

Best Practices for Multilingual Linked Open Data

Dominic Jones, Jose E. Labra, Jorge Gracia

The result of numerous MLW workshops
and MLODE, Leipzig, Sept 2012

Purpose:

- Presentations from experts in the field.
- Open discussion around a number of topics.
- Collaborative editing of draft best-practices.
- Continuing (post-workshop editing of doc).
- Publishing via the MLW website for reference.

You choose to whether to put your name against the reference document as a **contributor**.

Discussion Points:

- Naming, URIs / IRIs
 - Use of full IRI's vs. ASCII
 - Opaque vs. descriptive URI's
 - Selection of the namespace
- Labeling content
 - Language tags
 - Labels vs. Longer Descriptions
 - Target User (author, developer, end user)
- Interlinking
 - Enriching vocabularies
 - Linking the same concepts in different languages (Different lexicalizations)
 - Leverage english resources for non-english LD
 - Language content negotiation.
- Quality issues
 - Datasets
 - Vocabularies
 - Quality benchmarking & provenance

Agenda:

- 14.30- 15.45 - 8 * 5min presentations + Q&A.
- 15.45- 16.15 – Coffee
- 16.15 - 17.15 – Discussion Collaborative editing of shared google doc.

Post workshop – continued editing and publishing as a reference document.

Order of Presentation

- Ivan Herman, “Towards Multilingual Data on the Web?” Semantic Web Activity Lead, W3C.
- Gordon Dunsire, “Multilingual bibliographic standards in RDF: the IFLA experience”, Independent Consultant; Chair of IFLA Namespaces Technical Group (Remote speaker).
- Daniel Vila, “Naming and Labeling Ontologies in the Multilingual Web”, Universidad Politécnica de Madrid, Spain.
- Dave Lewis, “XLIFF workflow and Multilingual Provenance in Linked Data”, Trinity College, Dublin, Ireland.
- Charles McCathie Nevile, Web Standards, Yandex.
- Roberto Navigli, "BabelNet: a multilingual encyclopedic dictionary as LOD", Sapienza University of Rome, Italy.
- Haofen Wang, “The state of the art of Chinese LOD development”, APEX labs, China Zhishi.me
- Jose E. Labra, “Patterns for Multilingual LOD: an overview”, University of Oviedo, Asturias, Spain.

Web Link

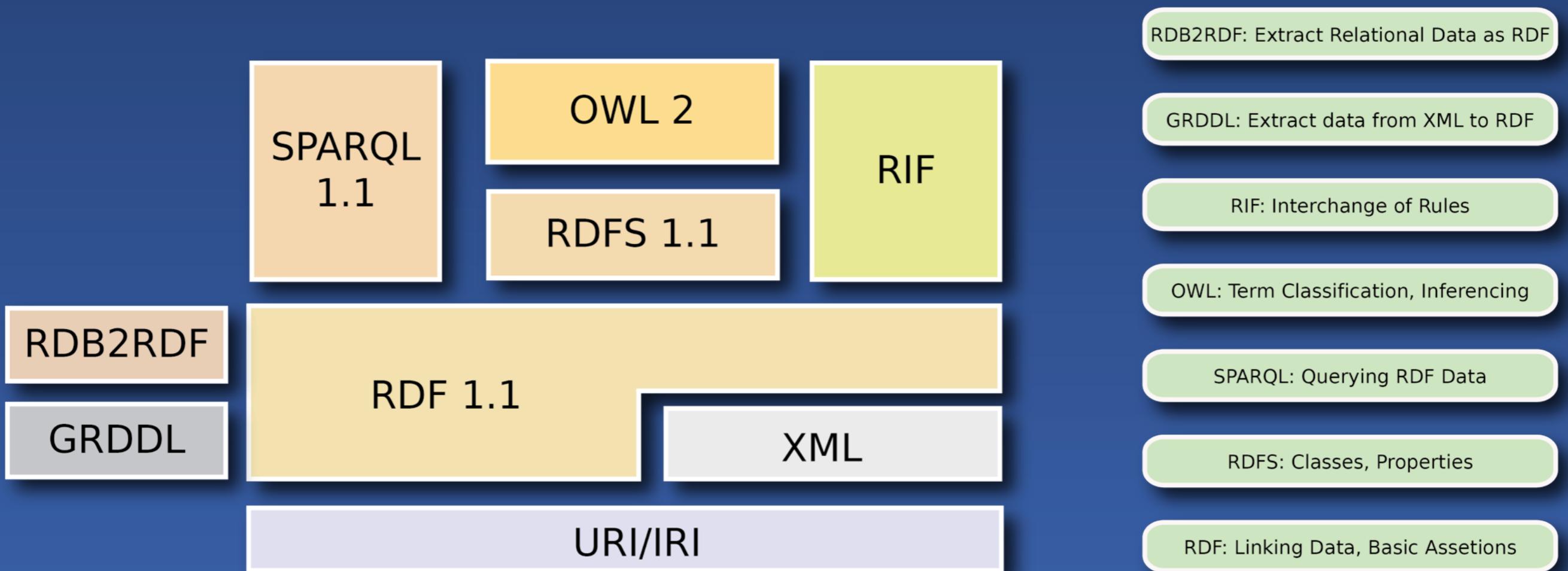
Go here: <http://goo.gl/Th2VA> to be part of the discussion!

Towards the (multilingual?) Data on the Web

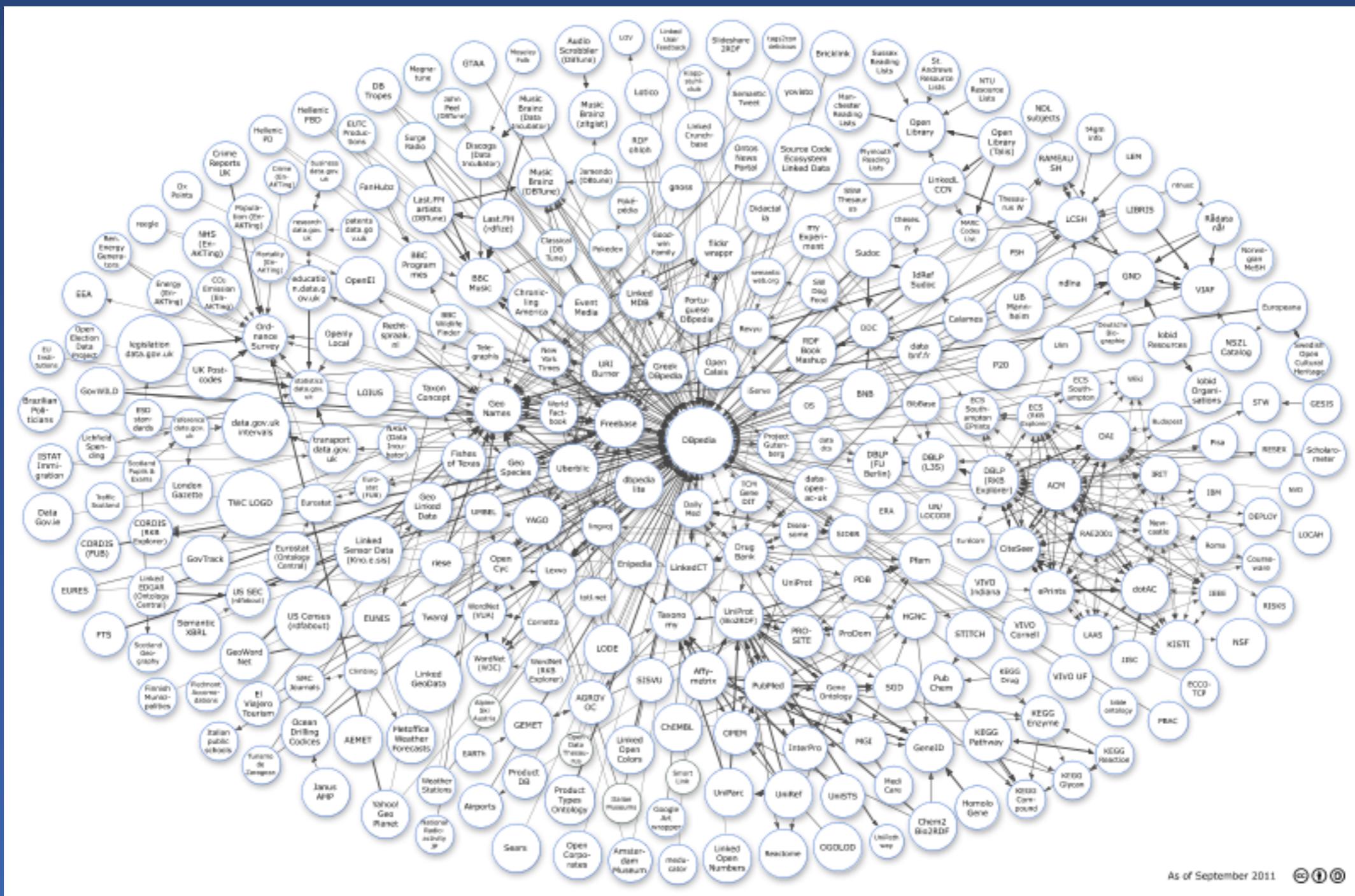
Ivan Herman
W3C

W3C

What we have today: technologies



What we have today: lots of datasets



- What the community needs is more *deployment*
 - use cases
 - more data
 - more *linked* data
 - etc.
- It is important that the underlying technology would be seen as stable

W3C's immediate plans

- Not to concentrate on new technology specifications
- Instead, look at the deployment issues
 - vocabulary definition, usage
 - outreach to different data formats

Vocabulary definitions

- The W3C Community Group structure gives an excellent environment to build vocabularies
 - good example: Open Annotation CG
- We would like to greatly extend this practice, possibly offering other tools (e.g., hosting of vocabularies)
- We are considering setting up some sort of a registry with metadata on the vocabularies
 - what would be a good set of metadata on the usability of a vocabulary in a multilingual environment?

