CCURL-2014 @ LREC-2014, May 26, 2014, Reykjavik, Iceland

Linguistic Linked Open Data

From collection to application

(for under-resourced languages)

Christian Chiarcos

chiarcos@uni-frankfurt.de

Steven Moran

steven.moran@uzh.ch

Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era

- Collaboration
- Sustainability
- Publication and maintenance
- Benefits of Linked Data

Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era

- Collaboration
- Sustainability
- Publication and maintenance
- Benefits of Linked Data
  - How can research on underresourced languages benefit from Semantic Web technologies, and specifically the Linked Data framework?

Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
    - e.g., Chalkan (Turkic, Altay, 1180 speakers)

Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, …)
    - e.g., Shor (Turkic, Siberia, 2800 speakers)

Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, …)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
    - e.g., Hausa [2010] (Chadic, West Africa, 34-53 mio speakers)

Defining under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, …)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions
    - e.g., Russian [2005] (Slavic, Eurasia, 150 mio speakers)

Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web
    - protocols and standards
    - links between data sets

Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web

=> Information integration

  - Structural interoperability
    - comparable formats and protocols to access data

    => use the same query language for different data sets

Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web

=> Information integration

- Structural interoperability
- Conceptual interoperability
  - develop and (re-)use a shared vocabularies for equivalent concepts

=> the same query on different data sets

Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web

=> Information integration

- Structural interoperability
- Conceptual interoperability
- Federation
  - data published on the web
    - under an open license
    - with a query interface (SPARQL end point)

=> use a single query to query different datasets

Linked Data & under-resourced Languages

- Linked Data
  - rules of best practice for publishing data on the web

=> Information integration

  - Structural interoperability
  - Conceptual interoperability
  - Federation

Now: Non-technical intro to Linked Data

Later: How does this help under-resourced languages ?

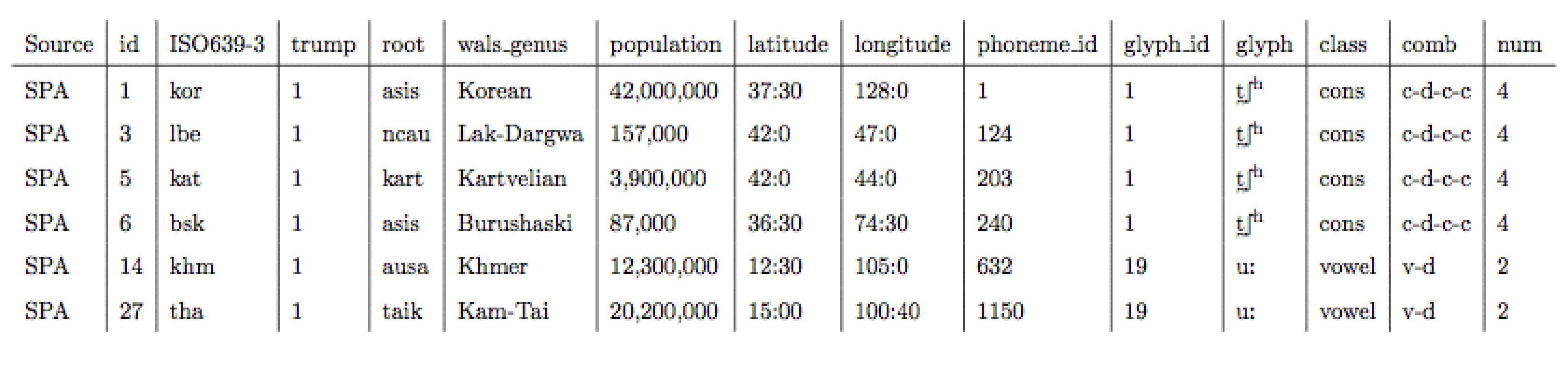
Linked Data

A non-technical introduction

From Tables to RDF to Linked Data

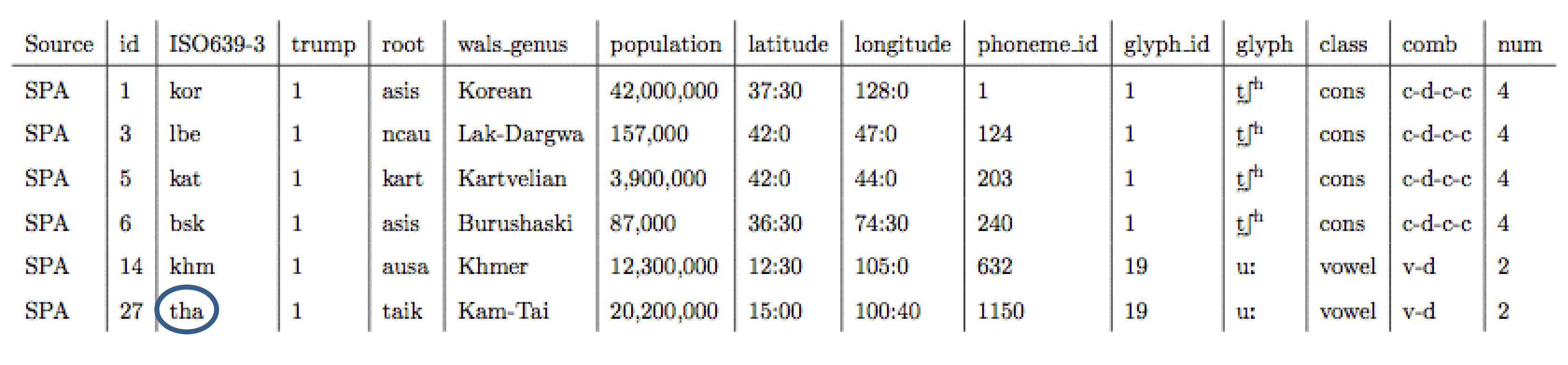
- PHOnetics Information Base and LExicon (PHOIBLE)
  - Moran, S. 2012. Using Linked Data to Create a Typological Knowledge Base. In Chiarcos, C., Nordhoff, S., and Hellmann, S. (eds), *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.
- Phoneme inventories and phonological features
  - Covers ~20% of the world's spoken languages
  - Compiled from various sources, originally a flat table (list)

