towards
pan european lexicology and lexicography
by means of linked (open) data

eveline wandl-vogt + thierry declerck
ICLTT @ austrian academy of sciences @ vienna. AT
german research institute for artificial intelligence @ saarbrücken. DE
COST IS 1305: ENeL 2014. september 29th

outline

I frame conditions:
pan european lexicology + lexicography

outline

I frame conditions:
pan european lexicology + lexicography
linked (open) data

outline

I frame conditions:
pan european lexicology + lexicography
linked (open) data

II modeling: first results

outline

I frame conditions:
pan european lexicology + lexicography
linked (open) data

II modeling: first results

III follow up challenges

presented by eveline. thierry

outline

I frame conditions:
pan european lexicology + lexicography
linked (open) data

II modeling: first results

III follow up challenges

outline

I frame conditions:
pan european lexicology + lexicography
linked (open) data

II modeling: first results

III follow up challenges

EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

EUROPEAN NETWORK OF e-LEXICOGRAPHY

outline

I frame conditions:
pan european lexicology + lexicography
linked (open) data

II modeling: first results

III follow up challenges

EUROPEAN COOPERATION IN SCIENCE AND TECHNOLOGY

EUROPEAN NETWORK OF e-LEXICOGRAPHY

DARIAH-EU Digital Research Infrastructure for the Arts and Humanities

frame conditions I

pan european
lexicology + lexicography

point of view

- 
- 

frame conditions I

pan european
lexicology + lexicography

point of view

- national ←→ supranational

pan european
lexicology + lexicography

consequences: focus on commonalities

- structures
- concepts
- comparative linguistics
- etymology
- cultural background

frame conditions I

pan european
lexicology + lexicography

consequences: focus on commonalities

- structures
- concepts
- comparative linguistics
- etymology
- cultural background

eurolinguistics

pan european
lexicology + lexicography

consequences: focus on commonalities

- multilingual
- structure
- cultural diverse
  cultural frame is europe

frame conditions I

pan european
lexicology + lexicography

consequences: focus on commonalities

- multilingual
- structure
- cultural diverse
  cultural frame is europe

`eurolexicography´

by means of linked data towards pan european lexicology and lexicography: examples

examples

1) pan european words?
2) pan european concepts?

by means of linked data towards pan european lexicology and lexicography: examples

examples

1) pan european words?
2) pan european concepts?

3) aligned pan european corpora

by means of linked data towards pan european lexicology and lexicography: examples

examples

1) pan european words?
2) pan european concepts?

3) aligned pan european corpora

4) interlinking of dictionaries

by means of linked data towards pan european lexicology and lexicography: examples

examples

1) pan european words?
2) pan european concepts?

} WG4

3) aligned pan european corpora

} WG3

4) interlinking of dictionaries

} WG1

towards eurolexicography
pan european words?

research questions

- common roots ← etymology
- common neologisms

towards eurolexicography
pan european concepts?

research questions

- quantitative analysis of representation of a concept
- concept based dictionary access

towards eurolexicography
aligned pan european corpora

research questions

- lexical acquisition
  aligned pan european corpora as source for a pan european dictionary