Vocabulary validation

- Discussing the possibility of a workshop on vocabulary validation
 - “structural” validation against some schema-like definition
 - “quality” validation on data values, etc.
- Issue: how would one validate multilingual vocabularies?

Reaching out to other types of data

- *Data on the Web* is the really important thing
 - data may be in other formats: table, CSV, etc.
- There is a disconnect between:
 - the Linked Data and the more general Data on the Web worlds
 - the Web Developers' community and the Linked Data world

Open Data on the Web, 23 – 24 April 2013, London (Home)

FollowUp My Mercurial LocalData TR TR List 2013 Private Social Mailing lists SW Python RDFa it! Bookmarklets TOC

Reader

W3C www.w3.org/2013/04/odw/

W3C ODI Open Knowledge Foundation

Open Data on the Web 23 - 24 April 2013, Google Campus, Shoreditch, London



Home Program Committee Topics Who Should Attend? How to Participate Logistics Important Dates

Call for Participation

"Analysts believe an open data strategy should be a top priority for any organization that uses the Web as a channel for delivering goods and services. Open data strategies support outside-in business practices that generate growth and innovation."

— Press release "Gartner Says Big Data Makes Organizations Smarter, But Open Data Makes Them Richer", August 2012

Whether your business has an open data strategy today, or you want to learn from global experts how to publish or consume open data, the **World Wide Web Consortium (W3C)**, the **Open Data Institute** and the **Open Knowledge Foundation** invite you to join a discussion about realizing the promise of open data. Participants will help prioritize W3C's agenda in the area of data on the Web at a time when a number of working groups have either completed or are close to completing their charters. The Workshop, hosted by Google, is open to all and will take place in the heart of London's Tech City of Shoreditch.

Main Topics

From Toronto to Taipei, from Washington to Wellington via Westminster, data hubs are being created by public

Host

W3C gratefully acknowledges Google for hosting this workshop.



Sponsor



If you're interested in being a sponsor, please contact Alan Bird. For additional information, please visit the Sponsorship program.

Important Dates

Today

Questions arising for such a workshop

- What about the multilinguality of non-RDF, non-Linked Data?
 - how to check
 - how to create
 - is there *any* way to manage that properly?
 - reaching out to other types of data *across* languages?

**Looking forward to the
discussions!**

Multilingual bibliographic standards in RDF: the IFLA experience

Gordon Dunsire

Independent Consultant; Chair of IFLA Namespaces Technical Group

Presented at breakout session Requirements Gathering: Best practices for Multilingual Linked Open Data (BP-MLOD),
as part of the W3C Multilingual Web Workshop, Rome, 2013

<input type="text"/>	<input type="button" value="Search Vocabularies"/>
<input type="text"/>	<input type="button" value="Search Element Sets"/>



International Federation of Library Associations and Institutions (IFLA) maintains global standards for the library/bibliographic environment:

- Functional Requirements for Bibliographic Records (FRBR)/ Authority Data (FRAD)/Subject Authority Data (FRSAD)
 - International Standard Bibliographic Description (ISBD)
 - UNIMARC
- ... as RDF element sets and value vocabularies

open metadata registry

Supporting Metadata Interoperability

<input type="text"/>	Search Vocabularies
<input type="text"/>	Search Element Sets

Element Sets: Show detail for ISBD elements

[Detail](#)[Elements](#)[History](#)[Maintainers](#)

Opaque URLs

Label ⓘ	Type	URI ⓘ	Status	Updated ⓘ	Last Updated by ⓘ	Actions
has content form	property	.../ns/isbd/elements/P1001	Published	2011-06-10 13:22	Gordon Dunsire (ifla)	
has content qualification	property	.../ns/isbd/elements/P1002	Published	2011-06-10 13:24	Gordon Dunsire (ifla)	
has media type	property	.../ns/isbd/elements/P1003	Published	2011-06-10 13:25	Gordon Dunsire (ifla)	
has title proper	subproperty	.../ns/isbd/elements/P1004	Published	2011-06-13 12:43	Gordon Dunsire (ifla)	
has parallel title	subproperty	.../ns/isbd/elements/P1005	Published	2011-06-13 12:51	Gordon Dunsire (ifla)	
has other title information	subproperty	.../ns/isbd/elements/P1006	Published	2011-06-13 12:44	Gordon Dunsire (ifla)	
has statement of responsibility				Dunsire (ifla)		
has edition statement				Dunsire (ifla)		
has parallel edition statement				Dunsire (ifla)		
has statement of responsibility				Dunsire (ifla)		
has additional edition statement				Dunsire (ifla)		
has title				Dunsire (ifla)		
has mathematical data				Dunsire (ifla)		
has music format statement				Dunsire (ifla)		
has numbering area				Dunsire (ifla)		
has place of publication				Dunsire (ifla)		
has name of publisher, producer, distributor	property	.../ns/isbd/elements/P1017	Published	2011-06-11 14:10	Gordon Dunsire (ifla)	
has date of publication, production, distribution	property	.../ns/isbd/elements/P1018	Published	2011-06-11 14:14	Gordon Dunsire (ifla)	
has place of printing or manufacture	property	.../ns/isbd/elements/P1019	Published	2011-06-11 14:15	Gordon Dunsire (ifla)	
has name of printer or manufacturer	property	.../ns/isbd/elements/P1020	Published	2011-06-11 14:16	Gordon Dunsire (ifla)	

Element Sets: ISBD elements

Elements: has content form

Detail

Statements

History

Profile property	Show object	Lang	Status	Updated 	Updated by 	Actions
name	hasContentForm	English	Published	7 May 2011 16:50	Gordon Dunsire (ifla)	
label	has content form	English	Published	3 March 2011 14:41	Gordon Dunsire (ifla)	
type	property	English	Published	7 May 2011 16:51	Gordon Dunsire (ifla)	
uri	http://iflastandards.info/n...	English	Published	10 June 2011 13:23	Gordon Dunsire (ifla)	
status	Published	English	Published	10 June 2011 13:22	Gordon Dunsire (ifla)	
description	Relates a resource to a cat...	English	Published	3 March 2011 14:42	Gordon Dunsire (ifla)	
domain	http://iflastandards.info/n...	English	Published	3 March 2011 14:42	Gordon Dunsire (ifla)	
label	tiene forma del contenido	Spanish	Published	3 April 2011 13:45	Gordon Dunsire (ifla)	
description	Relaciona un recurso con un...	Spanish	Published	3 April 2011 13:45	Gordon Dunsire (ifla)	

9 results

Scope
Style
Reference source
Disambiguation

Translations of RDF representations of
IFLA standards

IFLA Cataloguing Section, [ISBD Review Group's ISBD/XML Study Group](#)

Version 1.0, 9 April 2012

Preamble

Background

Partial translation

Local schedule

Detail	Properties	History			
skos:property	Value	Language	Status	Updated	Actions
preferred label	dataset	English	Published	3 March 2011 13:47	
definition	Content expressed by digital... Examples include numeric da...	English	Published	14 March 2011 13:29	
scope note	Examples include numeric da...	English	Published	14 March 2011 13:29	
preferred label	conjunto de datos	Spanish	Published	14 March 2011 13:30	
definition	Contenido expresado por dat...	Spanish	Published	14 March 2011 13:30	
preferred label	Ejemplos son los datos num...	Spanish	Deprecated	25 August 2012 12:04	
preferred label	skup podataka	Croatian	Published	11 May 2011 11:01	
preferred label	dataset	Italian	Published	28 August 2012 8:58	
definition	Un contenuto espresso media...	Italian	Published	28 August 2012 8:58	
scope note	Gli esempi includono i dati...	Italian	Published	28 August 2012 8:57	
preferred label	набор от данни	Bulgarian	New-Proposed	24 August 2012 5:03	
definition	Съдържание, из... Примерите вклю...	Bulgarian	New-Proposed	24 August 2012 16:49	
scope note	Примерите вклю...	Bulgarian	New-Proposed	24 August 2012 16:51	
preferred label	электронные да...	Russian	New-Proposed	25 August 2012 7:00	
definition	Содержание, вы... Примеры включа...	Russian	New-Proposed	25 August 2012 11:49	
scope note	Примеры включа...	Russian	New-Proposed	25 August 2012 11:49	
preferred label	skup podataka	Serbian	New-Proposed	25 August 2012 11:55	
scope note	Takvi primeri su numerički...	Serbian	New-Proposed	25 August 2012 11:59	
definition	Sadržaj izražen digitalno...	Serbian	New-Proposed	25 August 2012 12:02	
scope note	Ejemplos son los datos num...	Spanish	Published	25 August 2012 12:05	

Vocabulary: Show detail for Multilingual dictionary of cataloguing terms and concepts (MulDiCat)

RDF value vocabulary

[Detail](#)[Concepts](#)[History](#)[Versions](#)[Maintainers](#)

Detail

Owner: Gordon Dunsire

Name: Multilingual dictionary of cataloguing terms and concepts (MulDiCat)

URL: <http://www.ifla.org/publications/multilingual-dictionary-of-cataloguing-terms-and-concepts-muldicat>

Note: Intended to be used for authoritative translations of IFLA cataloguing standards and related documents. The terms reflect international agreements on terms to use for these cataloguing and classification concepts

Community:

Status: Published

Language: English

... for authoritative
translations of IFLA
cataloguing standards and
related documents.