From Tables …



Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

From Tables to RDF …

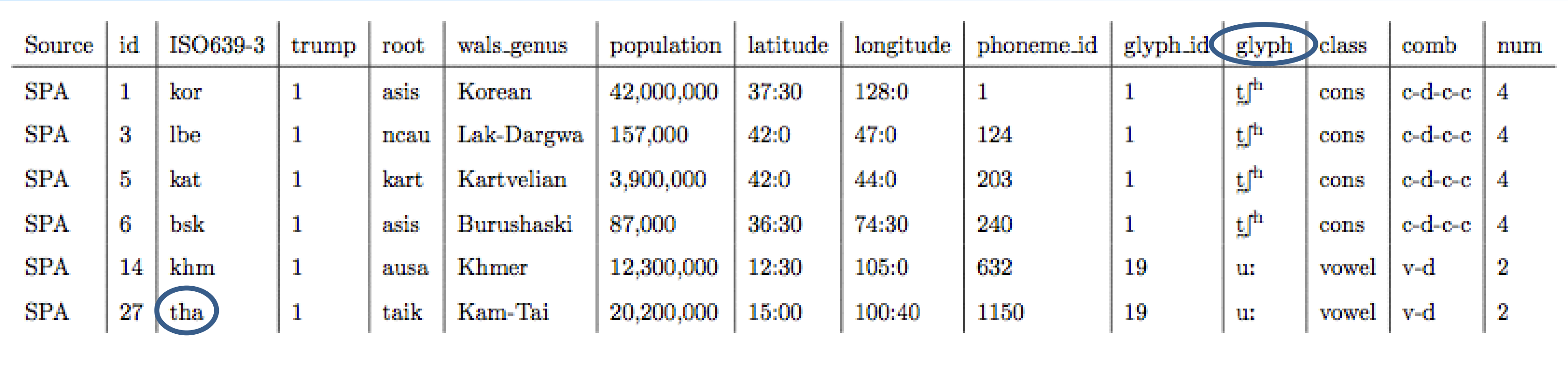


Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t̪ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t̪ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t̪ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t̪ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject
(primary key)

From Tables to RDF …

Property
(„Relation“)

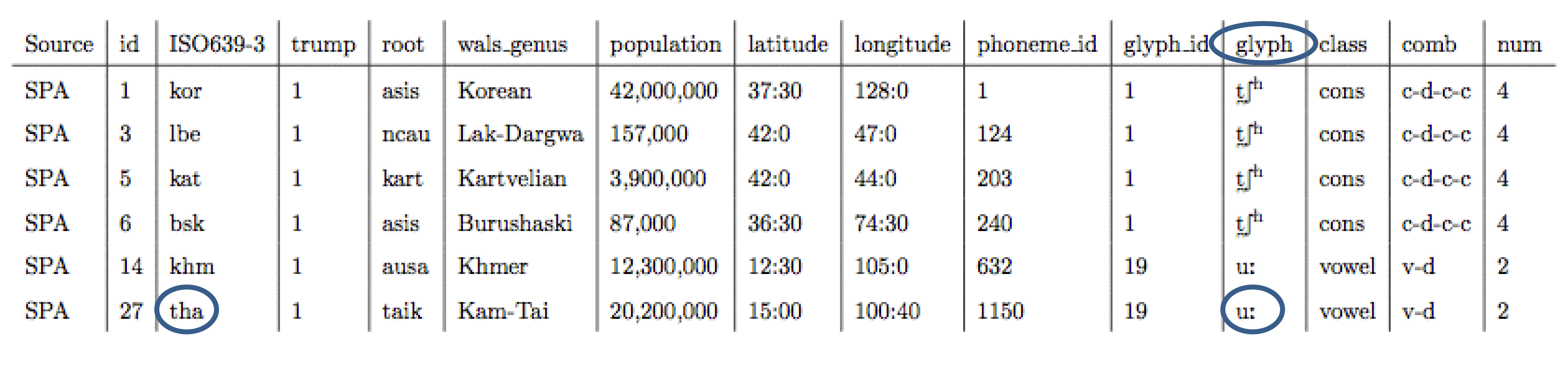


Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t͡ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject