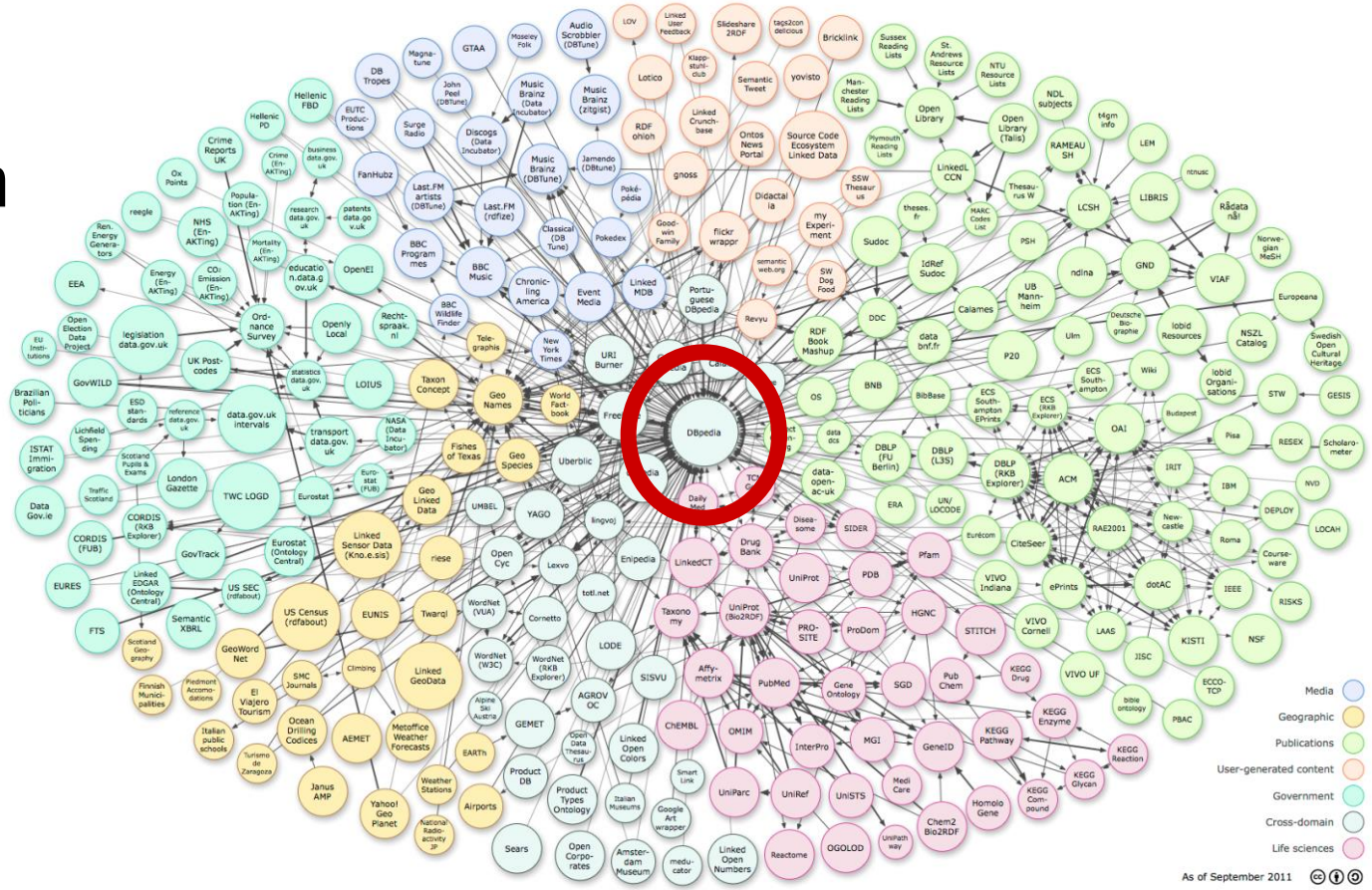
  eg EUROPARL

towards eurolexicography
interlinking of dictionaries

research questions

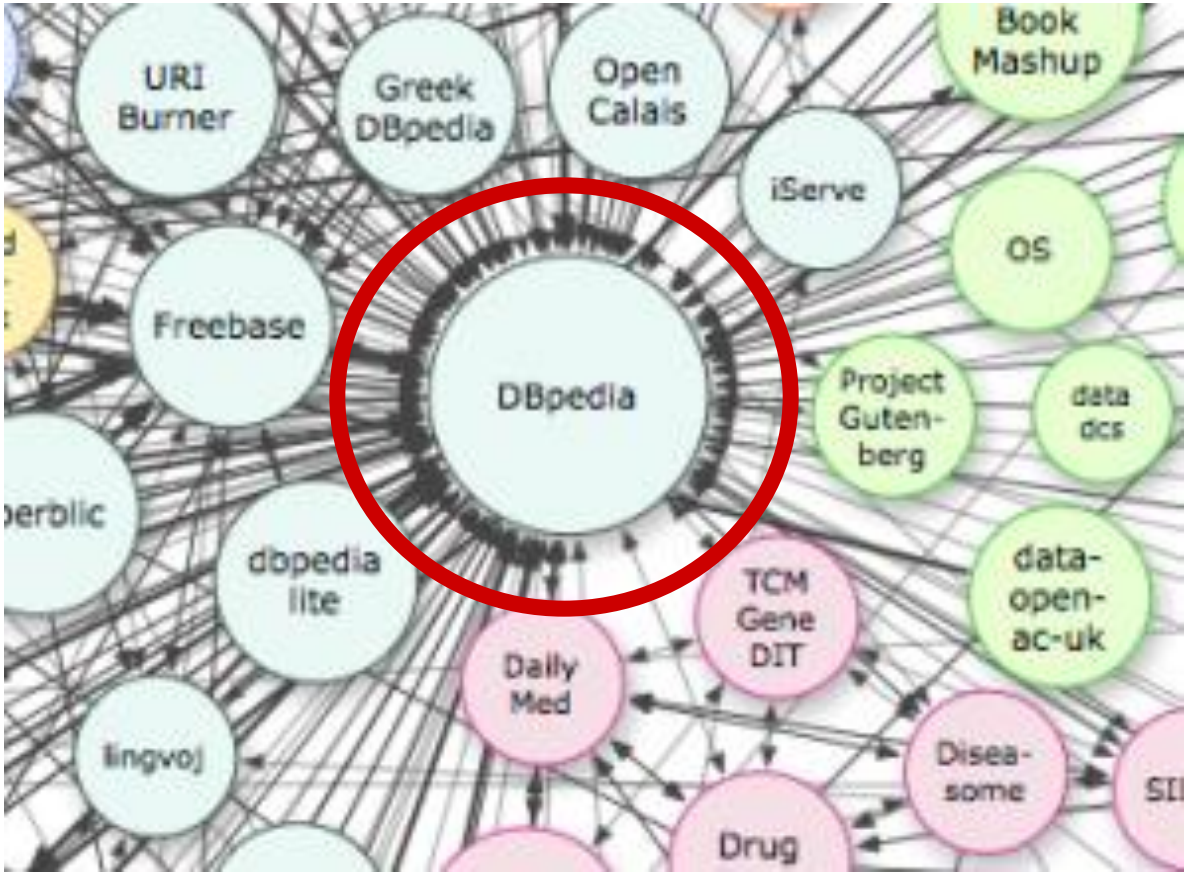
- interfaces of dictionaries
- data aggregation
- data reuse

LOD graph



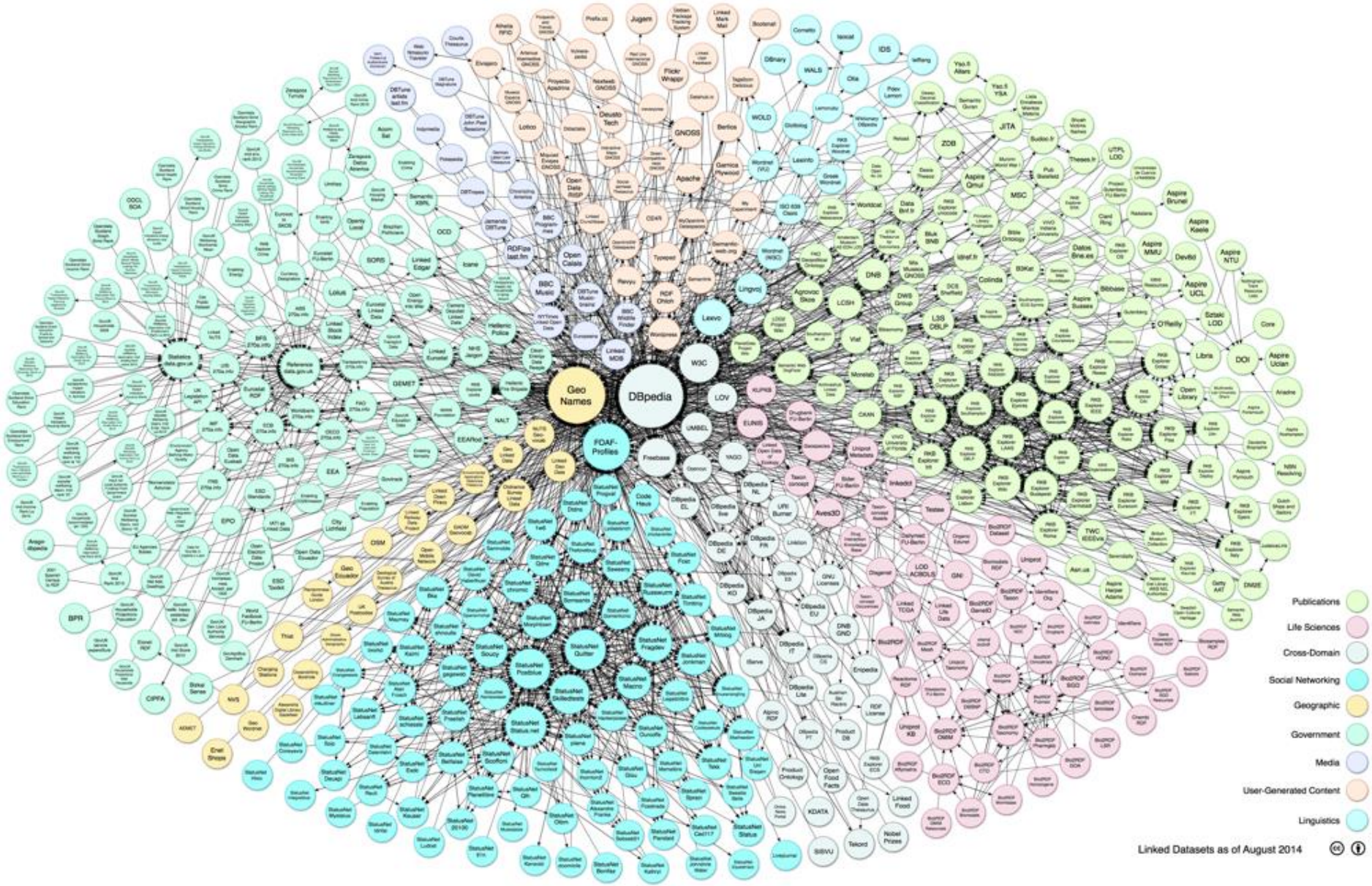
http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html