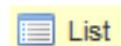
URI

Base Domain: <http://iflastandards.info/ns/>

Token: muldicat

URI: <http://iflastandards.info/ns/muldicat>

Users +

[List](#)[Get RDF](#)[Get XML Schema](#)

Vocabulary: Multilingual dictionary of cataloguing terms and concepts (MulDiCat)

Concepts: Content type

26+ languages

Detail	Properties	History	Language	Status	Updated	Actions
skos:property	Value					
definition	هو تحديد يعكس ش...		Arabic	Published	7 May 2012 4:58	
preferred label	نوع المحتوى		Arabic	Published	2 May 2012 7:53	
definition	Означение за о...		Bulgarian	Published	7 May 2012 4:58	
preferred label	Вид на съдържа...		Bulgarian	Published	2 May 2012 7:54	
definition	Designació que reflecteix ...		Catalan	Published	7 May 2012 4:59	
alternative label	Contingut, Tipus de		Catalan	Published	29 August 2012 5:24	
preferred label	Tipus de contingut		Catalan	Published	2 May 2012 7:54	
definition	Označení vystihující z...		Czech	Published	10 May 2012 3:24	
preferred label	Typ obsahu		Czech	Published	2 May 2012 7:56	
definition	Eine Bezeichnung, die die g...		German	Published	10 May 2012 3:25	
preferred label	Art des Inhalts		German	Published	2 May 2012 7:57	
alternative label	Type of content		English	Published	1 May 2012 12:15	
preferred label	Content type		English	Published	1 May 2012 10:15	
definition	A designation that reflects...		English	Published	1 May 2012 10:16	
definition	Una designación que reflej...		Spanish	Published	10 May 2012 3:27	
preferred label	Tipo de contenido		Spanish	Published	2 May 2012 8:03	
definition	Merkintö, joka kuvastaa si...		Finnish	Published	7 May 2012 4:59	
alternative label	Sisällön tyyppi		Finnish	Published	29 August 2012 5:27	
preferred label	Sisältötyyppi		Finnish	Published	2 May 2012 7:57	
preferred label	Contenu (Type de)		French	Published	28 August 2012 8:40	

End

- ❖ gordon@gordondunsire.com
- ❖ <http://iflastandards.info/ns/fr/>
- ❖ <http://iflastandards.info/ns/isbd/>
- ❖ MulDiCat
 - ❖ <http://metadataregistry.org/vocabulary/show/id/299.html>
- ❖ UNIMARC
 - ❖ Real soon now

Naming and Labeling in the Multilingual Web of Data

Daniel Vila-Suero

Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo sn, 28660 Boadilla del Monte, Madrid

<http://www.oeg-upm.net>

dvila@fi.upm.es

Acknowledgements: BabelData Project (TIN2010-17550), Elena Montiel-Ponsoda, Elena Escolano, Boris Villazón-Terrazas, Gordon Dunsire, Asunción Gómez-Pérez, Jorge Gracia

W3C Multilingual Web workshop:
Making the multilingual web work
Rome, 13.03.2013

- Based on

"Style guidelines for naming and labeling ontologies in the multilingual Web" Montiel-Ponsoda, Vila-Suero, Villazón-Terrazas, Dunsire, Escolano and Gómez-Pérez. DC Conference 2011

- + some practical examples/issues from:
 - <http://datos.bne.es>, and
 - IFLA vocabularies translation into Spanish

NAMING

Some general URI design guidelines

Guideline	Source
Don't put too much semantics in the URI	(Berners-Lee, 1998)
Don't bind URIs to some classification or topic (you may change your point of view)	(Berners-Lee, 1998)
Use short URLs as much as possible	(Théreaux, 2003) *
Choose a case policy (all lowercase or first letter uppercase)	(Théreaux, 2003)
Use lower case	(Cabinet Office, 2009)
Separate words by hyphens	(Cabinet Office, 2009)
Where the reference local name may change, or becomes overly verbose, a code may be more appropriate	(Cabinet Office, 2009)

Naming: Preliminary guidelines for a multilingual scenario

Guidelines for the Multilingual Web	Example
Use preferably opaque URI local names	http://aims.fao.org/aos/agrovoc/c_330988
When using meaningful URI local names, use your own language with diacritics, etc.	http://geo.linkeddata.es/ontology/rio
When using meaningful URI local names, use dash or CamelCase for word delimiter, depending on what reads more easily in your own language.	http://geo.linkeddata.es/ontology/aguas-corrientes

Some tools are not prepared for opaque URIs (Pubby)...

* <http://datos.bne.es/resource/XX1718747>

ifla-frbr:P3039	▪ Cervantes Saavedra, Miguel de
ifla-frbr:P3040	▪ 1547-1616
ifla-frad:P4031	▪ Cervantes ▪ Cervantes di Saavedra, Michele ▪ Cervantes, Miguel de ▪ Servantes Saavedra, Migel ▪ Sîrfantis Saafedrâ, Mîgîl dî ▪ Therbantes, Minkel nte ▪ Zerbantes eta Saabedra, Mikel ▪ Θερβάντες, Μιγκέλ ντε ▪ Сервантес Сааведра, Мигель де ▪ سرفانتس ساافدرا، میگیل دی

Semantic Web Journal reviewer about datos.bne.es' paper* :

"It is pity that local names of chosen IFLA-FRBR properties are cryptic codes ... but **authors of this paper are not to blame about that**"

* <http://www.semantic-web-journal.net/content/datosbnees-library-linked-data-dataset>

Some others are better prepared (Puelia)...

* <http://datos.bne.es/frontend/persons>

Display labels are configurable using a Turtle config file

tiene fechas de persona	1547-1616	
tiene nombre de persona	Cervantes Saavedra, Miguel de	
tipo	Person / Persona	

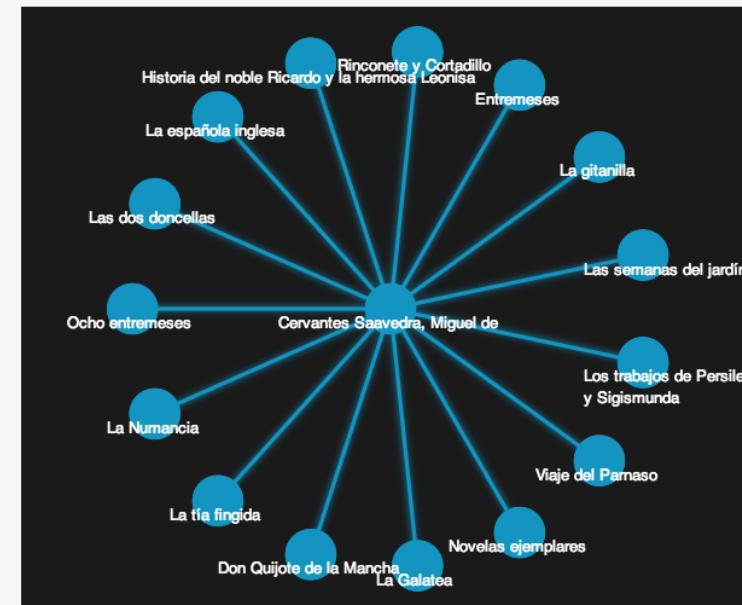
Label not selected based on User's locale

frbr:C1005 a rdfs:Class;
rdfs:label "Person"@en, "Persona"@es

Some personal experiences using opaque URIs

The screenshot shows a search interface with the following elements:

- Logos for Ontology Engineering Group and BNE.
- Navigation links: autores, obras, entidades, themes.
- Search bar: Cervantes Saavedra, Miguel de.
- Result summary: Cervantes Saavedra, Miguel de
- List of titles (titles):
 - [Visita a las galerías y Nómadas]
 - Don Quijote de la Mancha
 - El celoso extremo
 - Novela
 - Leben und Taten des s... Quixote de la Mancha
 - Las mejores páginas de estudios y comentarios seguidas de un vocabulario
 - La española inglesa
 - La cueva de Salamanca
- Portrait of Miguel de Cervantes.
- Section: Atributos (Attributes), listing various name variants and dates.



- How many times you write an URI when developing an application?
 - e.g. var workURI = "http://ifla.ns..../C1001"
- For issuing queries to open SPARQL endpoints opaque URIs are painful.
- **isbd:hasStatementOfResponsibilityRelatingToEdition**
Or
isbd:P1010?

LABELING

Labeling: Initial guidelines for the Multilingual Web

- How to represent labels?

rdfs:label, SKOS, SKOS-XL, Lemon?

- How to structure the content?

