

## 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 29<sup>th</sup>



I frame conditions: pan european lexicology + lexicography





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





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



II modeling: first results



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



- II modeling: first results
- III follow up challenges

presented by eveline. thierry



- I frame conditions: pan european lexicology + lexicography linked (open) data
- II modeling: first results
- III follow up challenges





- I frame conditions: pan european lexicology + lexicography linked (open) data
- II modeling: first results
- III follow up challenges



IN SCIENCE AND TECHNOLOGY e-LEXICOGRAPHY

EUROPEAN COOPERATION



- I frame conditions: pan european lexicology + lexicography linked (open) data
- II modeling: first results
- III follow up challenges



**NETWORK** OF

IN SCIENCE AND TECHNOLOGY e-LEXICOGRAPHY

**ARIAH-EU** | Digital Research Infrastructure for the Arts and Humanities

EUROPEAN COOPERATION



#### point of view









#### point of view

• national  $\leftarrow \rightarrow$  supranational







#### consequences: focus on commonalities

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



#### consequences: focus on commonalities

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

eurolinguistics



#### consequences: focus on commonalities

- multilingual
- structure
- cultural diverse cultural frame is europe



#### consequences: focus on commonalities

- multilingual
- structure
- cultural diverse cultural frame is europe







#### examples

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





#### examples

- 1) pan european words?
- 2) pan european concepts?
- 3) aligned pan european corpora





#### examples

- 1) pan european words?
- 2) pan european concepts?
- 3) aligned pan european corpora
- 4) interlinking of dictionaries





#### examples

pan european words?
 pan european concepts?
 aligned pan european corpora
 WG3
 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



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

EURALEX 2014

#### frame conditions II



LOD graph



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

EURALEX 2014

## LOD graph 2014-08-30





## LOD graph 2014-08-30



#### frame conditions II

DESCICLTT

the linguistic linked open data graph



2014-07-15



## 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)





## concepts W3C Ontolex CG

Membership or staff.

🗿 Wandl-Vogt, Eveline - Outl 🗴	📡 Achtung - Englisch - Deuts × 🔱 cloose look - Google-Suche 🗴 🐝 Ontology-Lexica Commun × +		
www.w3.org/community/ontol	≪/ ♥ C S * ontole		<b>^</b>
W3C 🐝	W3C Community and Business Groups	Search blogs	Q
CURRENT GROUPS	REPORTS ABOUT Home / Ontology-Lexica Community		
📔 Mailing List	Ontology-Lexica Community Group	Get involved!	
SS RSS	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	Anyone may join this Community Group. All participants in this group have signed the W3C Community Contributor License Agreement (CLA).	
	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	JOIN THIS GROUP or learn how to join or request an account. Note: Community Groups are proposed and	
	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)	run by the community. Although W3C hosts these conversations, the groups do not necessarily represent the views of the W3C	

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 levica and



## model W3C Ontolex CG

lexicalForm



×

×

₩ *P* 

# mulider



+



## european project www.lider-project.eu

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



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

(<u>http://www.topquadrant.com/downloads/topbraid-composer-install/</u>; there select: free edition)

 One can also use the Protégé editor (<u>http://www.topquadrant.com/downloads/topbraid-composer-install/</u>) 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

3/

S

0

N

e

0



TopBraid - Pheme/euralex_tutorial.ttl - TopBraid Composer FE	COT, and day, Mal. (	and Theory and the	and second lines in			_ 0 <mark>_ x</mark> _
File Edit Navigate Project Model System Inference Resource	Window Help					
E <sup>*</sup> • 🛛 🖻 😐 🖉 • 🖉 🖕 🔶 •	🗘 🔻 🖻 📥 🗠	)	A	Quick Access		🏠 Resource 💊 TopBraid
😫 Classes 🛛 🔰 🔁 🗖 🗖	■ euralex_tutorial.ttl 🛛					Properties 🛛 🗖 🗖
<ul> <li>classes A</li> <li>owl:Thing (1105)</li> <li><http: purl.org="" voaf#vocabulary="" vocommons=""> (1)</http:></li> <li>Definition (14)</li> <li>Form (20)</li> <li>Iemon:LemonElement (402)</li> <li>LexicalEntry (18)</li> <li>LexicalEntry (18)</li> <li>Lexicon (10)</li> <li>SourceLexicon (1)</li> <li>TargetLexicon (1)</li> <li>owl:NamedIndividual (349)</li> <li>owl:Nothing</li> <li>semiotics:Expression (18)</li> <li>semiotics:Meaning (11)</li> <li>LexicalSense (11)</li> <li>SenseLexicon (3)</li> <li>skos:Collection</li> </ul>	Resource Form       Image: Comparison of the second s					<pre>entry isDenotedBy isReferenceOf isSenseOf lemon:condition lemon:condition lemon:definition lemon:edge lemon:edge lemon:element lemon:entry lemon:example lemon:entry</pre>
skos:ConceptScheme	[] ▓a tt [▲ tt ♡]				l 💶 🛃	
synsem:Argument (3)		Domain Relevant Prope	rties 🐑 Error Log 🔫 SPARQI		Basket 2	
	[Resource]	rdf:type	rdfs:label	rdfs:comment		
Se. Navigator SS	<ul> <li>lexicon_en</li> <li>lexicon_es</li> <li>lexicon_eudelex</li> <li>lexicon_kdictionaries_de</li> <li>lexicon_kdictionaries_fr</li> <li>lexicon_kslova_slk</li> <li>lexicon_sardic</li> <li>lexicon_sld_slo</li> </ul>	owl:NamedIndividual, Lexic owl:NamedIndividual, Lexic Lexicon owl:NamedIndividual, Lexic owl:NamedIndividual, Lexic Lexicon owl:NamedIndividual, Lexic	Wörterbuch Deutsch-Baski	Transforming onto a preli Slovene Lexical Database w		