From Tables to RDF …

Property
(„Relation“)



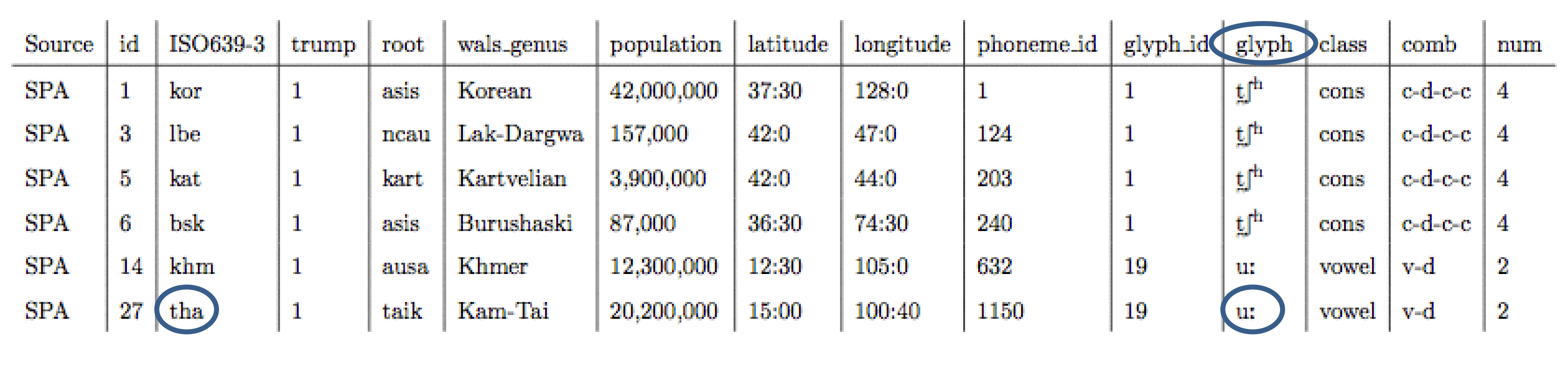
Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject

Object

From Tables to RDF …

Property („Relation“)



Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject

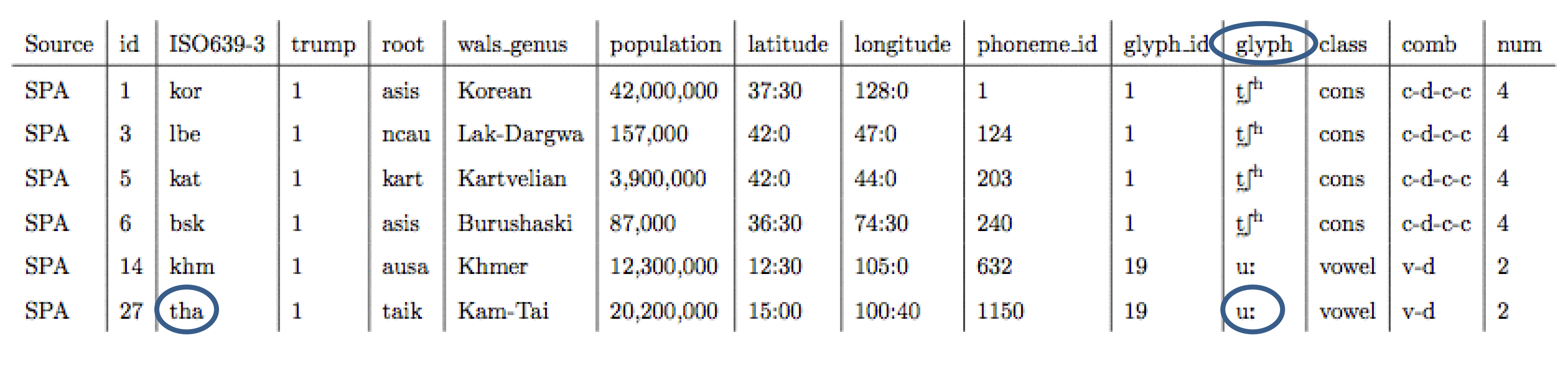
Object

1. Decompose tables into triples, i.e.,

- entity attribute value resp.
- Subject Property Object

From Tables to RDF …

Property
(„Relation“)



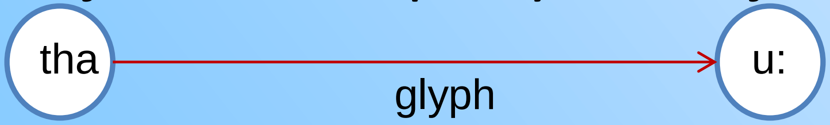
Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t͜ʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19	uː	vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai	20,200,000	15:00	100:40	1150	19	uː	vowel	v-d	2