Guidelines for the Multilingual Web	Example
Use the singular form for nouns	
Use space as word delimiters, or follow your own language conventions	skos:prefLabel "financial asset"@en
Use upper or lower case according to your own language conventions	rdfs:label "autor"@es rdfs:label "Miguel de Cervantes"@es
Add as many labels as needed for classes and properties	ex:FAOlabel1 rdf:type skosxl:Label; skosxl:literalForm "Food and Agriculture Organization"@en. ex:FAOlabel2 rdf:type skosxl:Label; skosxl:literalForm "FAO"@en

Example: ISBD Cartographic

Preferred Label:	cartographic
Language:	English
URI:	http://iflastandards.info/ns/isbd/terms/contentqualification/type/T1001 (RDF)
Top Concept?:	✓
Status:	Published

Properties

scope note	Uključuje geografske karte, atlase, globuse, reljefne modele i tako dalje.	Serbian	New-Proposed
scope note	Includes maps, atlases, globes, relief models, etc.	English	Published
scope note	Include le carte geografiche, gli atlanti, i globi, i modelli in rilievo, etc.	Italian	Published
scope note	包括地图、地图册、球仪、地形模型等。	Chinese	New-Proposed
scope note	Incluye mapas, atlas, globos, modelos en relieve, etc.	Spanish	Published
scope note	Включва карти, атласи, глобуси, модели на релеф и др.	Bulgarian	New-Proposed
scope note	Ietver kartes, atlantus, globusus, reljefa modejus, u.c.	Latvian	New-Proposed
scope note	Включает в себя карты, атласы, глобусы, модели рельефа и т.п.	Russian	New-Proposed

alternative label	cartográfico	Spanish	New-Proposed
alternative label	cartográfica	Spanish	New-Proposed
alternative label	kartografski	Croatian	New-Proposed
alternative label	kartografska	Croatian	New-Proposed

```
isbd:T1001 lemon:isReferenceOf [ lemon:isSenseOf :cartographic] .  
:cartographic lemon:LexicalEntry ;  
    lemon:form [ lemon:writtenRep "cartográfico"@es ;  
                isocat:grammaticalGender isocat:mASCULINE ] ;  
    lemon:form [ lemon:writtenRep "cartográfica"@es ;  
                isocat:grammaticalGender isocat:fEMININE ] .  
  
:isocat:grammaticalGender rdfs:subPropertyOf lemon:property .
```

Some questions arise:

- What pattern should IFLA use to name/manage these new URIs (:cartographic) ?
- Are we making it more difficult for data consumers (more HTTP requests, etc.)?

THANK YOU!!

Slides are available at **slideshare.net/DanielVilaSuero**

email: **dvila@fi.upm.es**

Twitter: **@dvilasuero**



XLIFF Workflow and Multilingual Provenance in Linked Data

David Lewis
CNGL at Trinity College Dublin
Co-chair W3C MLW-LT WG



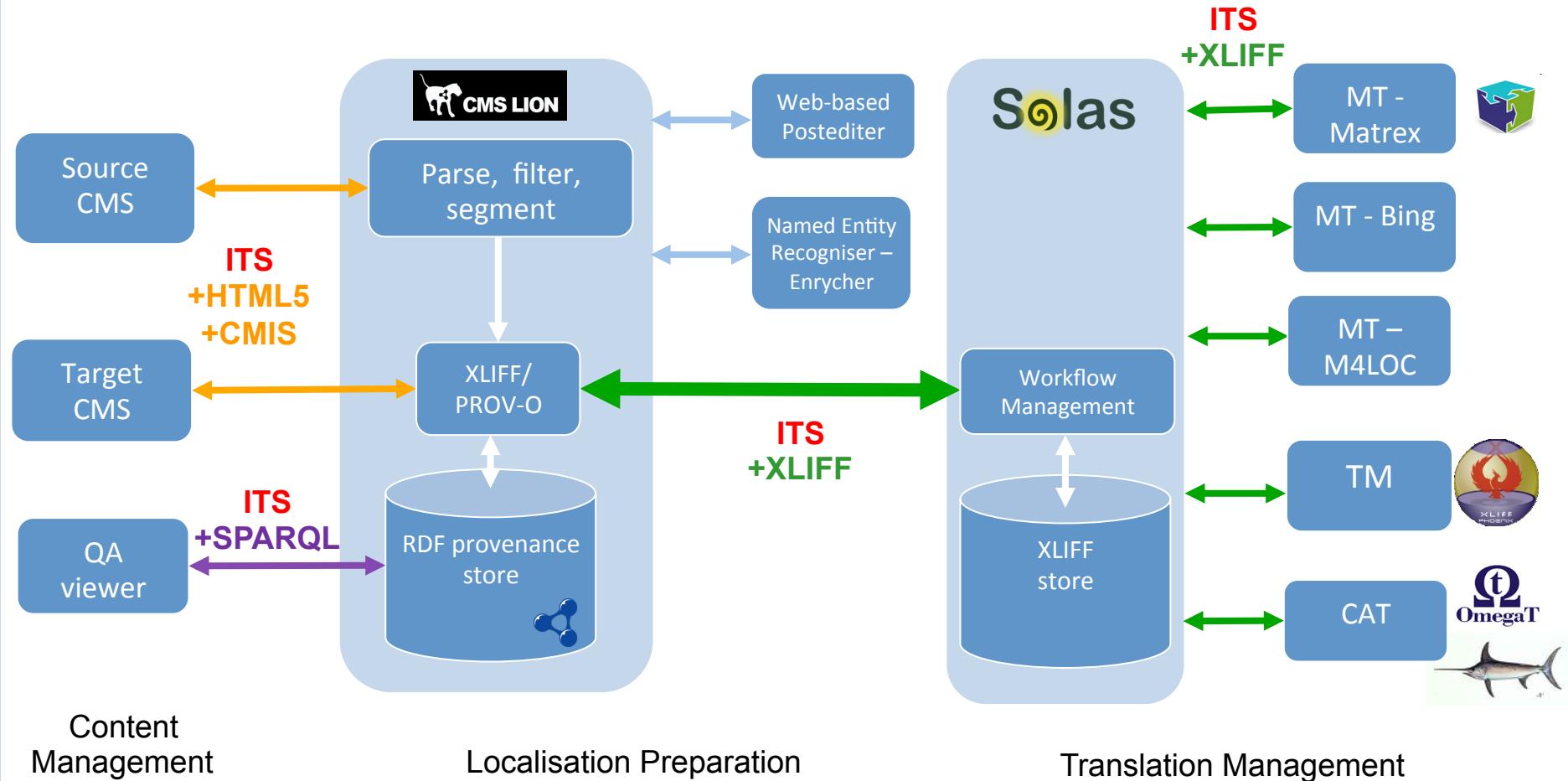
The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

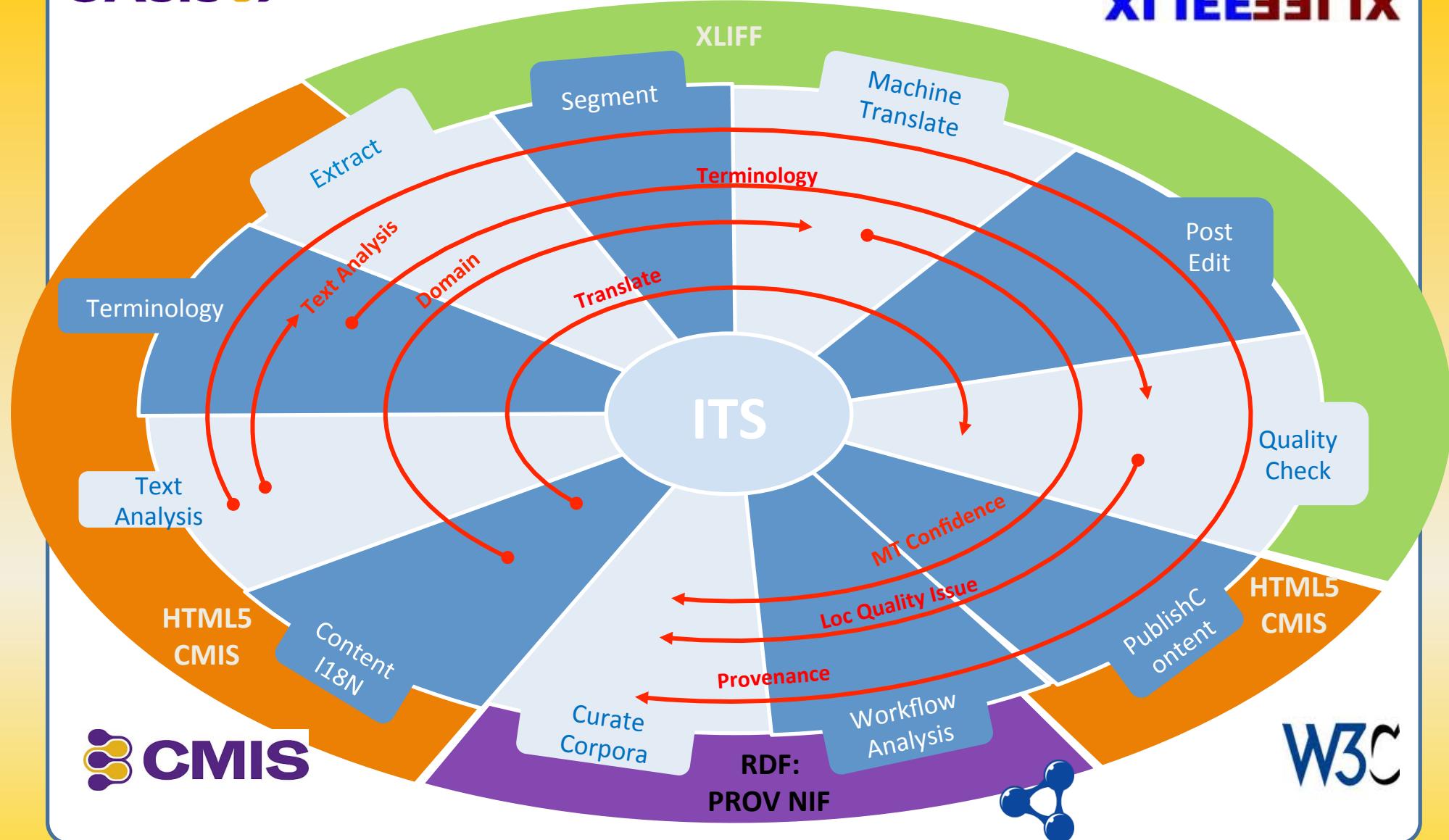


Curation of Language Resources

- Language Resource Curation
 - Relies on specific funded actions
 - Episodic funding, sustainability and coverage concerns
- LOD promises:
 - Improved sharing and annotation of existing resources
 - Some opportunistic curation, e.g. mining Dbpedia
- Can there be better synergy with commercial language services?
 - Use case: Localisation workflows

