



# Linked Data in Linguistics for NLP and Web Annotation

UNIVERSITÄT LEIPZIG

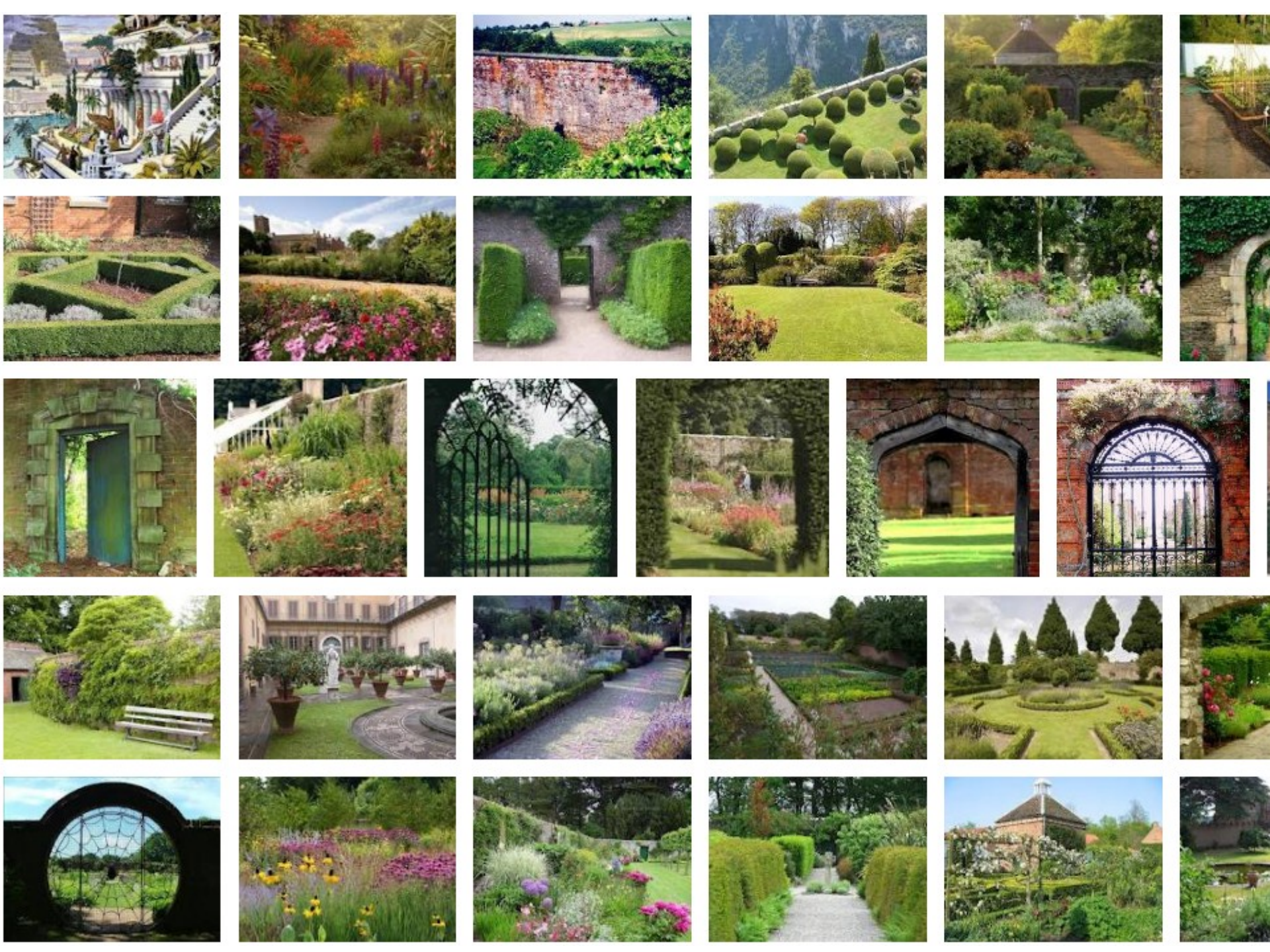
<http://nlp2rdf.org>

<http://lod2.eu>

Sebastian Hellmann

AKSW, Universität Leipzig



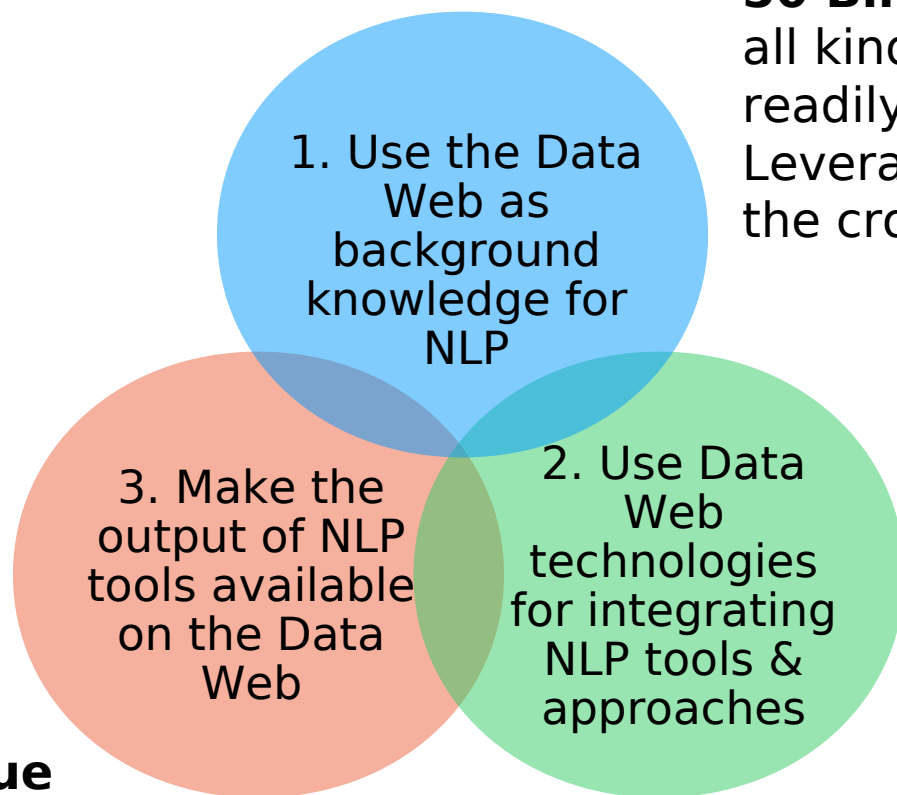