LOD graph

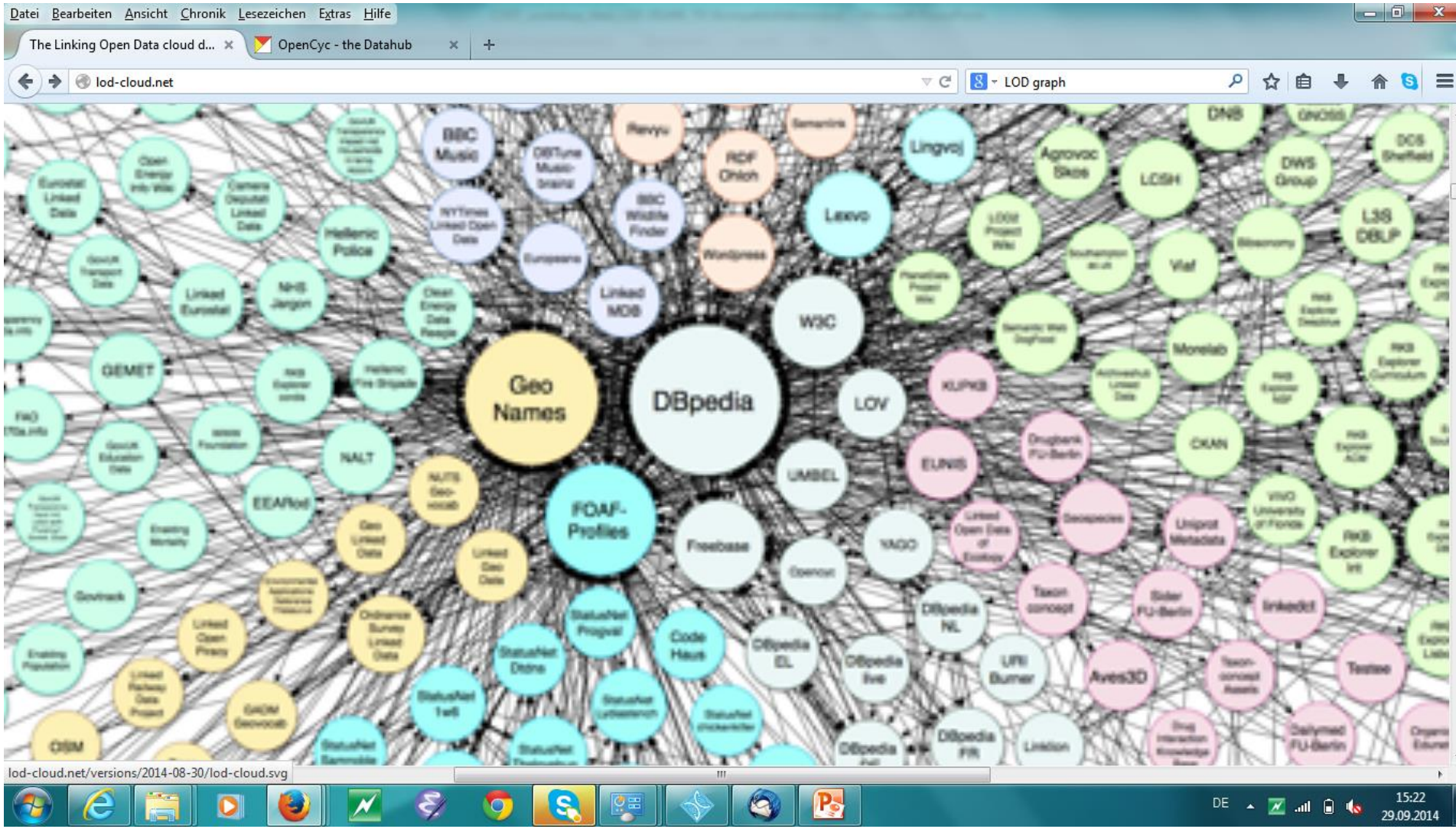


http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html

LOD graph 2014-08-30

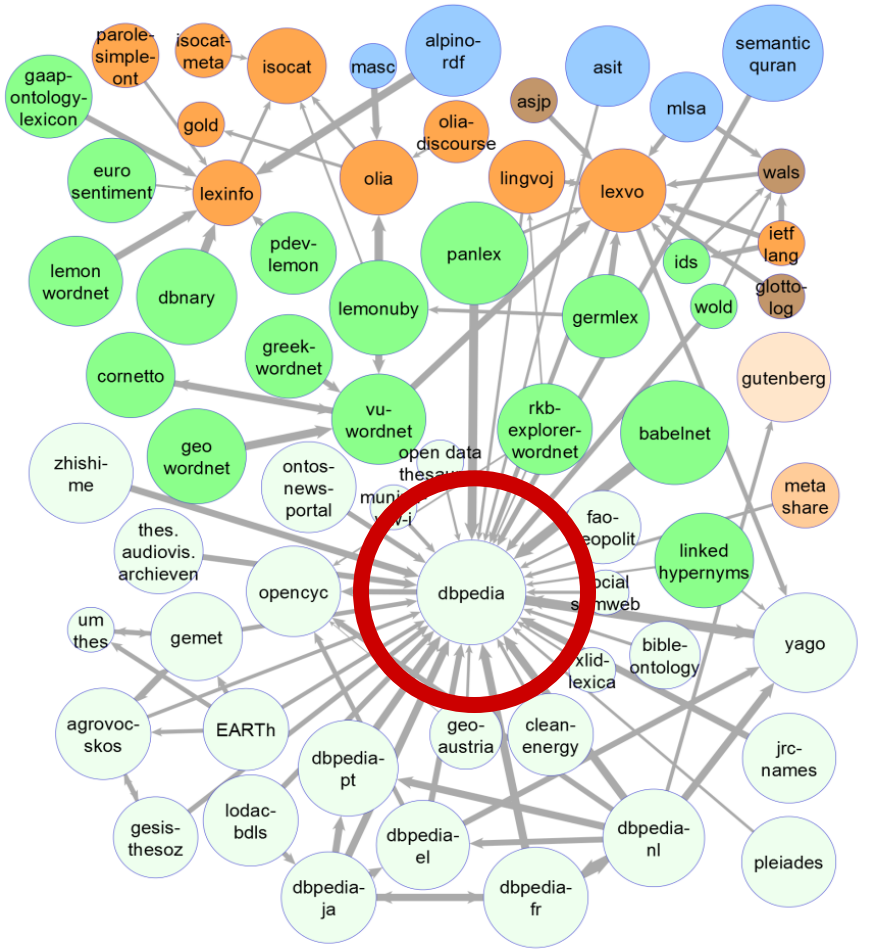


LOD graph 2014-08-30



frame conditions II

the linguistic linked open data graph



principles
linked open data

a „light“ or „shallow“ or „robust“ version of the semantic web
(the 5 stars mug, ☞ http://www.w3.org/DesignIssues/LinkedData.html)

frame conditions II

concepts

W3C Ontolex CG

Wandl-Vogt, Eveline - Outl... | Achtung - Englisch - Deuts... | cloose look - Google-Suche | Ontology-Lexica Commun...

www.w3.org/community/ontolex/

ontolex

LOG IN GET AN ACCOUNT MY ACCOUNT SKIP

W3C Community and Business Groups

Search blogs

CURRENT GROUPS REPORTS ABOUT

Mailing List
Wiki
RSS
Contact Group

Home / Ontology-Lexica Community...