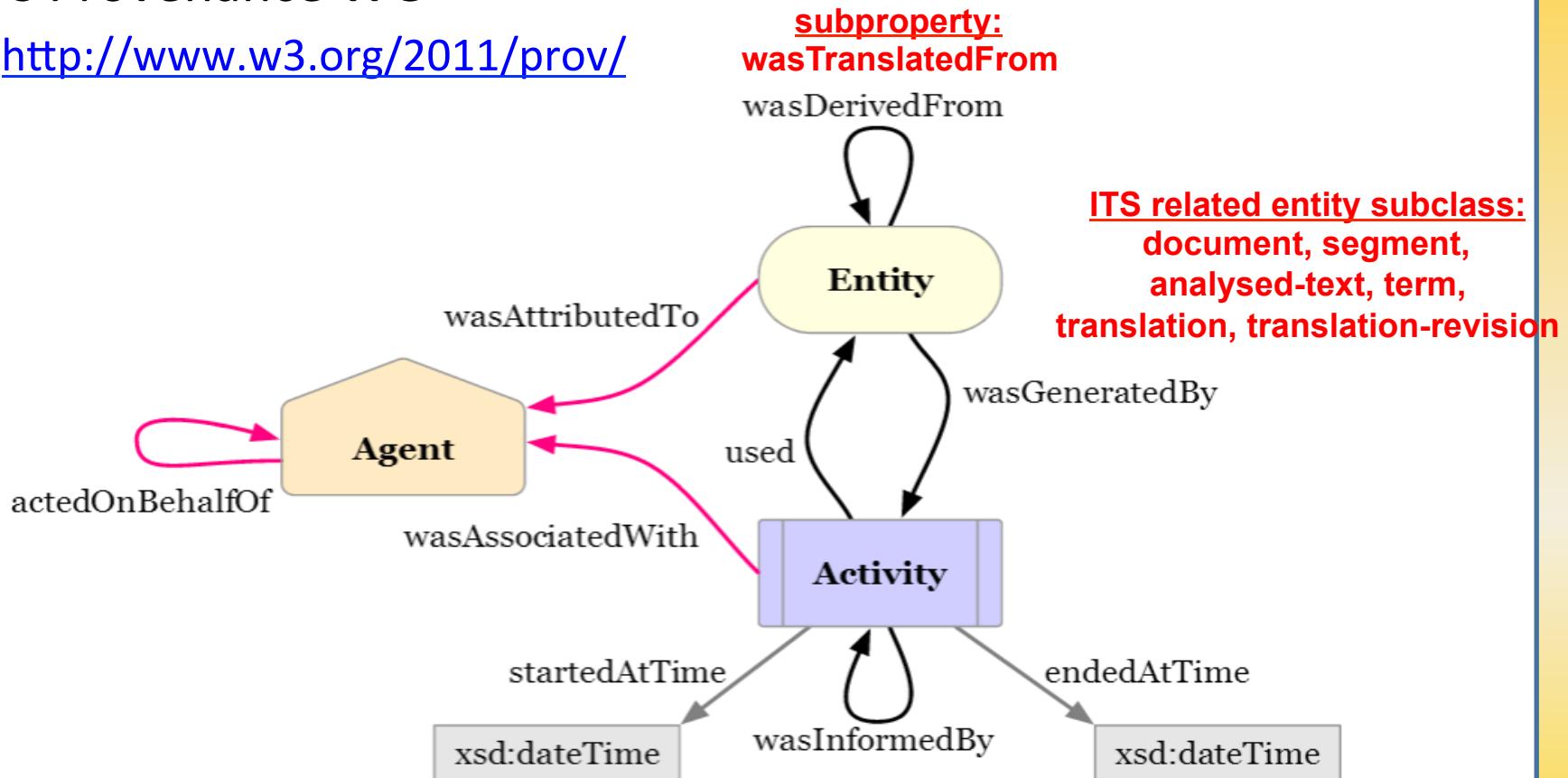
System Architecture





Provenance Data: RDF-based logging

- W3C Provenance WG
 - <http://www.w3.org/2011/prov/>



From: <http://www.w3.org/TR/prov-primer/>

The MultilingualWeb-LT Working Group receives funding by the European Commission (project name LT-Web) through the Seventh Framework Programme (FP7) in the area of Language Technologies. Grant Agreement No. 287815.

Active Curation of L3Data

- Generate Linked Language and Localisation Data ('L3Data')
- Pooling to boost EU's many SME language service providers
- Encourage commercial users/generators/curators of L3Data
- EC and other public bodies could bootstrap open data
- Current Gaps:
 - Standards based instrumentation and analytics – ITS
 - Attribution/license annotation and access control
 - Open source tool chain - linked data generating CAT/TMS/CMS
 - More academic-industrial collaboration on use cases
- Follow ITS-XLIFF-PROV-O Use Case at:
 - [http://www.w3.org/International/multilingualweb/lit/wiki/Provenance Best Practice](http://www.w3.org/International/multilingualweb/lit/wiki/Provenance_Best_Practice)

Questions?

BabelNet: a Multilingual Encyclopedic Dictionary as LOD

Roberto Navigli

DIPARTIMENTO
DI INFORMATICA

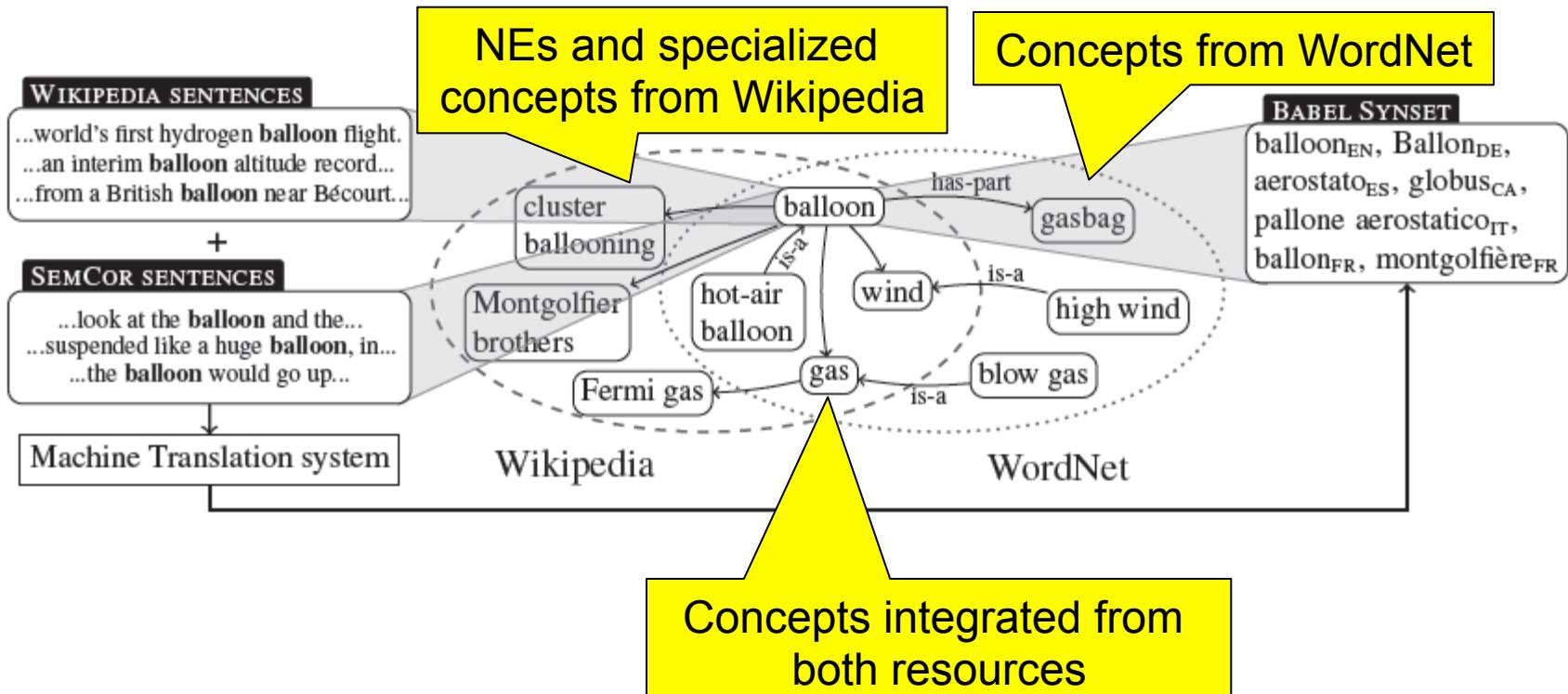


SAPIENZA
UNIVERSITÀ DI ROMA

Linguistic Computing Laboratory
<http://lcl.uniroma1.it>

BabelNet

- A wide-coverage multilingual semantic network including both encyclopedic (from Wikipedia) and lexicographic (from WordNet) entries



Roberto Navigli and Simone Ponzetto: BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. **Artificial Intelligence**, 193, Elsevier, 2012.