Subject

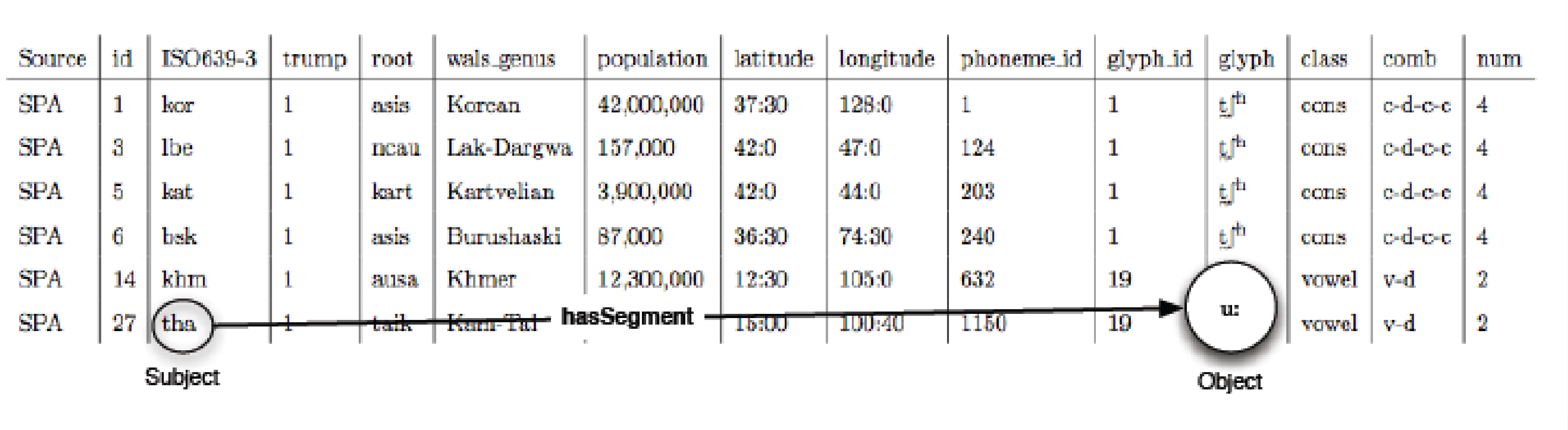
Object

1. Decompose tables into triples, i.e.,

- ❑ entity attribute value resp.
- ❑ Subject Property Object



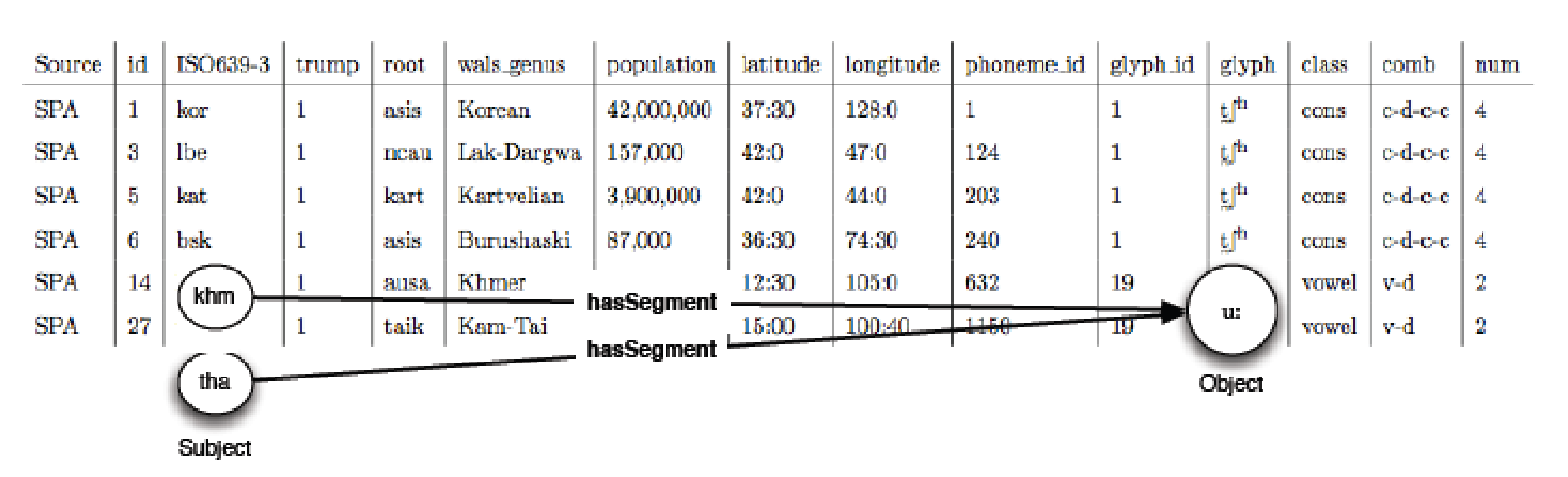
From Tables to RDF …



Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	tʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	tʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	tʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	tʃʰ	cons	c-d-c-c	4
SPA	14	khm	1	ausa	Khmer	12,300,000	12:30	105:0	632	19		vowel	v-d	2
SPA	27	tha	1	taik	Kam-Tai		15:00	100:40	1150	19	u:	vowel	v-d	2