Ontology-Lexica Community Group

The mission of the Ontology-Lexicon community group is to: (1) Develop models for the representation of lexica (and machine readable dictionaries) relative to ontologies. These lexicon models are intended to represent lexical entries containing information about how ontology elements (classes, properties, individuals etc.) are realized in multiple languages. In addition, the lexical entries contain appropriate linguistic (syntactic, morphological, semantic and pragmatic) information that constrains the usage of the entry. (2) Demonstrate the added value of representing lexica on the Semantic Web, in particularly focusing on how the use of linked data principles can allow for the re-use of existing linguistic information from resource such as WordNet. (3) Provide best practices for the use of linguistic data categories in combination with lexica. (4) Demonstrate that the creation of such lexica in combination with the semantics contained in ontologies can improve the performance of NLP tools. (5) Bring together people working on standards for representing linguistic information (syntactic, morphological, semantic and pragmatic) building on existing initiatives, and identifying collaboration tracks for the future. (6) Cater for interoperability among existing models to represent and structure linguistic information. (7) Demonstrate the added value of applications relying on the use of the combination of lexica and

Get involved!

Anyone may join this Community Group. All participants in this group have signed the W3C Community Contributor License Agreement (CLA).

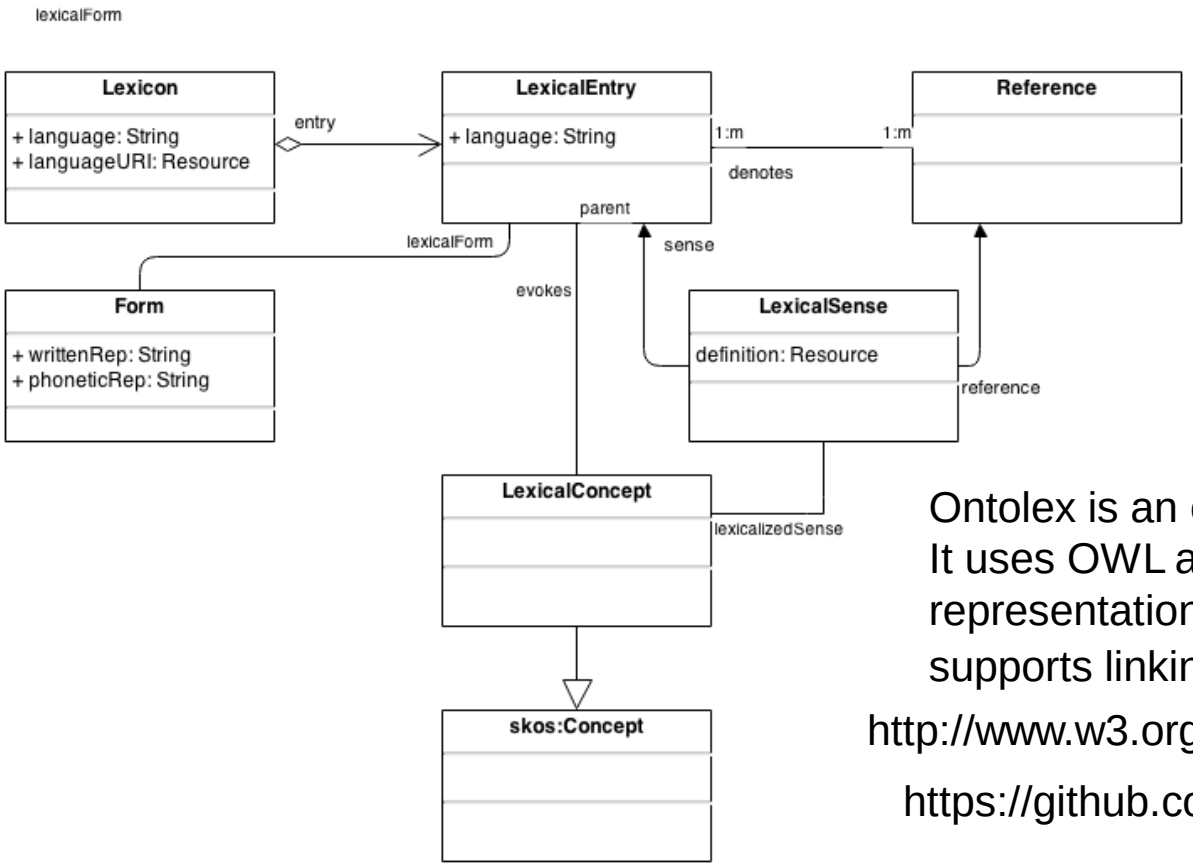
JOIN THIS GROUP

or learn how to join or request an account.

Note: Community Groups are proposed and run by the community. Although W3C hosts these conversations, the groups do not necessarily represent the views of the W3C Membership or staff.

model

W3C Ontolex CG



Ontolex is an extension of LMF
It uses OWL and RDF as representation languages, and supports linking to LOD data sets.