P

#### DE 🔺 📈 .all 🗈 🍫 12:53 29.09.2014

#### lexical entry in ontolex with intances

e

0

💠 TopBraid - Pheme/euralex_tutorial.ttl - TopBraid Composer FE	CONT. and Address and o	In state 1 in succession	and the second damage of the			- 0 -	x
File Edit Navigate Project Model System Inference Resource	Window Help						
CÎ • 🗌 🖻 😐 📥 🔕 • M 🔺 🖹 🏠 🐤 🔶 •	r ⇔ ▼   🖻 ♦ lex_sld_arhivin	rati-apr		Quick Access		Resource 📀 TopBraid	
😫 Classes 🛛 😵 🏷 🖳 🗖 🗖	euralex_tutorial.ttl					🚦 Properties 🛛 🗖	
<ul> <li>owl:Thing (1105)</li> <li><http: purl.org="" voaf#vocabulary="" vocommons=""> (1)</http:></li> <li>Definition (14)</li> <li>Form (20)</li> <li>lemon:LemonElement (402)</li> <li>LexicalEntry (18)</li> <li>LexicalEntry (18)</li> <li>Lexicon (10)</li> <li>SourceLexicon (1)</li> <li>TargetLexicon (1)</li> <li>owl:NamedIndividual (349)</li> <li>owl:Nothing</li> <li>semiotics:Expression (18)</li> <li>semiotics:Meaning (11)</li> <li>LexicalSense (11)</li> <li>SenseLexicon (3)</li> <li>skos:Collection</li> <li>skos:Concept</li> <li>skos:Concept</li> <li>skos:ConceptScheme</li> <li>synsem:Frame (3)</li> </ul>	lexinfo:partOfSpeech ▽         lexinfo:verb         rdf:type ▽         LexicalEntry         owl:topDataProperty ▽         owl:topObjectProperty ▽         canonicalForm ▽         definition ▽         denotes ▽         evokes ▽         isEvokedBy ▽         isLexicalizedSenseOf ▽         language ▽         Solovenian					Interpretation of the second sec	
• ~ ?	lexicalForm ▽					lemon:marker	
😘 Navigator 🛛 🧼 🗇 🗘 🖓 🖓 🍸 🗖 🗇	form_arhivirati					lemon:pattern	
> 🔁 Ontologies	lexicalizedSense ∨				-	I = 1	Þ
a 🗁 Pheme	Form Source Code						R
<ul> <li>Jestings</li> <li>Jestings&lt;</li></ul>	Imports     ◆ Instances       [Resource]       ▲ lay many	Domain Relevant Prop	perties 🔮 Error Log 🖈 SPAR( rdfs:label	QL 🔶 � 😒 ▽ 🗖 🗖 rdfs:comment	Basket		
catalog-v001.xml dlpo_data.rdfs.rdf file:///Pheme/dlpo_data.rdfs.rdf] dlpo_data.rdfs.rdf euralex_tutorial.ttl http://www.w3.org/ns/lemon/ontolex] catalog euralex_tutorial1.ttl	<ul> <li>Iex_marry</li> <li>Iex_privacy</li> <li>Iex_sardic_bleib</li> <li>Iex_sld_arhivirati-apr</li> </ul>	owi:NamedIndividual, Lexi owi:NamedIndividual, Lexi owi:NamedIndividual, Lexi owi:NamedIndividual, Lexi					