1. Decompose tables into triples, i.e.,
   - entity attribute value resp.
   - Subject Property Object

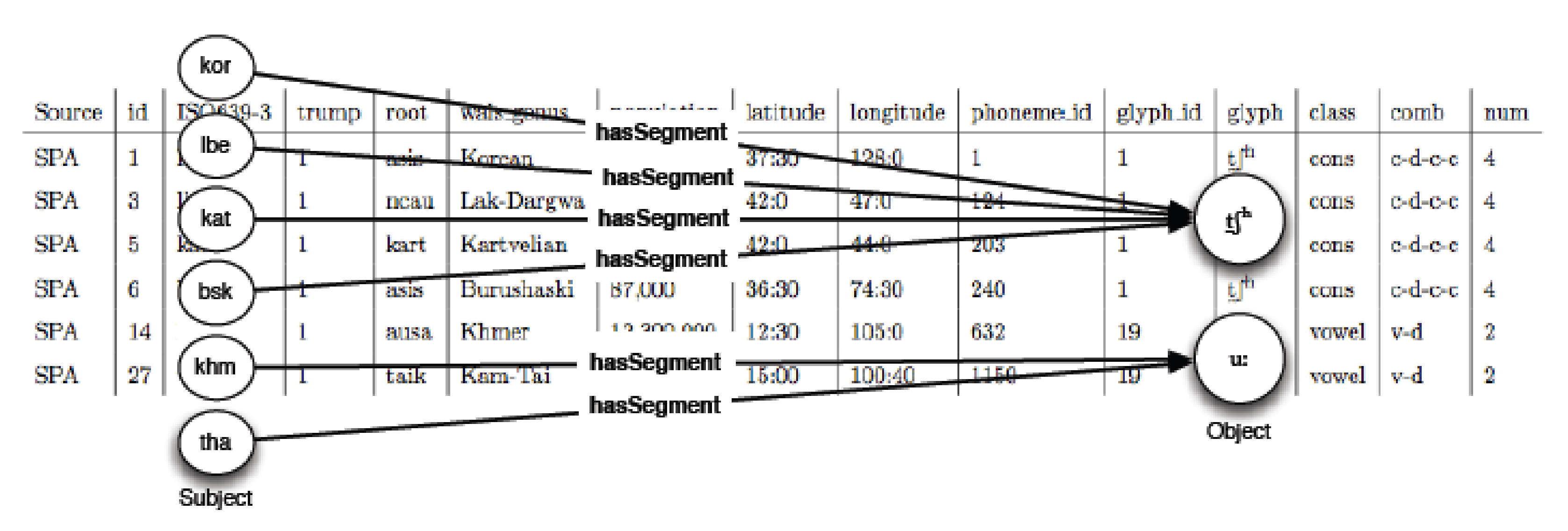
From Tables to RDF …



Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	kor	1	asis	Korean	42,000,000	37:30	128:0	1	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	3	lbe	1	ncau	Lak-Dargwa	157,000	42:0	47:0	124	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	5	kat	1	kart	Kartvelian	3,900,000	42:0	44:0	203	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	6	bsk	1	asis	Burushaski	87,000	36:30	74:30	240	1	t̠ʃʰ	cons	c-d-c-c	4
SPA	14		1	ausa	Khmer		12:30	105:0	632	19		vowel	v-d	2
SPA	27		1	taik	Kam-Tai		15:00	100:40	1150	19		vowel	v-d	2