# Turning Walled Gardens into Park Networks of Semantic Linguistic Data

How can we leverage the Data Web for natural language processing?



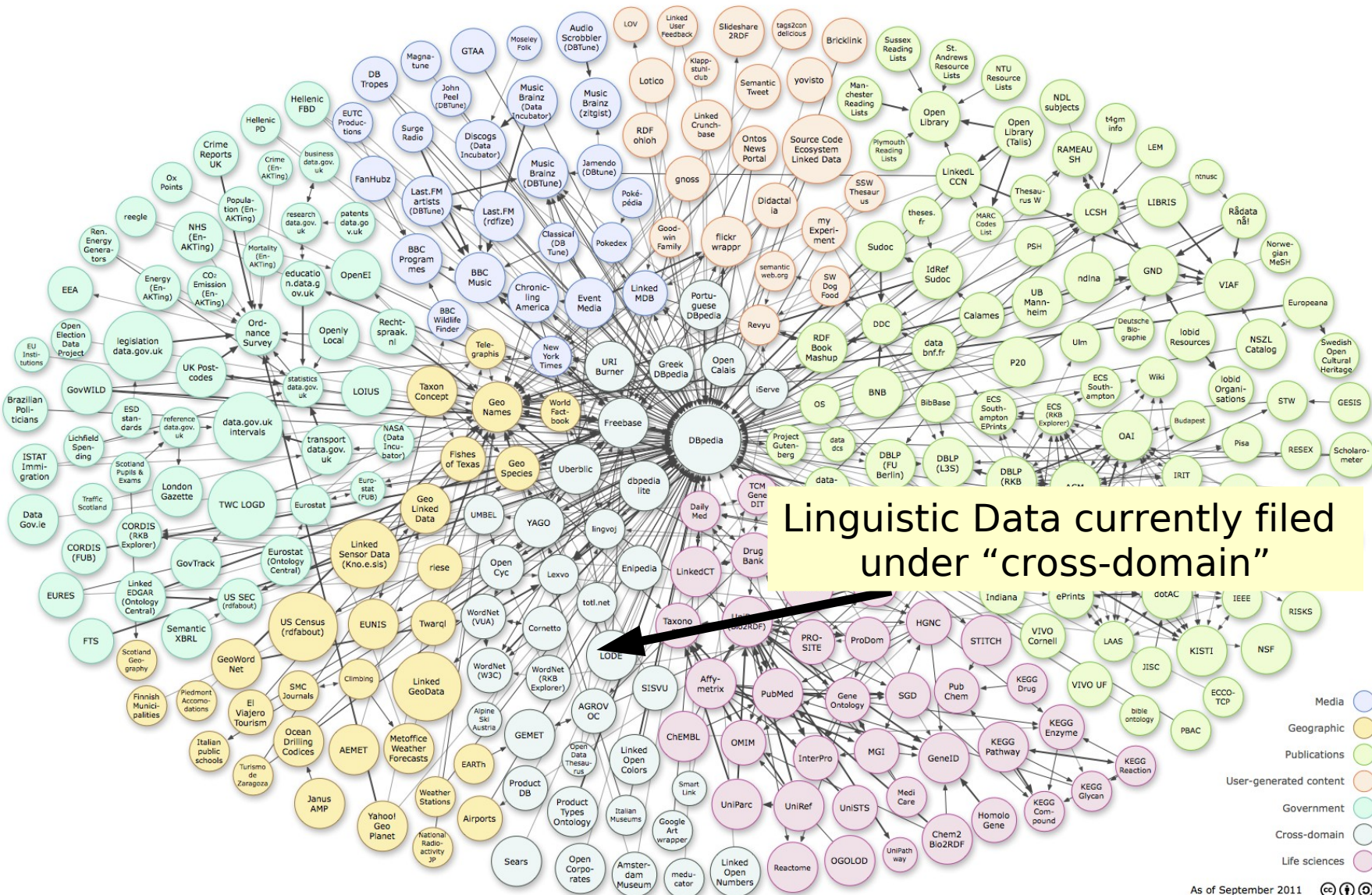
**50 Billion** facts covering all kinds of domains are readily available  
Leverage the wisdom of the crowds