BabelNet^{1.1.1}

A very large multilingual ontology

means: more accurate mappings and translations + Wikipedia categories + images + ...

A very large multilingual ontology with 5.5 million concepts • A wide-coverage "encyclopedic dictionary" • Obtained from the automatic integration of WordNet and Wikipedia • Enriched with automatic translations of its concepts • Connected to the Linguistic Linked Open Data cloud!



BabelNet is an output of the MultiJEDI ERC Starting Grant No. 259234. Concept and application by Roberto Navigli.

BabelNet and its API are licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.

For any commercial use, please contact us.



<http://babelnet.org>

BabelNet 1.1.1

A very large multilingual ontology

[SEARCH](#) · [EXPLORE](#) · [PUBLICATIONS](#) · [DOWNLOAD](#)

Type a term: English

(examples: plane, apple, star, Italian, bus driver, calcio, drive#n, bus#n#en, horse#en, mela#it)

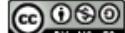
Noun

Title:	Multilingual Web	•	ID: bn:02289651n	•	Type: Named Entity	 explore
Senses:	  Multilingual Web					
Glosses:	  The Multilingual Web is an EU funded project whose primary goal is to provide: "Multilingual Web content management: standards and best practices".					
Categories:	  European Union and science and technology					
DBpedia:	 Multilingual Web					



BabelNet is an output of the MultiJEDI ERC Starting Grant No. 259234. Concept and application by Roberto Navigli.

BabelNet and its API are licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.

 For any commercial use, please contact us.



<http://babelnet.org>

(examples: plane, apple, star, Italian, bus driver, calcio, drive#n, bus#n#en, horse#en, mela#it)

Noun

Title: web ¹	• ID: bn:00080771n	• Type: Concept	 explore
Senses:	  web ¹		
Glosses:	  an intricate network suggesting something that was formed by weaving or interweaving; "the trees cast a delicate web of shadows over the lawn"		
Title: web ²	• ID: bn:00030977n	• Type: Concept	 explore
Senses:	  web ² , entanglement ¹   embolic,  enchevêtement,  verstrickung,  aggrovigliamento,  enredo		
Glosses:	  an intricate trap that entangles or ensnares its victim		
Title: vane ⁴	• ID: bn:00079552n	• Type: Concept	 explore
Senses:	  vane ⁴ , web ³		
Glosses:	  the flattened weblike part of a feather consisting of a series of barbs on either side of the shaft		
Title: network ¹	• ID: bn:00057379n	• Type: Concept	 explore
Senses:	  network ¹ , web ⁴   xarxa,  réseau,  netzwerk,  rete,  red		
Glosses:	  an interconnected system of things or people; "he owned a network of shops"; "retirement meant dropping out of a whole network of people who had been part of my life"; "tangled in a web of cloth"		
Title: World Wide Web ¹	• ID: bn:00080772n	• Type: Concept	 explore
Senses:	  World Wide Web ¹ , WWW ¹ , web ⁵   World Wide Web,  World Wide Web,  World Wide Web,  World Wide Web,  World Wide Web,  World Wide Web   Webs, WWW, Web, Teranyina mundial, W3,  The Web, World-Wide Web,		

Anatomy of BabelNet

5.5 million Babel synsets if we consider synsets with no English lexicalization

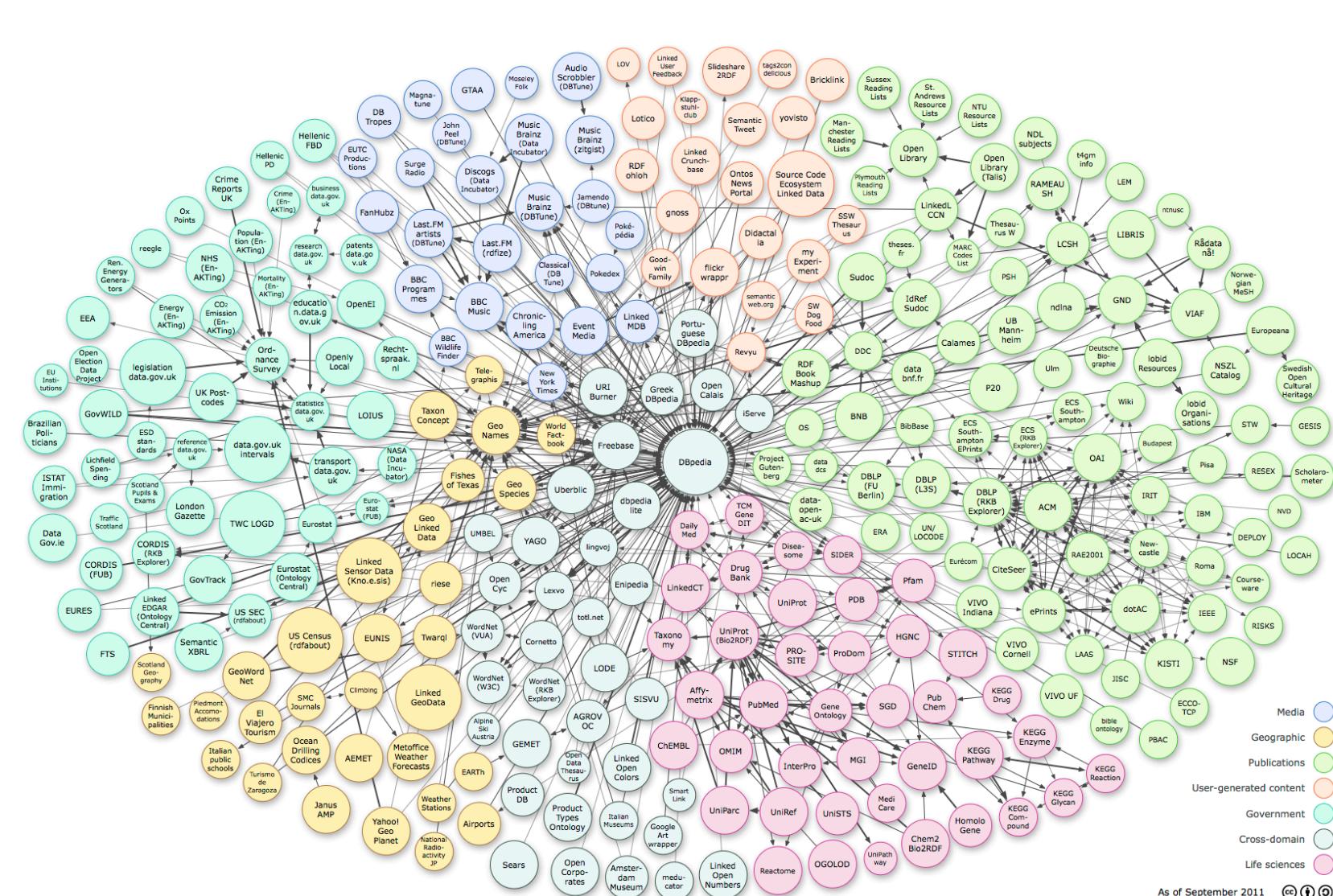
- 6 languages covered (moving to 40+)
- More than 3 million Babel synsets (i.e. concepts and NE)
- More than 26 million word senses:

	English	Catalan	French	German	Italian	Spanish	Total
English WordNet	206,978	-	-	-	-	-	206,978
Wikipedia	pages	2,955,552	123,101	524,897	506,892	404,153	4,863,970
	redirections	3,388,049	105,147	617,379	456,977	217,963	5,189,524
WordNet	translations	-	3,445,470	2,844,751	2,841,916	3,046,325	15,261,889
	monosemous	-	97,876	98,081	97,672	98,475	98,092
SemCor		-	6,852	6,855	6,850	6,856	34,268
Total	6,550,579	3,778,446	4,091,963	3,910,307	3,773,772	3,941,758	26,046,825

- About 70 million lexico-semantic relations:

	English	Catalan	French	German	Italian	Spanish	Total
WordNet	364,552	-	-	-	-	-	364,552
WordNet glosses	617,785	-	-	-	-	-	617,785
Wikipedia	50,104,898	971,379	5,594,590	5,931,099	3,598,733	3,397,754	69,598,453
Total	51,087,235	971,379	5,594,590	5,931,099	3,598,733	3,397,754	70,580,790

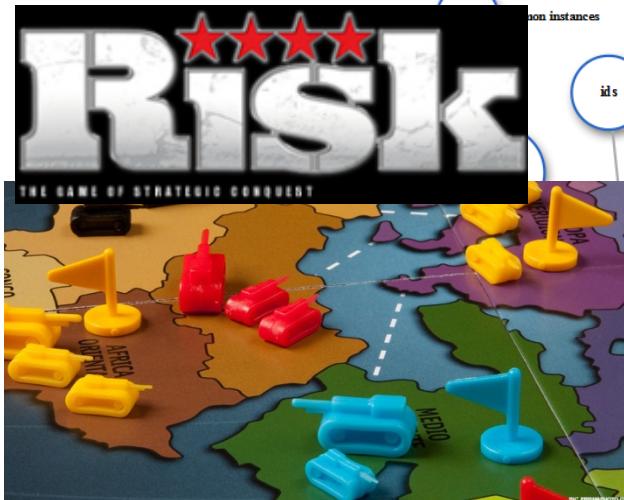
The Linked Open Data cloud...