1. Decompose tables into triples
2. Multiple triples constitute a graph

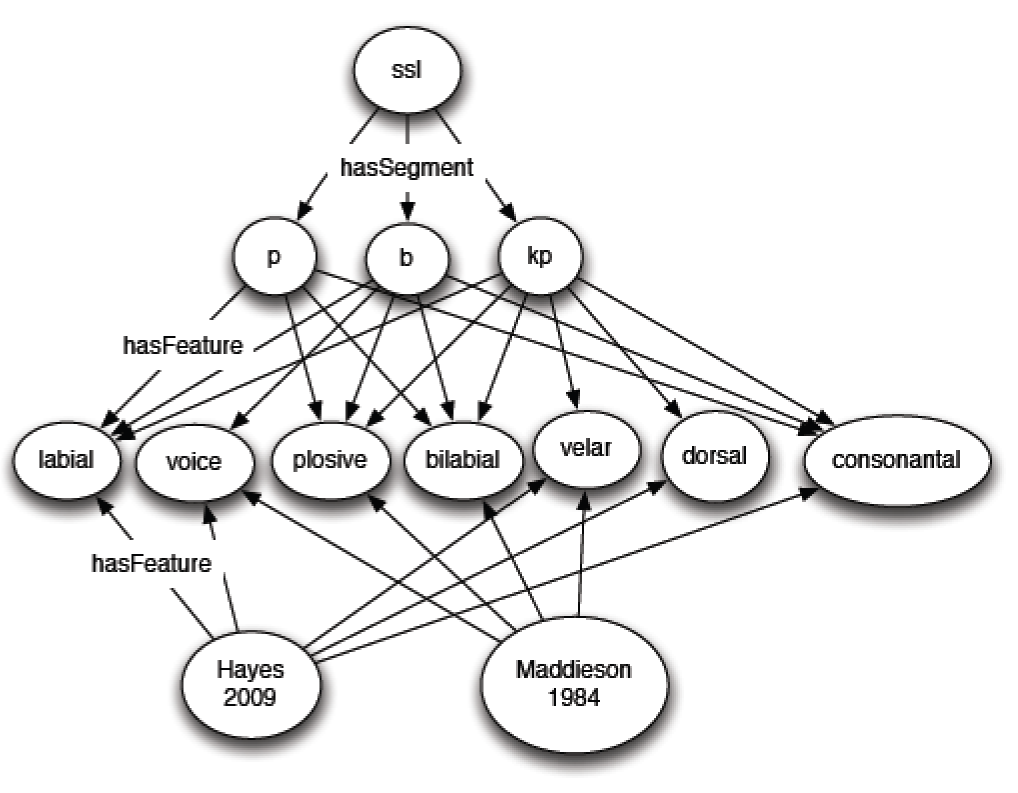
From Tables to RDF …



Source	id	ISO639-3	trump	root	wals_genus	population	latitude	longitude	phoneme_id	glyph_id	glyph	class	comb	num
SPA	1	[illegible]	1	asia	Korean	[illegible]	37:30	128:0	1	1	$t\int^h$	cons	c-d-c-c	4
SPA	3	[illegible]	1	ncau	Lak-Dargwa	[illegible]	42:0	47:0	124	1	[illegible]	cons	c-d-c-c	4
SPA	5	[illegible]	1	kart	Kartvelian	[illegible]	42:0	44:0	203	1	[illegible]	cons	c-d-c-c	4
SPA	6	[illegible]	1	asis	Burushaski	87,000	36:30	74:30	240	1	$t\int^h$	cons	c-d-c-c	4
SPA	14	[illegible]	1	ausa	Khmer	[illegible]	12:30	105:0	632	19	[illegible]	vowel	v-d	2
SPA	27	[illegible]	1	taik	Kam-Tai	[illegible]	15:00	100:40	1150	19	[illegible]	vowel	v-d	2

1. Decompose tables into triples
2. Multiple triples constitute a graph

From Tables to RDF …



1. Decompose tables into triples
2. Multiple triples constitute a graph
3. A graph can aggregate triples from other sources, as well

From Tables to RDF …

Graphs can be represented in other ways, **but** RDF allows us to

1. Provide explicit semantics (RDF Schema, Ontology)
2. Check consistency and infer implicit information
3. Merge (not only syntactically, but semantically)
4. Query
5. Link (enrich with external data)

From Tables to RDF …

Graphs can be represented in other ways, **but** RDF allows us to

1. Provide explicit semantics (RDF Schema, Ontology)
2. Check consistency and infer implicit information

RDFS, OWL

3. Merge (not only syntactically, but semantically)
4. Query
5. Link (enrich with external data)

From Tables to RDF …

Graphs can be represented in other ways, **but** RDF allows us to

1. Provide explicit semantics (RDF Schema, Ontology)
2. Check consistency and infer implicit information
3. Merge (not only syntactically, but semantically)
4. Query
5. Link (enrich with external data)

URIs & SPARQL

Uniform Resource Identifiers (URIs)

- Agree on a common vocabulary and names for entities
- **URIs** provide globally unique identifiers

“hasSegment”

string, not unambiguous

vs.

```
<http://mlode.nlp2rdf.org/resource/phoible/hasSegment>
```

URIs

vs.

```
@prefix phoible: <http://mlode.nlp2rdf.org/resource/phoible/>

... phoible:hasSegment ...
```

SPARQL

Merge data and query it using the W3C standard SPARQL (SPARQL Protocol and Query Language)

“the SQL of the Semantic Web”

```
SELECT DISTINCT ?language
WHERE {
        ?language phoible:hasSegment ?segment .
        ?segment phoible:hasFeature phoible:delayed_release
}
```

From Tables to RDF to Linked Data

- use URIs as names for things (1)
  - links to external URIs (links) allow us to retrieve more information from these sites
- **if** they can be resolved via HTTP (2)
- **and** provide information as RDF* (3)
- **and** they include links to other URIs (4)