On the Web, by sharing and copying the **value of information increases**

RDF is all about semantic **interoperability**



# 1. Use the Data Web as background knowledge for NLP



Linguistic Data currently filed under "cross-domain"



# 1. Use the Data Web as background knowledge for NLP

Three communities with three resources:

- Working Group for Open Linguistics Data (OWLG)
- DBpedia Internationalization Committee
- Wiktionary2RDF Wrappers

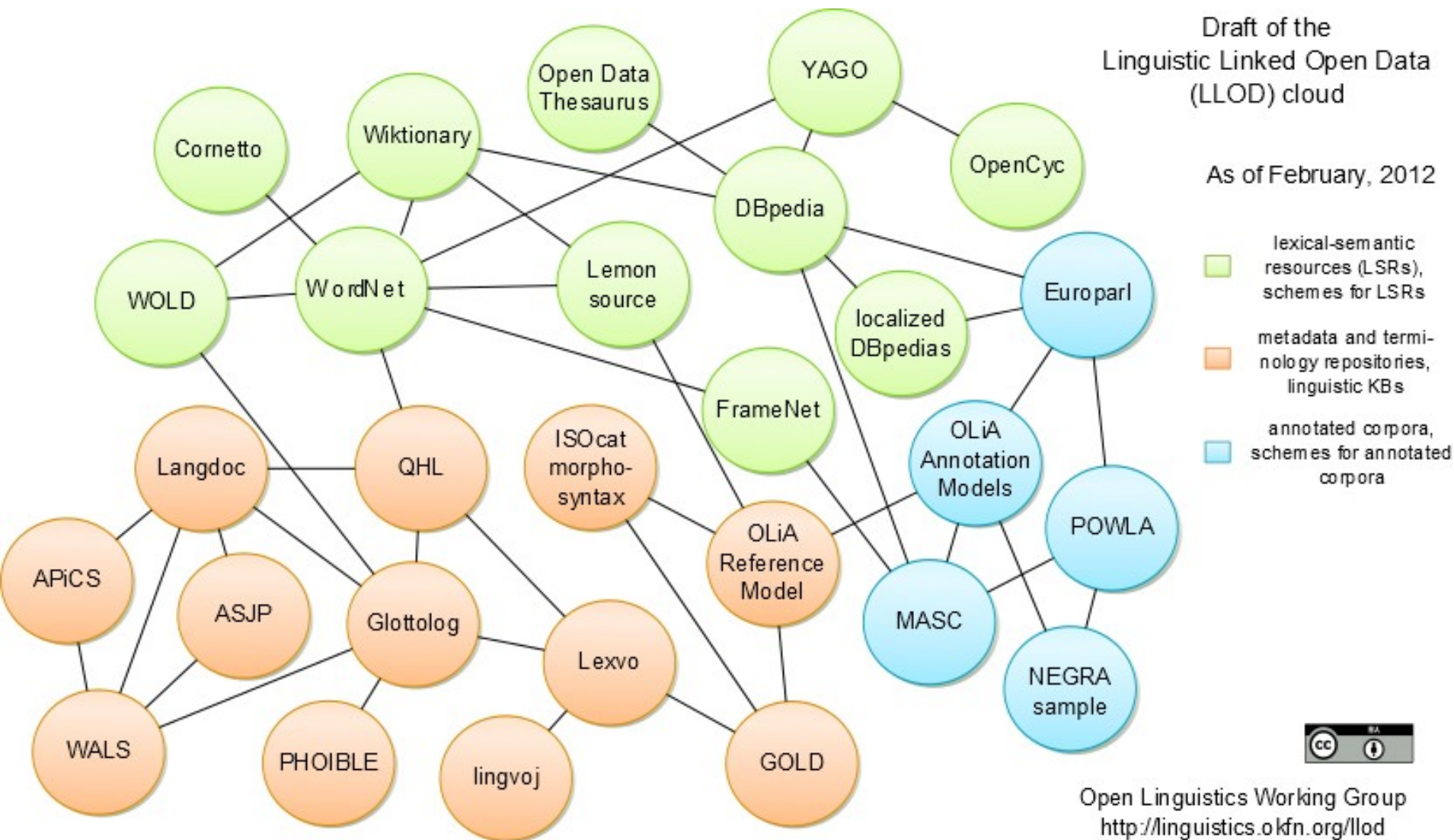




# The Linguistic Linked Open Data Cloud

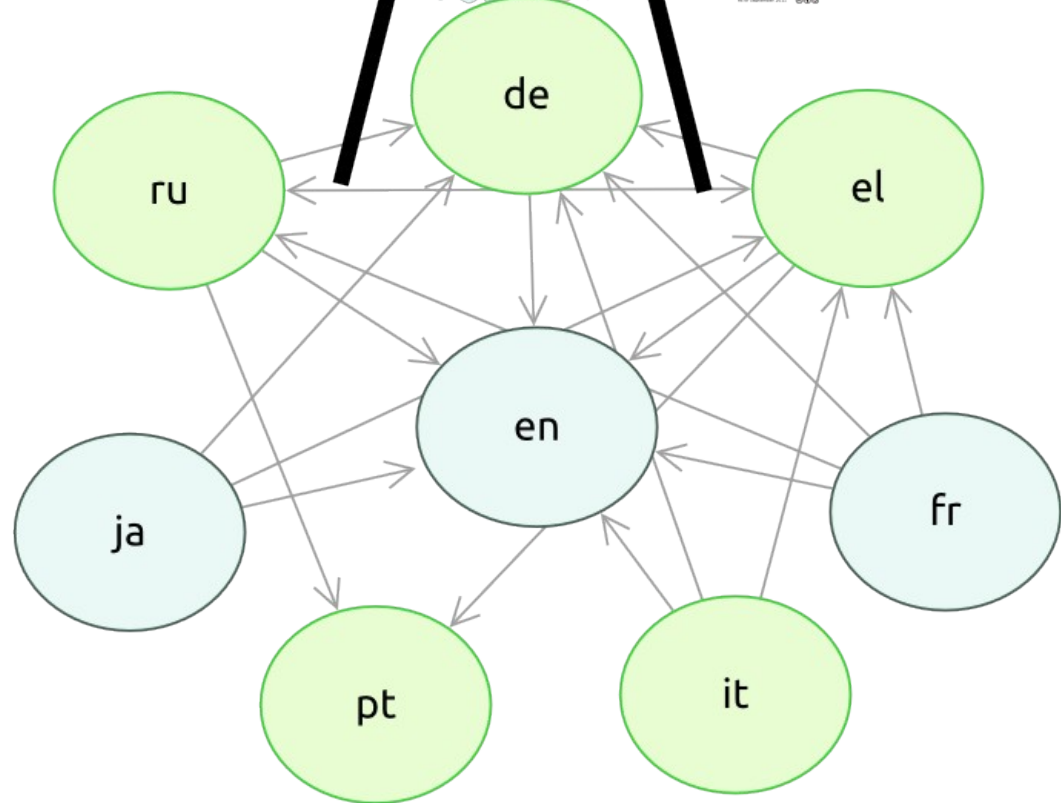
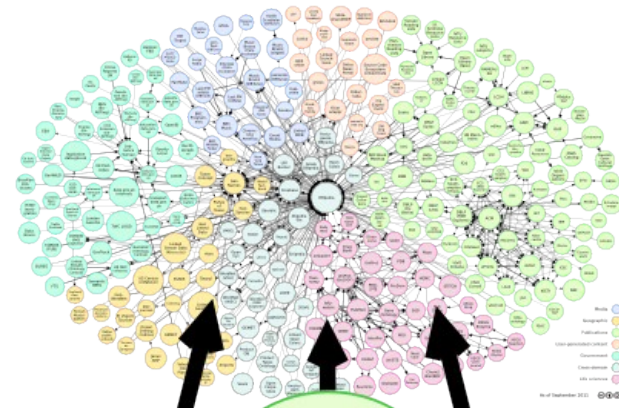
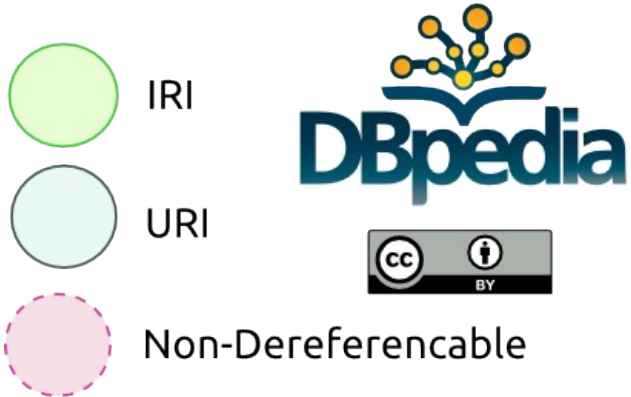
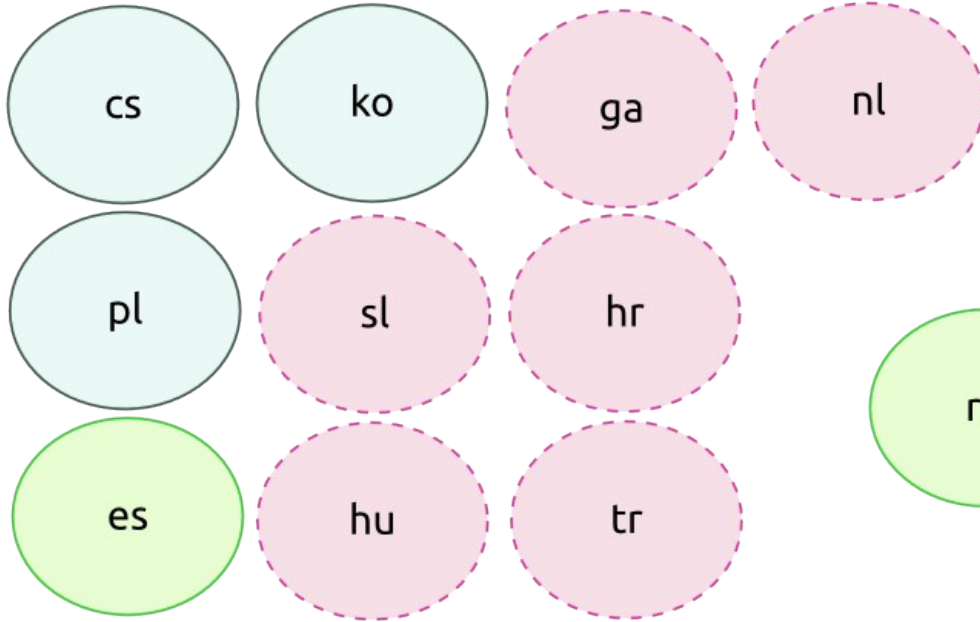
Draft of the Linguistic Linked Open Data (LLOD) cloud