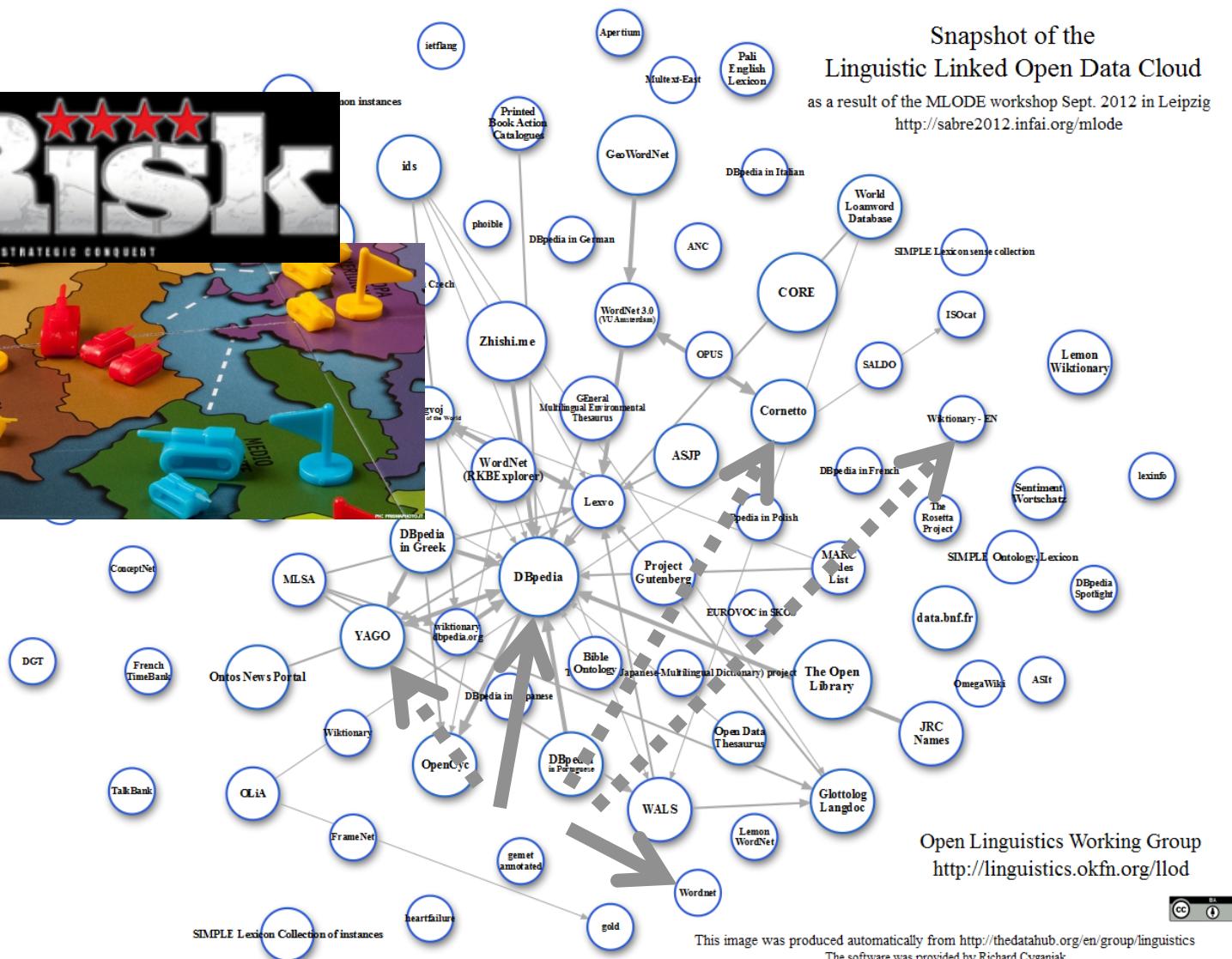
As of September 2011



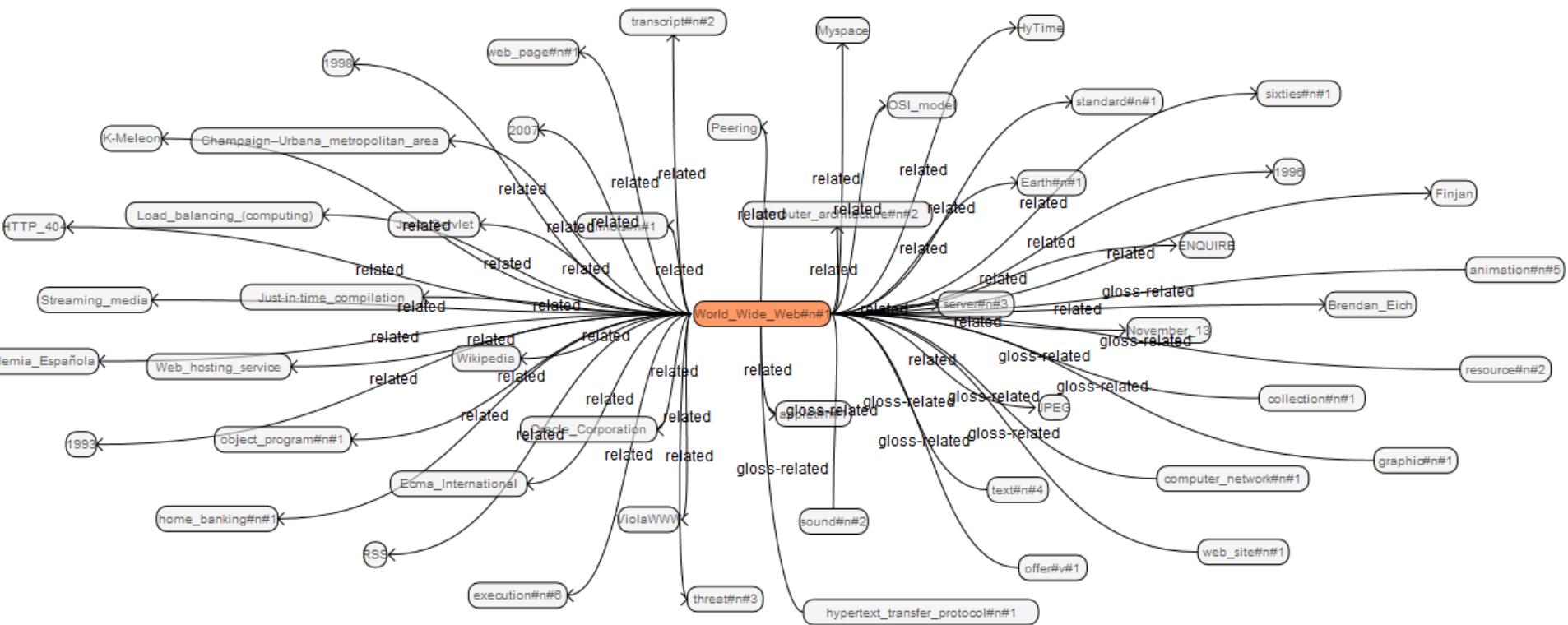
The *Linguistic* Linked Open Data cloud...