http://www.w3.org/community/ontolex/

https://github.com/cimiano/ontolex

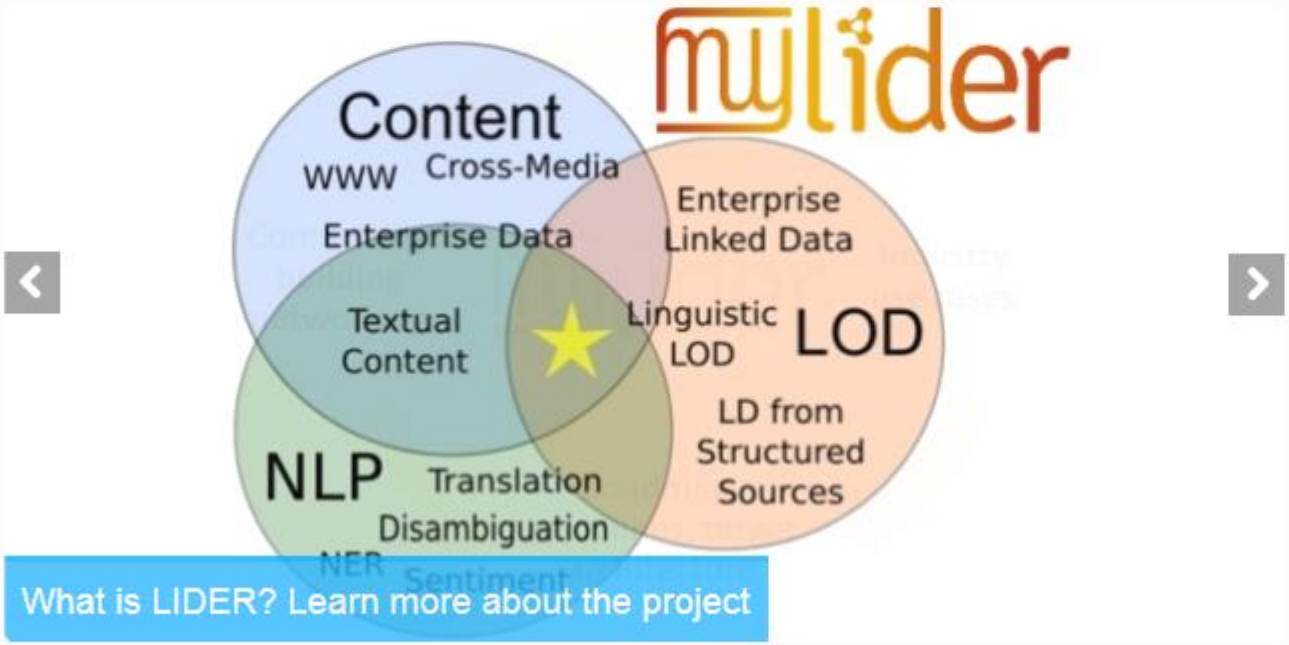
frame conditions II

LIDER | elexicography.eu | WG1-Object...

lexvoo

mylider

Home | What is LIDER | Documentation | News | Next Events | Get involved | Contact us



What is LIDER

The project's mission is to provide the basis for **Expected outcomes**:

LIDER NEWS

- Report from the W3C "MultilingualWeb" workshop in Madrid is now available
- Linked Data meets Content Analytics: 4th LIDER & LD4LT event, 2nd September, Leipzig
- Roadmapping Linked Data in the Localisation Industry Dublin 4th June'14
- LD4LT at LREC'14
- Two surveys released on the future of language technologies on the web

NEXT EVENTS

- MLODE 2014 - Leipzig, September, 1-2

european project
www.lider-project.eu

- co-operation
- LIDER use case on lexicography:
  transform data sets from
  COST ENeL-partners
  into LLOD

mylider

join in!

Modeling in Ontolex
First results

- We are currently dealing with following data:
  - 2 Austrian dialect dictionaries (Tustep/XML and Word)
  - 1 sample of a Slovak dictionary (XML and PDF/Word)
  - 1 Slovene dictionary (XML, LMF based)
  - 2 TEI encoded Arabic dialects
  - 1 Sample from a Bask-German dictionary (XML)
  - 1 Sample from a French lexicon (extracted from Wiktionary)
  - 1 Limburg questionaire/concept based list of words (Excel)
  - 1 Sample of a KDictionary (XML)
  - 1 Sample from the Digital Scottisch Lexicon (Old Scottisch, html + 1 example in TEI)
  - 1 Lexicon extracted from a corpus of „Baroque German“ (Austrian Academy of Sciences)

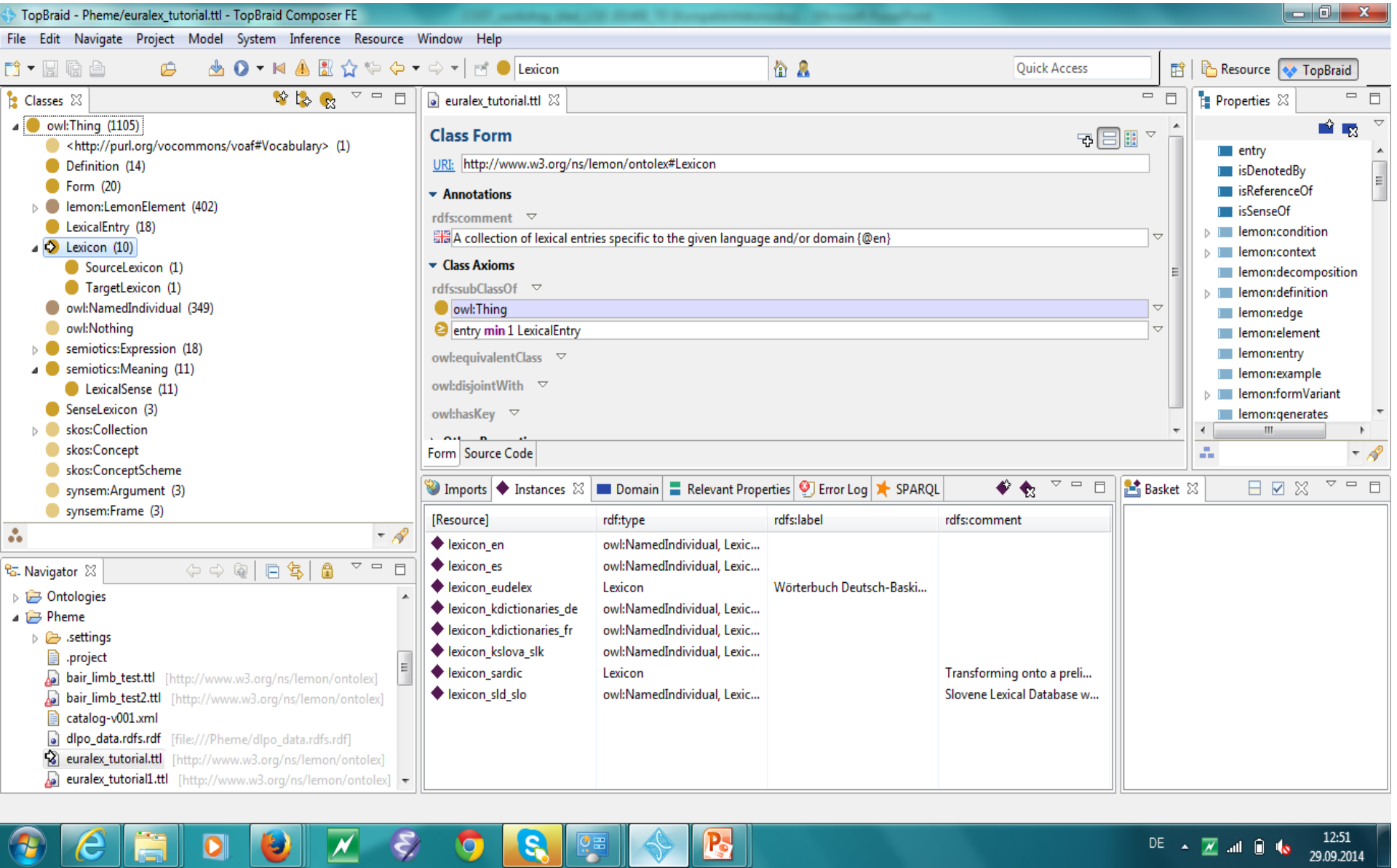
Steps in the modeling

- Manual analysis of the input dictionary data
- Comparison of the encoding of the original data and the ontolex model
- Manual „population“ of the ontolex model for some few elements of the original data, as „proof of concept“.
- Automatic „population“ of the ontolex model for the full original data set
- Manual linlkng of few entries in ontolex to dictionary external resources (to partially automatize)
  - Other lexical resources
  - Encyclopaedic resources
  - …
- Towards data aggregation/merging