As of February, 2012



# DBpedia Internationalization Committee

## List of DBpedia language Chapters (May 2012)



<http://wiki.dbpedia.org/Internationalization>



# Wiktionary2RDF - Mediator Wrapper

<http://dbpedia.org/Wiktionary>

## Wiktionary

### English

*The free dictionary*

2 197 000+ entries

### Français

*Le dictionnaire libre*

1 915 000 + articles

### 中文

自由的多語言詞典

952 000+ 條詞條

a multilingual tree  
encyclopedia

### Lietuvių

*Laisvasis žodynas*

552 000+ straipsniai

## Wiktionary

**['wɪksjənri]** *n.*,  
a wiki-based Open  
Content dictionary

### Türkçe

*Özgür sözlük*

270 000+ madde

### Malagasy

*Raki-bolana malalaka*

278 000+ teny

### Русский

*Свободный словарь*

267 000+ статей

**Wileo** **['wɪl kəʊl]**

### Tiếng Việt

*Từ điển mở*

229 000+ mục từ

### Polski

*Wolny słownik*

201 000+ stron

### தமிழ்

கட்டற்ற அகரமுதலி

192 000+ சொற்கள்

search • rechercher • 搜尋 • paieška • tadiavo • ara • поиск • tìm kiếm • szukaj • தேடு • serchez • 찾기 • pesquisa • haku • keresés • αναζήτηση • volltext • søk • sök • ricerca • zoeken





# Wiktionary2RDF - Mediator Wrapper

<http://dbpedia.org/Wiktionary>

## Wiktionary

### English

*The free dictionary*  
2 197 000+ entries

### Français

*Le dictionnaire libre*  
1 915 000 + articles

中文  
自由的多語言詞典  
952 000+ 條詞條

a multilingual tree  
encyclopedia

### Lietuvių

*Laisvasis žodynas*  
552 000+ straipsniai

**Wiktionary**  
['wɪksjənri] *n.*,  
a wiki-based Open  
Content dictionary

### Malagasy

*Raki-bolana malalaka*  
278 000+ teny

### Türkçe

*Özgür sözlük*  
270 000+ madde

Русский  
*Свободный словарь*  
267 000+ статей

Wileo ['wɪl kəʊl]

