language processing workflows
- Linguistic resources and Semantic Web
- Current & future tasks

Current and future tasks

- Semantic Web data cleansing
- Creation of multilingual Semantic Web data
- Definition of a lexico-linguistic layer in the LOD cloud
- Integration with industry workflows (e.g. in localization / web technologies)
- Development of research & industry usage scenarios

Semantic Web and Language Technologies

Felix Sasaki
DFKI / W3C Fellow
felix.sasaki@dfki.de

Assignment for Wednesday

ITS 2.0 "Disambiguation" and (Semantic Web sensitive) LT tools

- "Disambiguation" defines various pieces of metadata, see http://www.w3.org/TR/its20/#EX-disambiguation-html5-local-1
  - its-disambig-confidence, its-class-ref, its-ident-ref, its-granularity, its-source, its-ident
- Look for LT tools that produce disambiguation information
- Create a table comparing output of these tools to the ITS 2.0 metadata pieces
- Take usage of Semantic Web resources (e.g. dbpedia) into account
- Be prepared to state your view about the usefulness of the ITS 2.0 "disambiguation" information