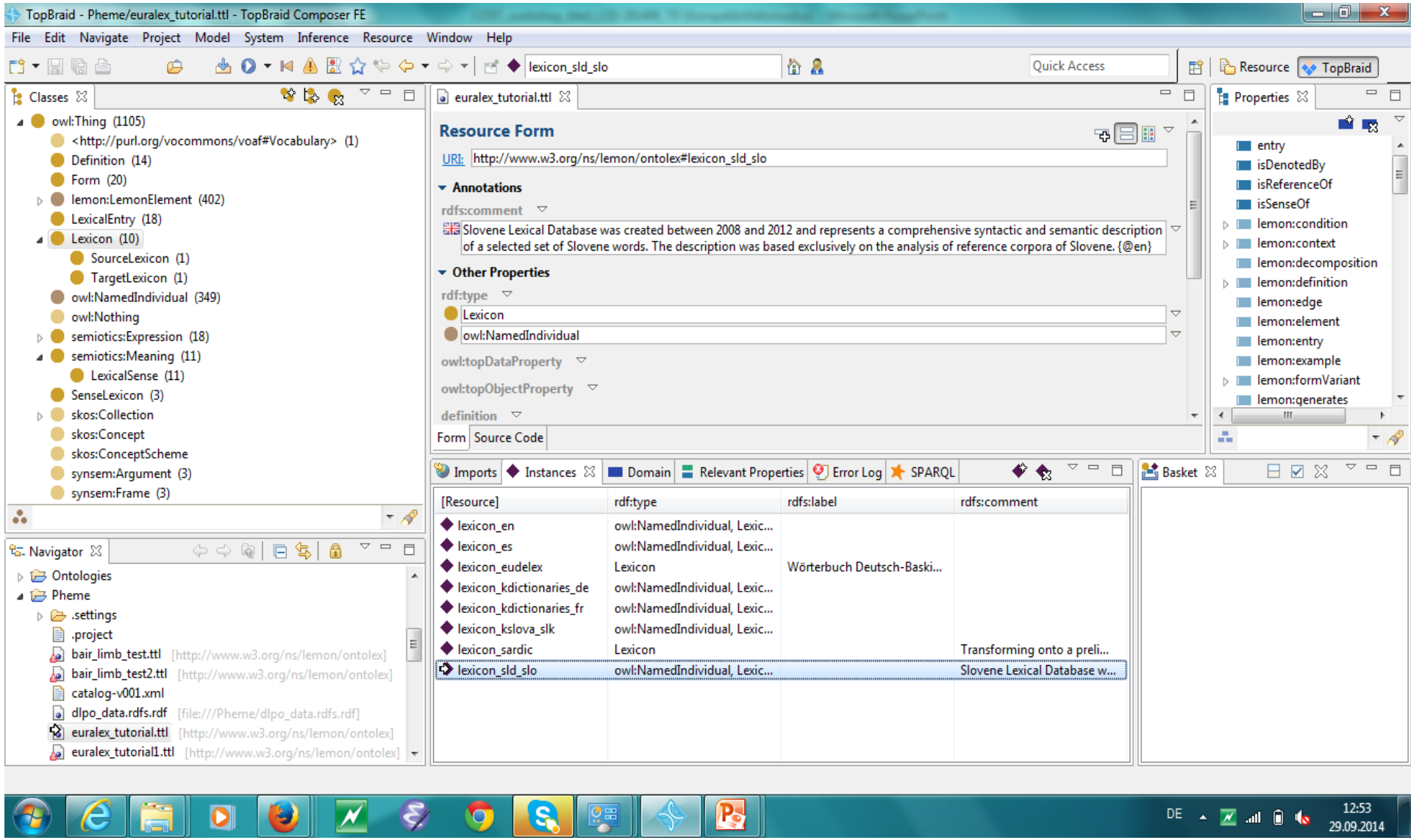
Examples

- Next slides are showing screen shots of the current implementation of the mapping between the original dictionary data and the Ontolex model.
  - We used the free edition of TopBraid for editing and visualization (http://www.topquadrant.com/downloads/topbraid-composer-install/; there select: free edition)
  - One can also use the Protégé editor (http://www.topquadrant.com/downloads/topbraid-composer-install/) or upload her/his OWL/RDF data onto Web Protégé – there are then published on the web (http://protegewiki.stanford.edu/wiki/WebProtege)