### Tiếng Việt

*Từ điển mở*  
229 000+ mục từ

### Polski

*Wolny słownik*  
201 000+ stron

### தமிழ்

கட்டற்ற அகரமுதலி  
192 000+ சொற்கள்

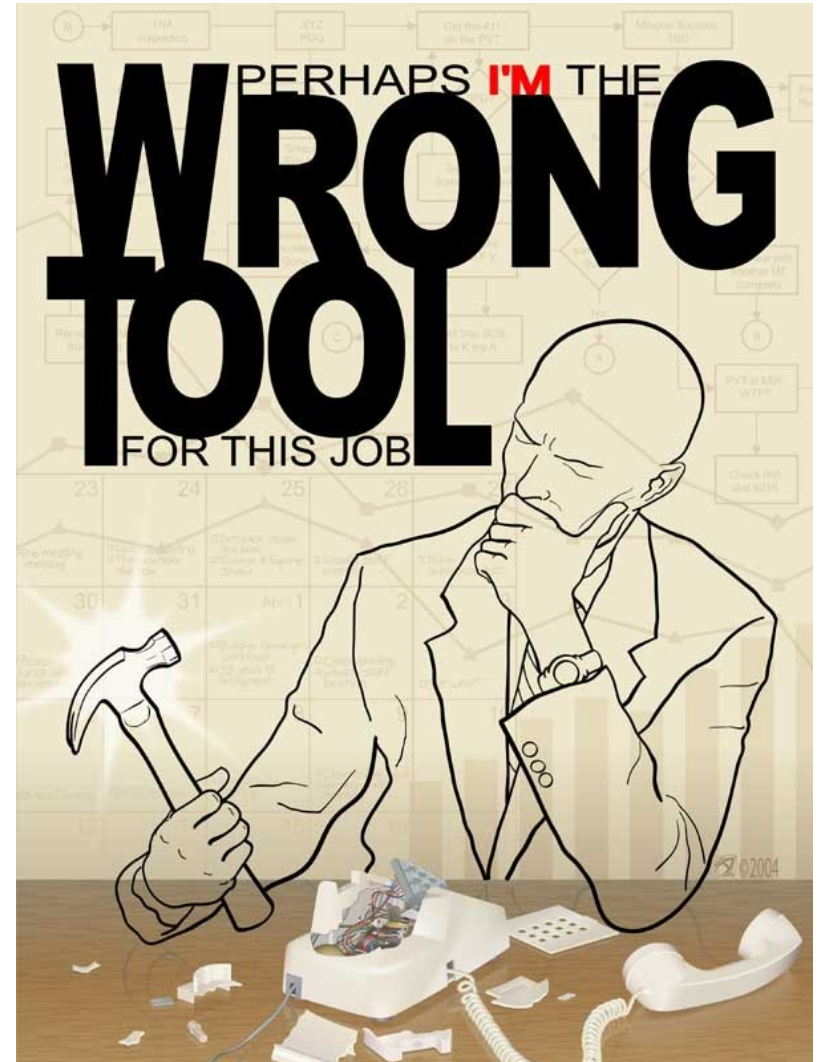




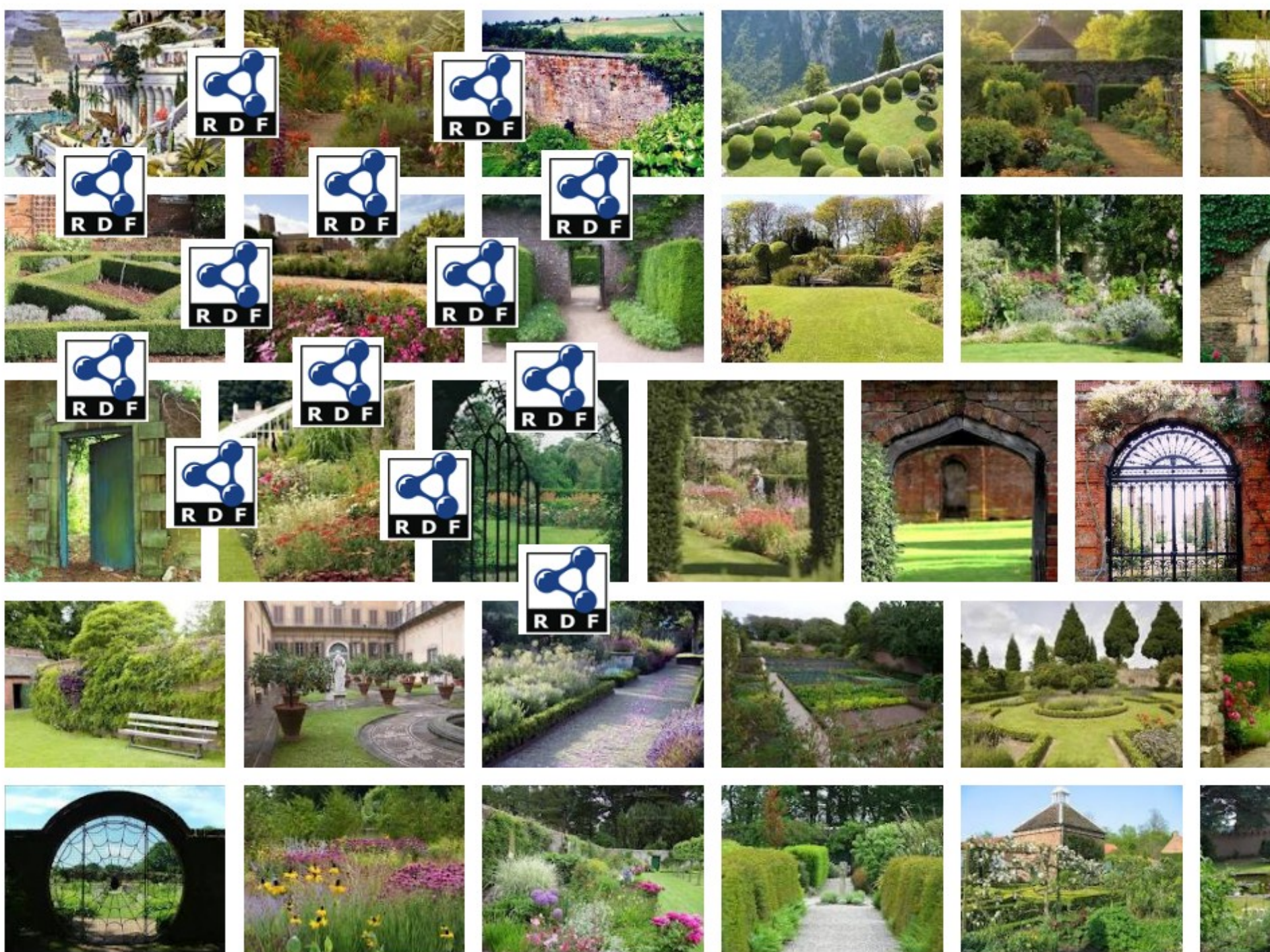
## 2. Use Data Web Technologies for Integrating NLP Tools and Approaches

### Golden Hammer Anti-pattern

The question is not **whether** to use RDF and Linked Data, but **when** to use...