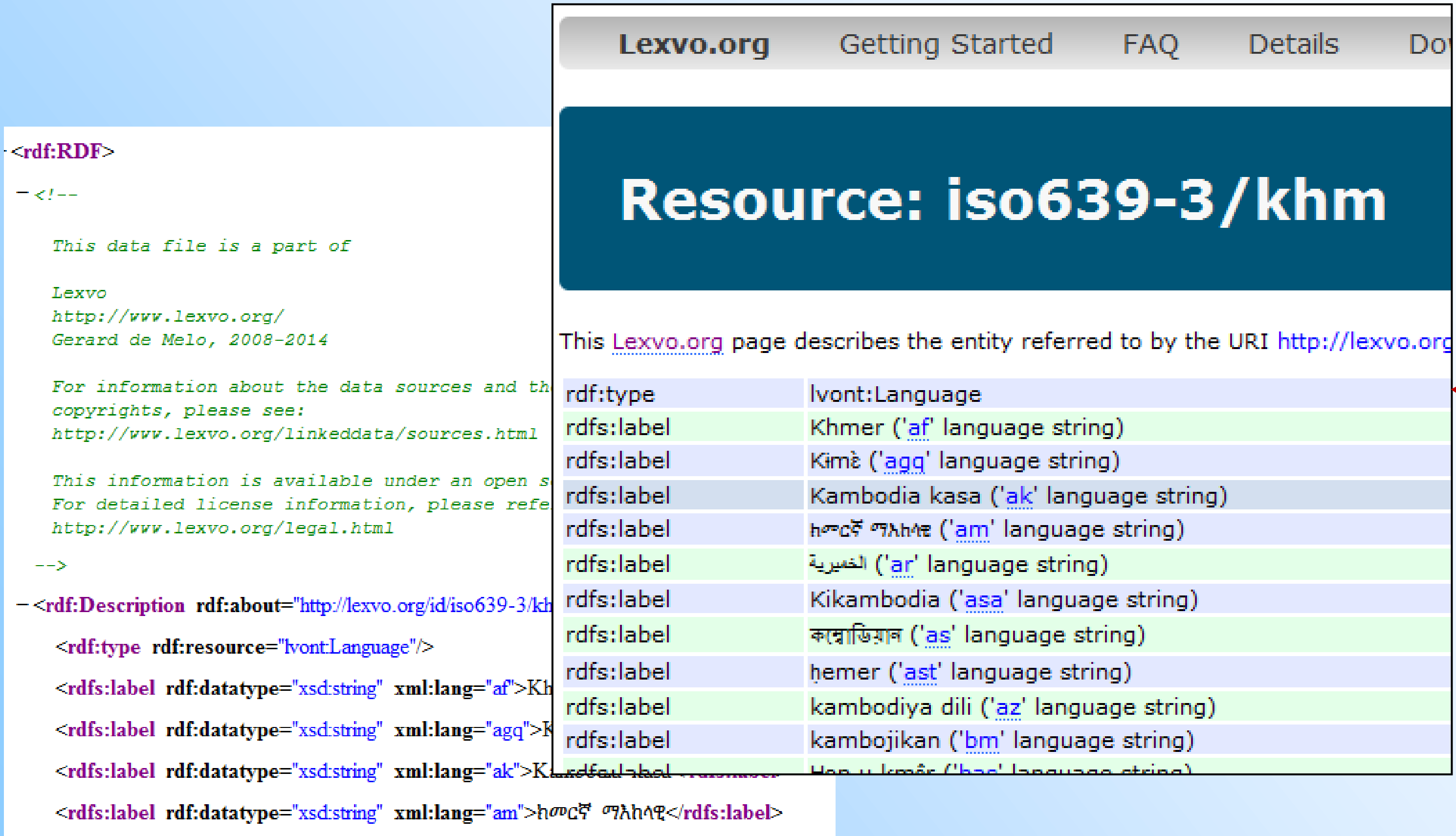
⇒ **then**, this is Linked Data (informally)

```
@prefix phoible: <http://mlode.nlp2rdf.org/resource/phoible/>
phoible:khm phoible:hasSegment "u:".
phoible:khm owl:sameAs <http://lexvo.org/id/iso639-3/khm>.
```

Turtle notation

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

From Tables to RDF to Linked Data



```
<rdf:RDF>
 - <!--

    This data file is a part of

    Lexvo
    http://www.lexvo.org/
    Gerard de Melo, 2008-2014

    For information about the data sources and th
    copyrights, please see:
    http://www.lexvo.org/linkeddata/sources.html

    This information is available under an open s
    For detailed license information, please refe
    http://www.lexvo.org/legal.html

   -->
 - <rdf:Description rdf:about="http://lexvo.org/id/iso639-3/kh
     <rdf:type rdf:resource="lvont:Language"/>
     <rdfs:label rdf:datatype="xsd:string" xml:lang="af">Kh
     <rdfs:label rdf:datatype="xsd:string" xml:lang="agq">K
     <rdfs:label rdf:datatype="xsd:string" xml:lang="ak">K
     <rdfs:label rdf:datatype="xsd:string" xml:lang="am">ክመርኛ ማእከላዊ</rdfs:label>
```

Resource: iso639-3/khm

This Lexvo.org page describes the entity referred to by the URI http://lexvo.org

rdf:type	lvont:Language
rdfs:label	Khmer ('af' language string)
rdfs:label	Kimɛ̀ ('agq' language string)
rdfs:label	Kambodia kasa ('ak' language string)
rdfs:label	ክመርኛ ማእከላዊ ('am' language string)
rdfs:label	الخميرية ('ar' language string)
rdfs:label	Kikambodia ('asa' language string)
rdfs:label	কম্বোডিয়ান ('as' language string)
rdfs:label	ḥemer ('ast' language string)
rdfs:label	kambodiya dili ('az' language string)
rdfs:label	kambojikan ('bm' language string)
rdfs:label	Hen u kmêr ('bas' language string)

```
@prefix phoible: <http://mlode.nlp2rdf.org/resource/phoible/>
phoible:khm phoible:hasSegment "u:".
phoible:khm owl:sameAs <http://lexvo.org/id/iso639-3/khm>.
```