P

8/

N

8

**!** 

0



OAW

ler Wissenschafter

#### written representation of an entry



Pa

#### DE 🔺 🗾 ..... 🗊 🔥 12:59 29.09.2014

## lexical sense of an entry+ link to external semantic references

8

N

0

5

0

TopBraid - Pheme/euralex_tutorial.ttl - TopBraid Composer FE							
File Edit Navigate Project Model System Inference Resource	Window Help						
[] • 🖓 😭 🙆 📥 🔕 • 🖷 🔠 😭 🐤 🔶	🖙 🗢 🔹 📄 📑 🔶 🔤 arhivirati_sens	el	🟠 🤱		Quick Access	Ē	🏠 Resource 💊 TopBraid
😫 Classes 💥 👘 🗖 🗖	🥃 euralex_tutorial.ttl 🛛 🗋	vartrans2.ttl 💿 all.owl	lime.owl	synsem.owl			Properties 🛛 🗖 🗖
a 😑 owl:Thing (1107)	<ul> <li>Annotations</li> </ul>					-	📫 🔜 🗸
<http: purl.org="" voaf#vocabulary="" vocommons=""> (1)</http:>	<ul> <li>Other Properties</li> </ul>						Implementation
Definition (14)	rdf:type ▽						Image:
Form (20)	LexicalSense						Iemon:decomposition
lemon:LemonElement (402)	owl:NamedIndividual						Image:
LexicalEntry (18)	owl+tonDataProperty ⊽						🔲 lemon:edge 😑
Lexicon (10)	owntopbatarroperty						lemon:element
Target evicon (1)	owl:topObjectProperty ▽						lemon:entry
owl:NamedIndividual (350)	definition $\bigtriangledown$						lemon:example
owl:Nothing	example $\bigtriangledown$					=	
semiotics:Expression (18)							Immonigenerates
semiotics:Meaning (12)	ISEVOKEOBY V						lemon:isSenseOf
LexicalSense (12)	isLexicalizedSenseOf $\bigtriangledown$						lemon:leaf
SenseLexicon (3)	isSenseOf ▽						Iemon:lexicalForm
skos:Collection							Image:
skos:Concept	language						lemon:marker
skos:ConceptScheme	languageURI ▽						lemon:nextTransform
synsem:Argument (3)	lexicalizedSense  □						Iemon:pattern
synsem:Frame (3)	reference 🗢						lemon:phraseRoot
•• • • •	<http: babelnet.org="" p="" sear<=""></http:>	ch?word=archivieren⟨=DE	>			~	Image:
😘 Navigator 🛛 🔅 🗇 🗁 🗇 🗐 🖛 🐄 🗖	svnsem:isA ▽						lemon:reference
	synsennisie						Filemon:semarg
	synsem:objOtProp					<u></u>	- 0
Settings	Form Source Code						• * *
.project	🥙 Imports 🔶 Instances 🖾						
bair_limb_test.ttl [http://www.w3.org/ns/lemon/ontolex]	[Perource]	rdf.th.ma	referiabel	relferice	mment A		
bair_limb_test2.ttl [http://www.w3.org/ns/lemon/ontolex]		iuntype	TUISIADEI	Turs.co			
catalog-v001.xml	▼ arto_sensel	owi:ivamedIndividual, Lexi					
dlpo_data.rdfs.rdf [file:///Pheme/dlpo_data.rdfs.rdf]		owinvamedindividual, Lexi					
euralex_tutorial.ttl [http://www.w3.org/ns/lemon/ontolex]		owinivamedindividual, Lexi					
euralex_tutorial1.ttl [http://www.w3.org/ns/lemon/ontolex]	consumption_sensel	owinamedindividual, Lexi			<b>T</b>		

P

3



ICLTT

### BabelNet als target of external semantic reference I

> Sabelnet.org/search?word=archivieren⟨=DE				8 - babelnet	٩	☆自	+	<b>^</b>	8
		② about • 🔟 statistics • ۞ preferences	1						
	Noun								
	Meaning:	archive <sup>1</sup> • ID: bn:00005448n • Type: Concept	W do	details] [explore]					
	Jelises.	<ul> <li>archive</li> <li>archive, archives, archivio, archivum, archivo</li> <li>Archive, Archives, Archives, Archive, Archive, Archive, Archive human, Archive article, Personal papers Archived, Archivehuman, Archive human, Archive Article, Dark archive Archive, Archivehuman, Archive human, Archive Article, Dark archive Archive, Archivierung, Archivieren, Archivibiliothek, Archivio</li> <li>archive, Archivierung, Archivieren, Archivibiliothek, Archivio</li> <li>archive, apxus, archive, Archive, Archiv, archive archive</li> <li>archivo</li> <li>archive, apxus, archives, Archiv, Levéltár, archivio</li> <li>archivo, archivum, archives, Archiv, Levéltár, archiva, archivo</li> <li>archive, Archivierung, Archives, Archiv, Levéltár, archiva, archivo, archivo, archive, archive, archive, archive, archive, archivo, archiva, archive, arch</li></ul>	ilvio, Arhiv, Arhiv, Arhiv, Arhiv, Archieve, a, Archieve, a, Archieve, a, Archiwum, archiwum, Archiwum, vio, Archiwum, chívum, levéttári, , cyfrowe hivo digital archiwum, Archiwum,						
	Glosses:	<ul> <li>W # A depository containing historical records and documents</li> <li>W # A place for storing earlier, and often historical, material. An archive records, newspapers, etc.) or other types of media kept for historia</li> <li># Accumulation of historical records</li> <li>A place for storing important and historical material.</li> </ul>	usually contains docu cal interest.	ments (letters,					

#### BabelNet als target of external semantic reference II



#### Verb



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-Non Commercial-Share Alike 3.0 License. COCOM For any commercial use, please contact us.



13:20

29.09.2014

DE 🔺 📈 .iil 🔒 🍫

ICLT"





- 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  $\leftarrow \rightarrow$  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 29<sup>th</sup>