## 2. Use Data Web Technologies for Integrating NLP Tools and Approaches

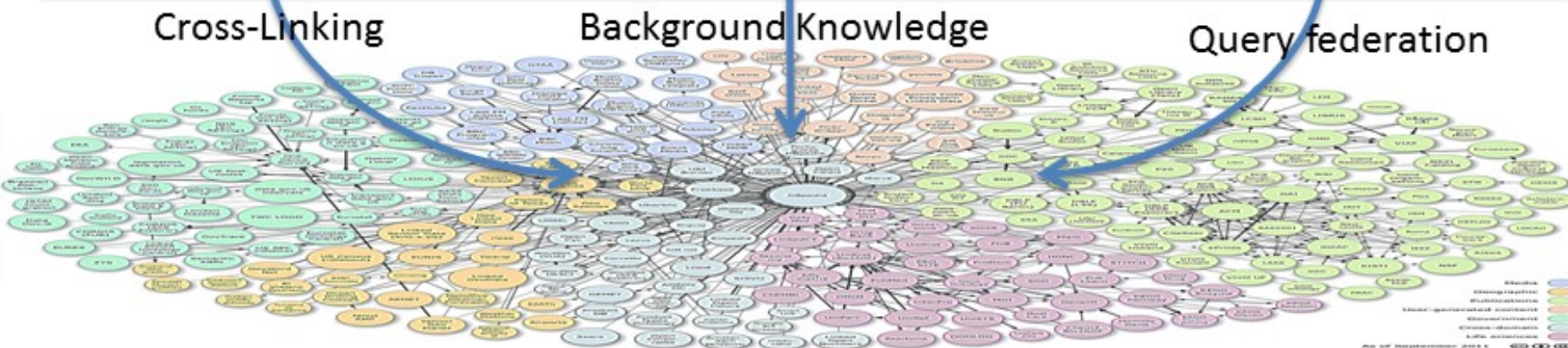
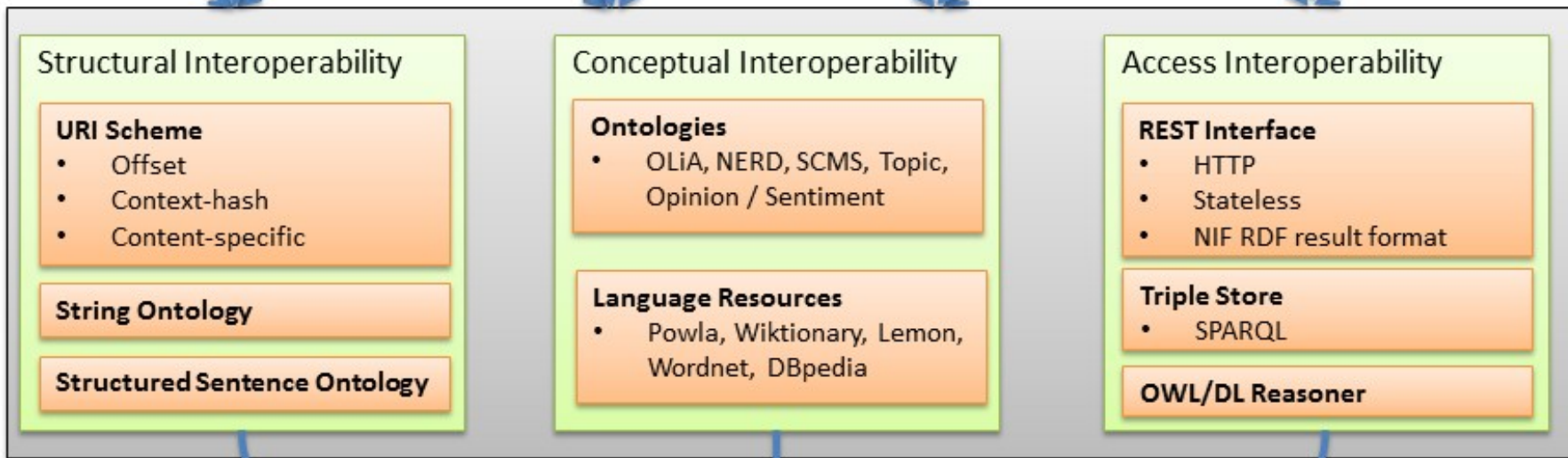
- Ontologies provide (formal) documentation (UML)
- Structure is easy to understand
- Wide range of RDF tools can be used, e.g. LOD2 Stack
- Indexing and querying as Big Picture possible



## 2. Use Data Web Technologies for Integrating NLP Tools and Approaches

The NLP Interchange Format (NIF) is an RDF/OWL-based format that aims to achieve interoperability between Natural Language Processing (NLP) tools, language resources and annotations.