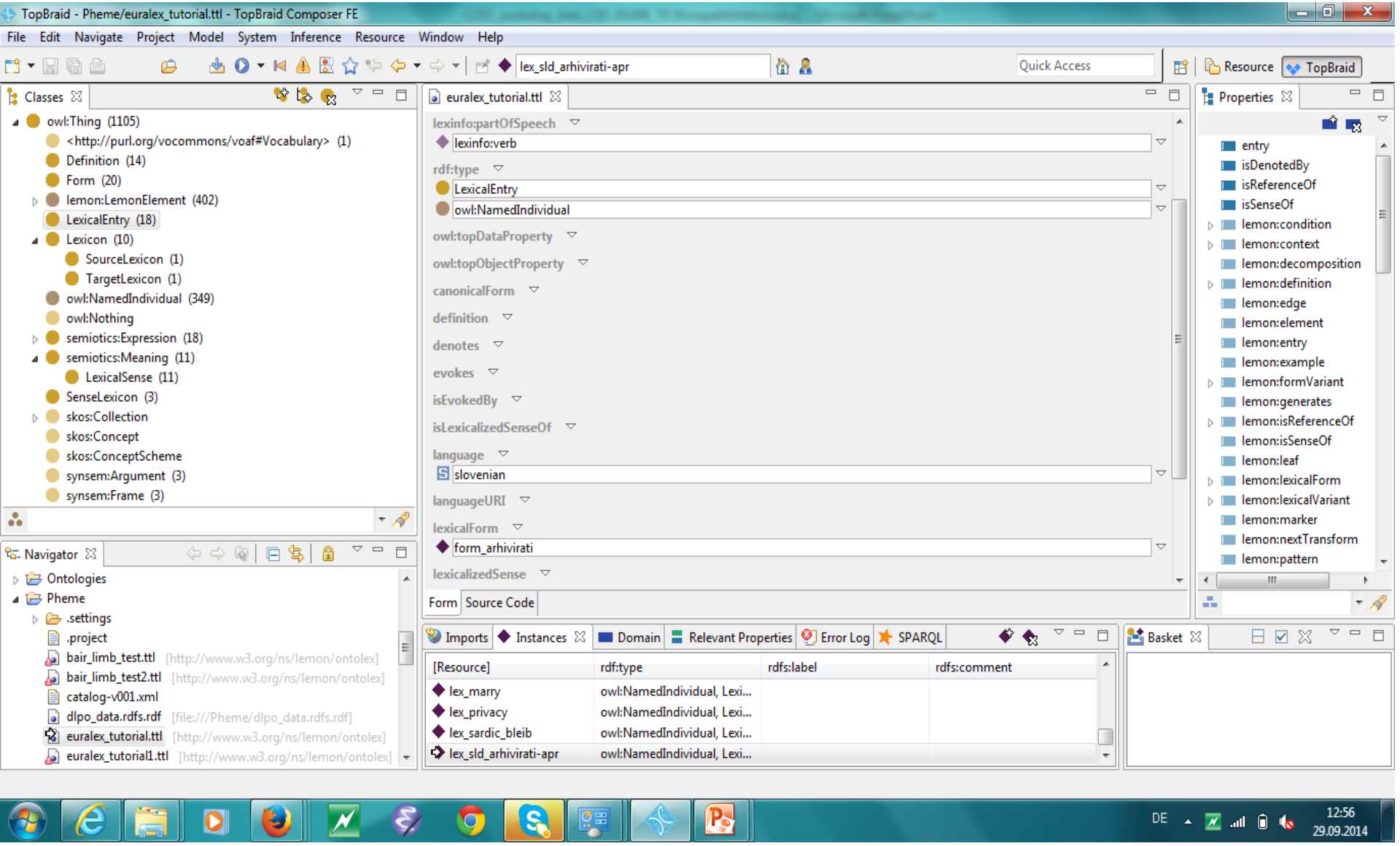
lexicon encoding in ontolex



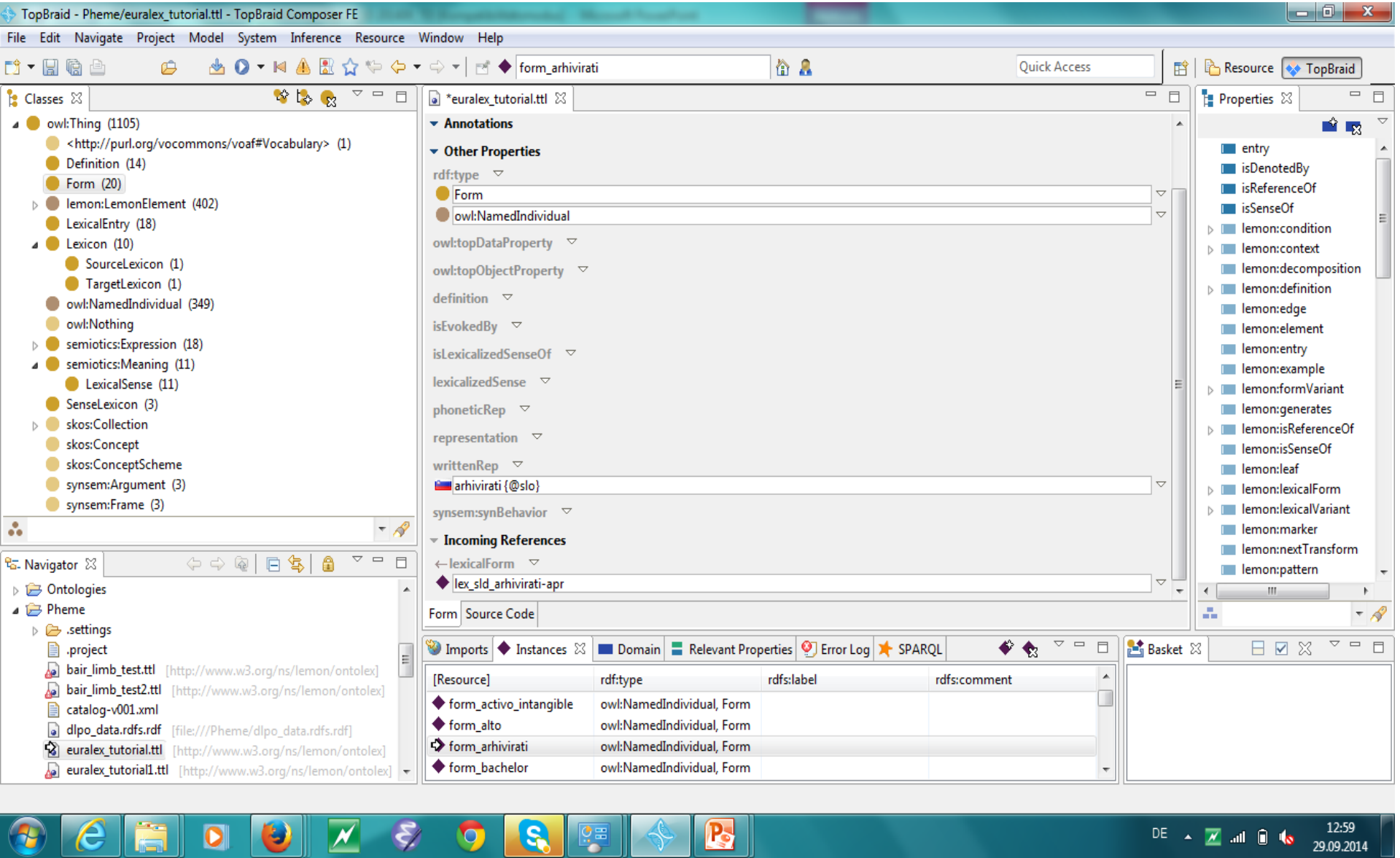
encoding of a lexicon instance in ontolex