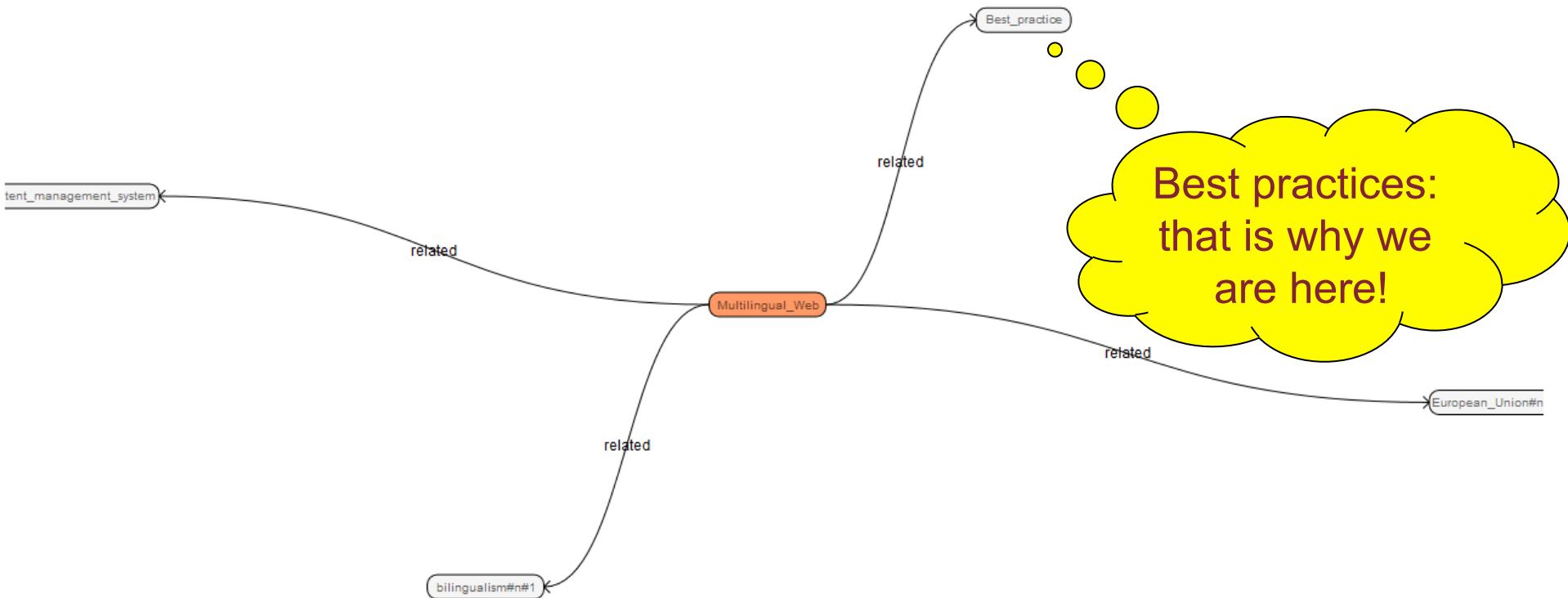
Snapshot of the
Linguistic Linked Open Data Cloud
as a result of the MLODE workshop Sept. 2012 in Leipzig
<http://sabre2012.infai.org/mlode>



An excerpt of the BabelNet graph centered on WWW



And... centered on the Multilingual Web!



The contribution of BabelNet to the LOD

- Going to be available in  RDF
- A real example of large **multilingual LOD** which could benefit from our discussion today
- Providing **different lexicalizations** for the same concepts in different languages
- Leveraging BabelNet for non-English LOD
- Encyclopedic and lexicographic information available in many languages used for **linking** LD multilingually?



SAPIENZA
UNIVERSITÀ DI ROMA

Roberto Navigli
Linguistic Computing Laboratory
<http://lcl.uniroma1.it>

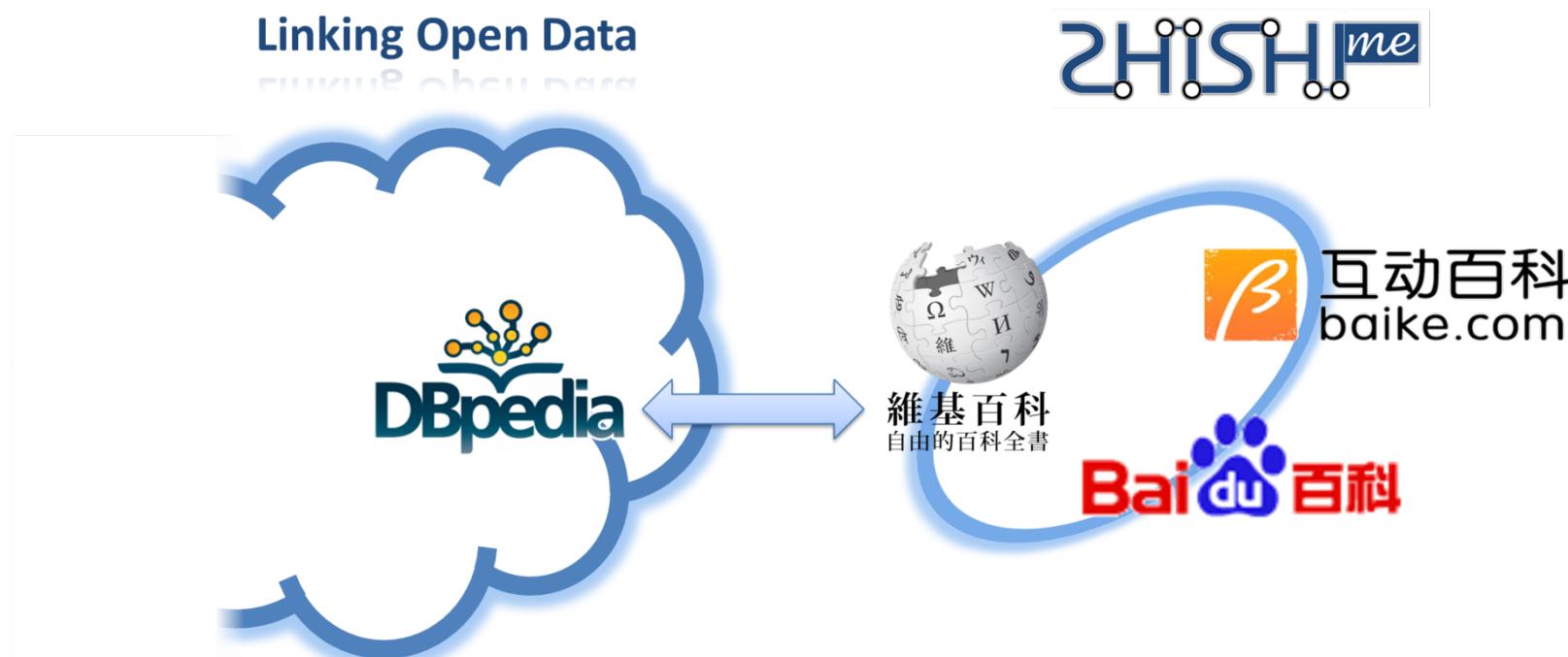


The state of the art of Chinese LOD development

Haofen Wang

Overview of Zhishi.me

- Zhishi.me (<http://zhishi.me>) is the first effort to publish large scale Chinese semantic data and link them together as a Chinese Linking Open Data (CLOD). The statistics are collected by Feb, 2013
 - Over 8 million distinct instances
 - Over 1 billion RDF triples



Multilingual Issues

- There are no one-size-fit-all mechanisms for providing resource identifiers.
- Using IRIs?
 - Chinese characters are non-ASCII, un-encoded words are reader-friendly.
 - XML specification: encoded URIs are not allowed to act as XML properties.
- Using URIs?
 - IRIs are incompatible with HTML 4, non-ASCII characters should be encoded with the URI escaping mechanism to generate legal URIs as “href” values.
 - Most Web browsers automatically encode IRIs which users entered into the address bar. Servers always receive URIs.
- A compromise
 - Storing IRIs in databases and transforming received URIs into IRIs. Trying our best to provide IRI embedded resource files when response users’ data request.

Future Work

- Assigning uniform resource identifiers for matched instances.
 - Naming resources properly
 - Uniting traditional Chinese names and simplified Chinese names
- Developing multilingual instance matching algorithms to discover more links between Zhishi.me and LOD.
- Integrating the e-commerce Web sites (360Buy and Taobao) and social Web sites (e.g., Weibo, Dianping) as the first effort of Chinese Linked Open Stream Data
- Extract ontologies from CLOD and linked with well-known thesaurus and taxonomies like schema.org
- Providing more APIs (like entity linking, complex relation finding and allowing users to upload their own data and return the links with CLOD for federated querying purpose)



Thanks!



Multilingual Linked Open Data Patterns

Jose Emilio Labra Gayo

University of Oviedo, Spain

Joint work with:

Dimitris Kontokostas

Sören Auer

Universität Leipzig, Germany

More info: <http://www.weso.es/MLODPatterns>



MLOD Patterns

A catalog of 20 MLOD patterns

Based on DBPedia I18n experience

Clasified by activity:

Naming

Dereference

Labeling

Longer descriptions

Linking

Reuse

More info: <http://www.weso.es/MLODPatterns>



MLOD Patterns

For each pattern:

Description

Context

Example

Discussion

Relationships

More info: <http://www.weso.es/MLODPatterns>

Overview

Goal	Pattern	Description
Naming	Descriptive URIs	Use descriptive URIs with ASCII characters, % encoding extended characters
	Opaque URIs	Use non human-readable URIs
	Full IRIs	Use IRIs with unicode characters
	Internationalized local names	Use Unicode characters only for local names
	Language in URIs	Include language information in the URI
Dereference	Return Language independent data	Return the same triples independently of the language
	Language content negotiation	Return different triples depending on user agent preferences
Labeling	Label everything	Define labels for all the resources
	Multilingual labels	Add language tags to labels
	Labels without language tag	Add labels without language tags in a default language
Longer descriptions	Divide longer descriptions	Replace long descriptions by more resources with labels
	Lexical information	Add lexical information to long descriptions
	Structured literals	Use HTML/XML literals for longer descriptions
Linking	Identity links	Use owl:sameAs and similar predicates
	Soft links	Use predicates with soft semantics
	Linguistic metadata	Add linguistic metadata about the dataset terms
Reuse	Monolingual vocabularies	Attach labels to vocabularies in a single language
	Multilingual vocabularies	Prefer multilingual vocabularies
	Localize existing vocabularies	Translate labels of existing vocabularies
	Create new localized vocabularies	Create custom vocabularies and link to existing ones

End of presentation



More info: <http://www.weso.es/MLODPatterns>