- Road map
  - Bootstrapped by LOD2, but a community project
  - First release in September 2011
  - Great resonance
    - Over 50 people joined the mailing list
    - First third party implementations and contributions
    - Several project discuss usage
  - Currently setting up advisory board, next draft in July







### 3. Make the Output of NLP Tools available on the Web

Currently there is no standard mechanism to transparently combine the WWW, GGG and NLP



## 3. Make the Output of NLP Tools available on the Web

http://www.w3.org/DesignIssues/LinkedData.html



Google



*Tim Berners-Lee*

*Date: 2006-07-27, last change: \$Date: 2009/06/18 18:24:33 \$*

*Status: personal view only. Editing status: imperfect but published.*

[Up to Design Issues](#)

### Linked Data

The **Semantic Web** isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.

Like the web of hypertext, the web of data is constructed with documents on the web. However, unlike the web of hypertext, where links are relationships anchors in hypertext documents written in HTML, for data they links between arbitrary things described by RDF,. The URIs identify any kind of object or concept. But for HTML or RDF, the same expectations apply to make the web grow:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF\*, SPARQL)
4. Include links to other URIs. so that they can discover more things.

Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps. This article discusses solutions to these problems, details of implementation, and factors affecting choices about how you publish your data.

### The four rules

I'll refer to the steps above as rules, but they are expectations of behavior. Breaking them does not destroy anything, but misses an opportunity to make data interconnected. This in turn limits the ways it can later be reused in unexpected ways. It is the unexpected re-use of information which is the value added by the web.





## 3. Make the Output of NLP Tools available on the Web



Confidence:  0.0

Contextual score:  0.0

Prominence (support):  0

No 'common words'

Default Disambiguation

Show best candidate

The [Semantic Web](#) isn't just about putting data on the [web](#). It is about making [links](#), so that a [person](#) or [machine](#) can explore the web of data. With [linked data](#), when you have some of it, you can find other, related, data.





## 3. Make the Output of NLP Tools available on the Web

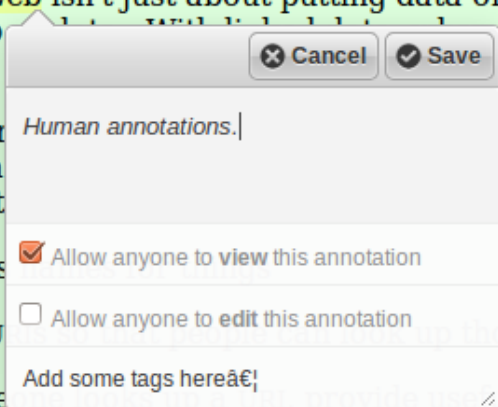
### Linked Data

The **Semantic Web** isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web. If you have some of it, you can find other, related, data.

Like the web of hypertext, where links connect between arbitrary documents, the Semantic Web has the same expectation: data is structured with documents on the web. However, unlike the web of hypertext documents written in HTML, for data they links URIs identify any kind of object or concept. But for HTML or RDF, the same expectation is to have a uniform way to describe data:

1. Use URIs as identifiers for concepts and objects.
2. Use HTTP URIs for documents that describe those names.
3. When some information is available, provide it in a uniform way, using the standards (RDF\*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

Simple. In fact, though, a surprising amount of data isn't linked in 2006, because of problems with one or more of the steps. This article discusses solutions to these problems, details of implementation, and factors affecting choices and





## 3. Make the Output of NLP Tools available on the Web

NLP Interchange Format (NIF) join the mailing list at:  
<http://nlp2rdf.org>

@PREFIX : <a href="http://www.w3.org/DesignIssues/LinkedData.html#">http://www.w3.org/DesignIssues/LinkedData.html#</a>	
Scheme 1: Offset-Based	<b>offset_717_729</b> Identifier _ Begin Index _ End Index
:offset_717_729 sso:oen dbpedia:Semantic_Web ; rev:hasComment "Hey Tim, good idea that Semantic Web!" .	
Scheme 2: Context-Hash- Based	<b>hash_10_12_60f02d3b96c55e137e13494cf9a02d06_Semantic%20Web</b> Identifier _ Context length _ String length _ MD5 Hash _ String MD5 Hash = md5 (" The (Semantic Web) isn't jus")
:hash_10_12_60f02d3b96c55e137e13494cf9a02d06_Semantic%20Web sso:oen dbpedia:Semantic_Web ; rev:hasComment "Hey Tim, good idea that Semantic Web!" .	



## Contact

### Address

University of Leipzig  
Faculty of Mathematics and Computer  
Science  
Institute of Computer Science  
Department of Business Information  
Systems

Postfach 100920  
04009 Leipzig  
Germany

UNIVERSITÄT LEIPZIG

AKSW



Project: <http://lod2.eu>

Organisation: <http://uni-leipzig.de>, <http://aksw.org>

Presenter: <http://bis.informatik.uni-leipzig.de/SebastianHellmann>

NLP2RDF page: <http://nlp2rdf.org>

Thanks for your  
attention!