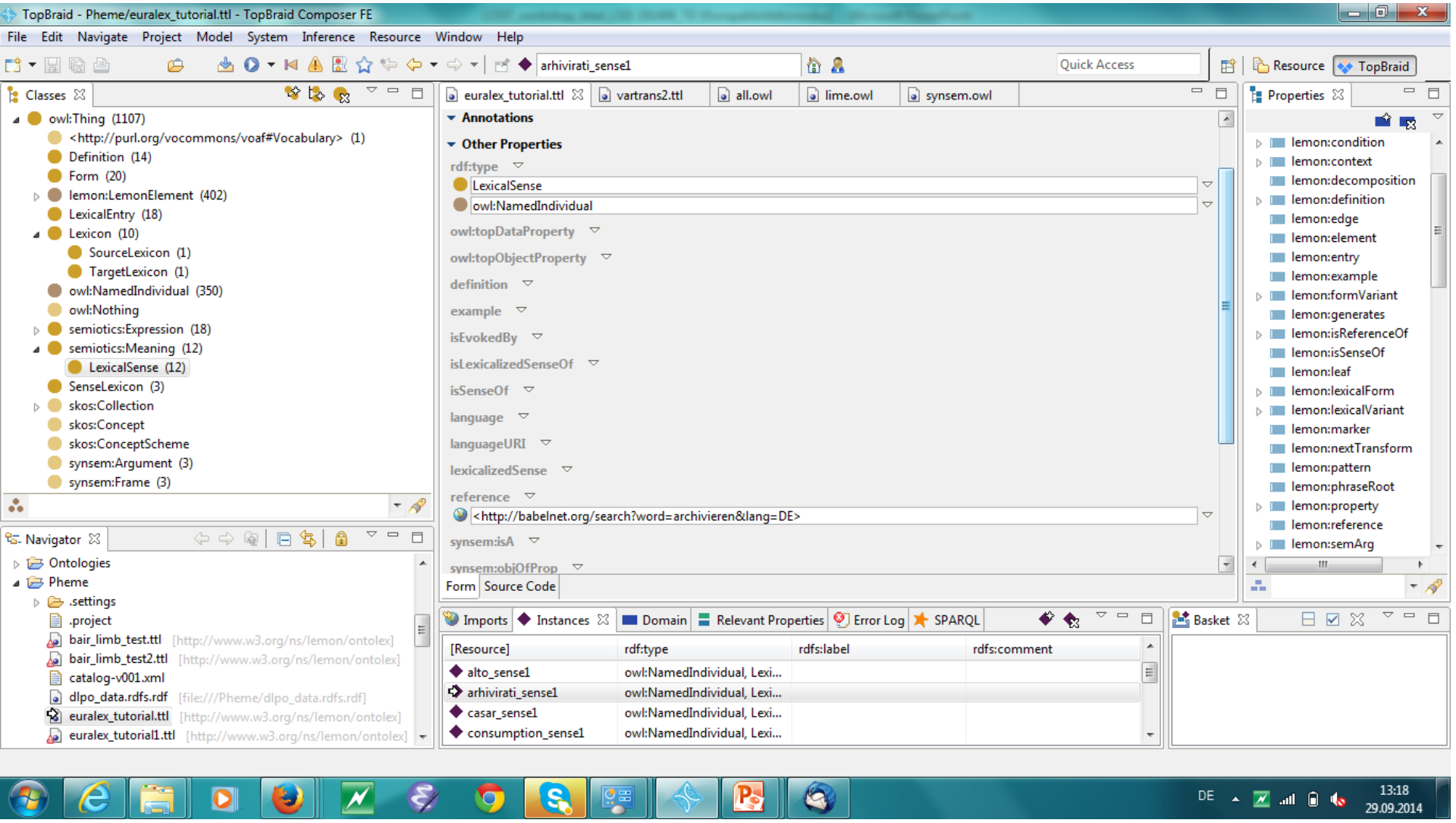
lexical entry in ontolex with intances



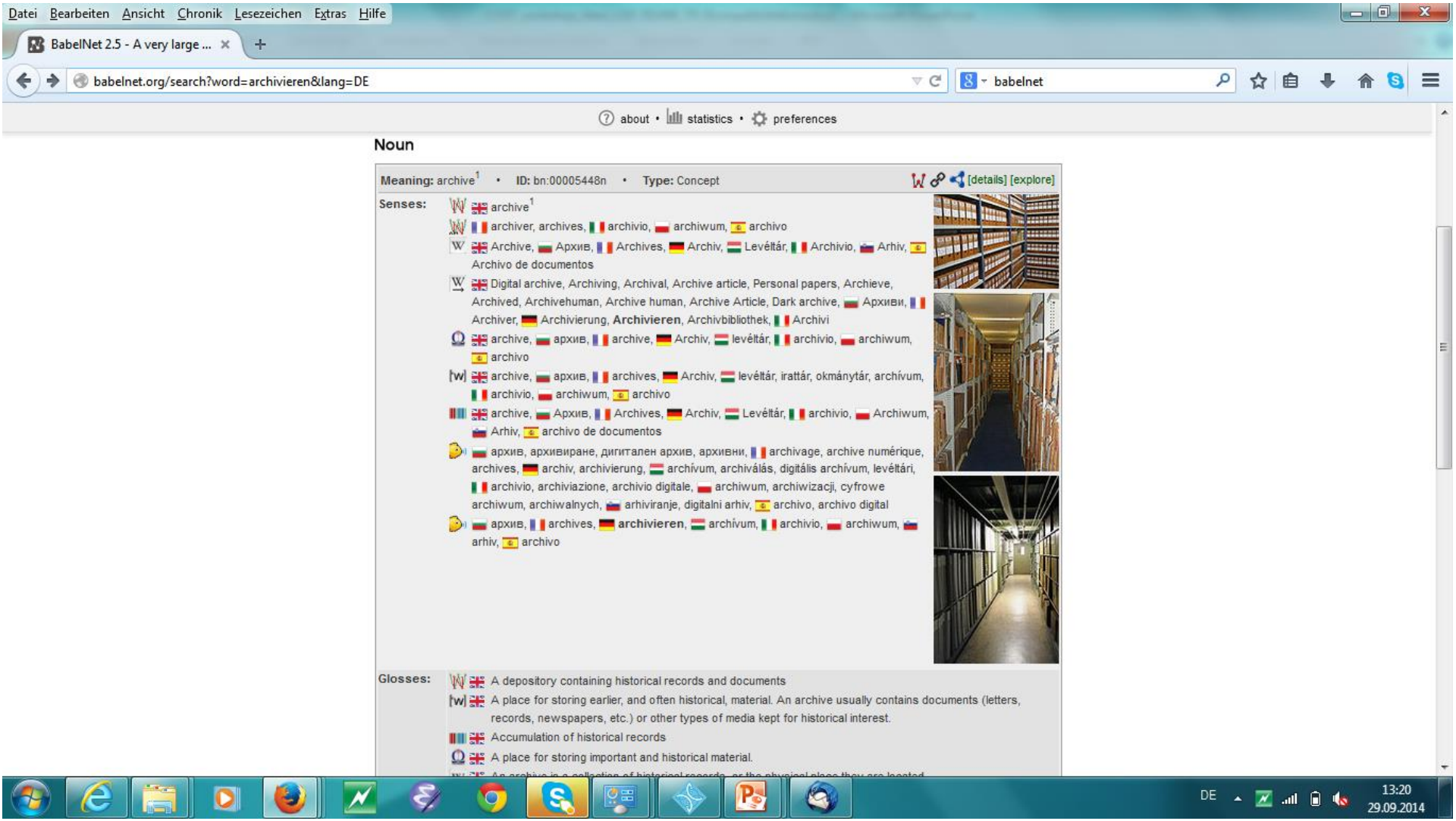
written representation of an entry



lexical sense of an entry
+ link to external semantic references



BabelNet
als target of external semantic reference I

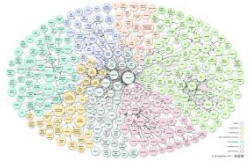


BabelNet
als target of external semantic reference II

challenges

- contribute into further developing of existing models, standards
- pilot project (portal + eurolinguistics)
- pilot project for using LOD for dictionary compiling
- increasing amount of data in the LD ←→ access
- licensing, towards open science
- towards collaborative scientific lexicography
  → virtual research environments

towards
pan-european lexicology and lexicography
by means of linked (open) data



eveline wandl-vogt + thierry declerck
ICLTT @ austrian academy of sciences @ vienna. AT
deutsches forschungszentrum für künstliche intelligenz @ saarbrücken. DE
COST IS 1305: ENeL 2014. september 29th