Turtle notation

Linked **Open** Data: The 5 star plan

★ Make your data available on the Web under an open license

★★ Make it available as structured data

(Excel sheet instead of image scan of a table)

★★★ Use a non-proprietary format

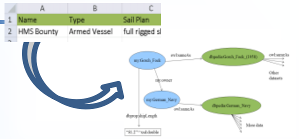
(CSV file instead of an Excel sheet)

★★★★ Use Linked Data format

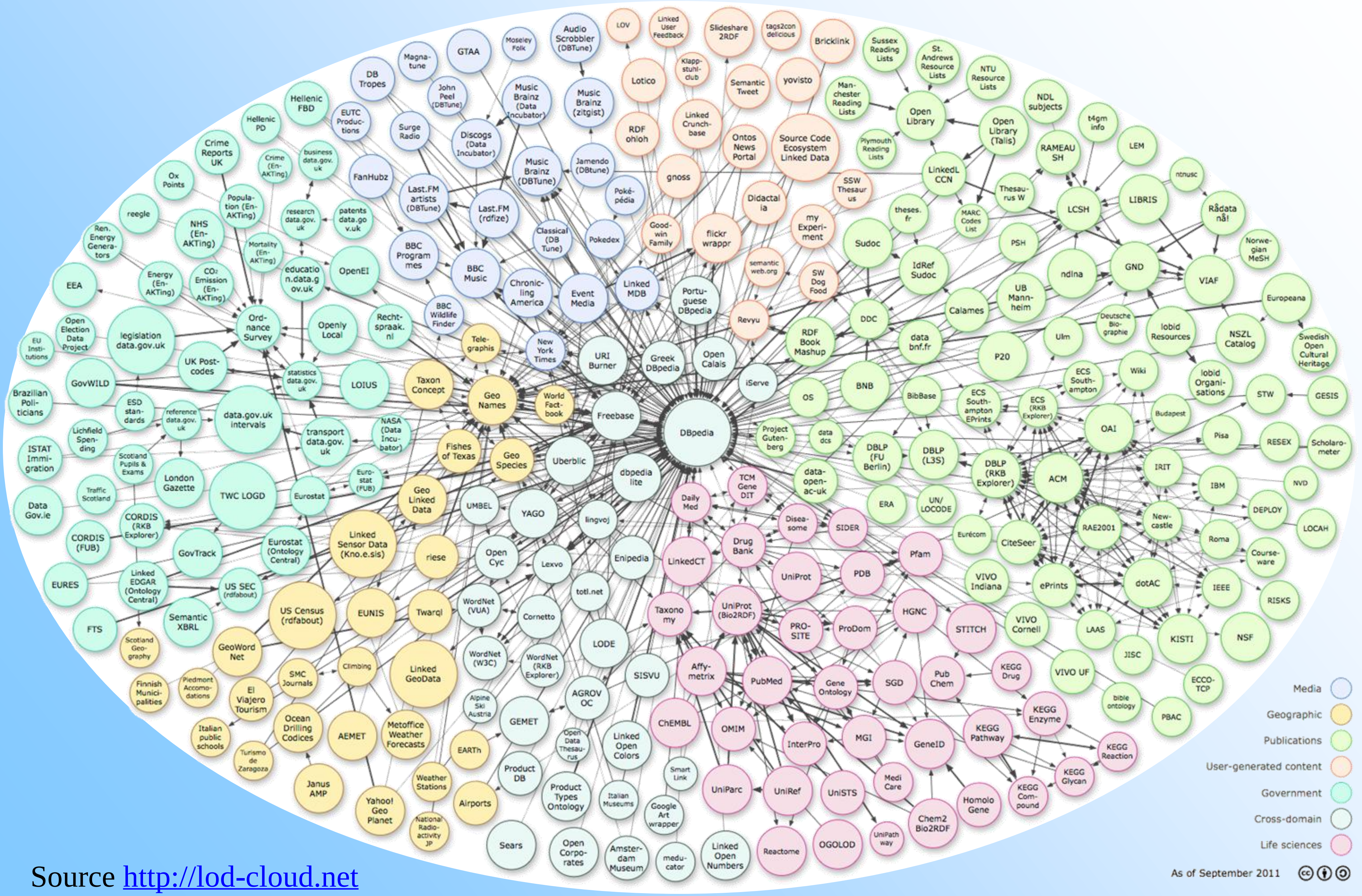
(URIs to identify things, RDF to represent data)

★★★★★ Link your data to other people's data to provide context

More: http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

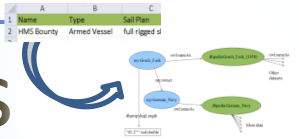


Linked Open Data cloud: Sep 2011

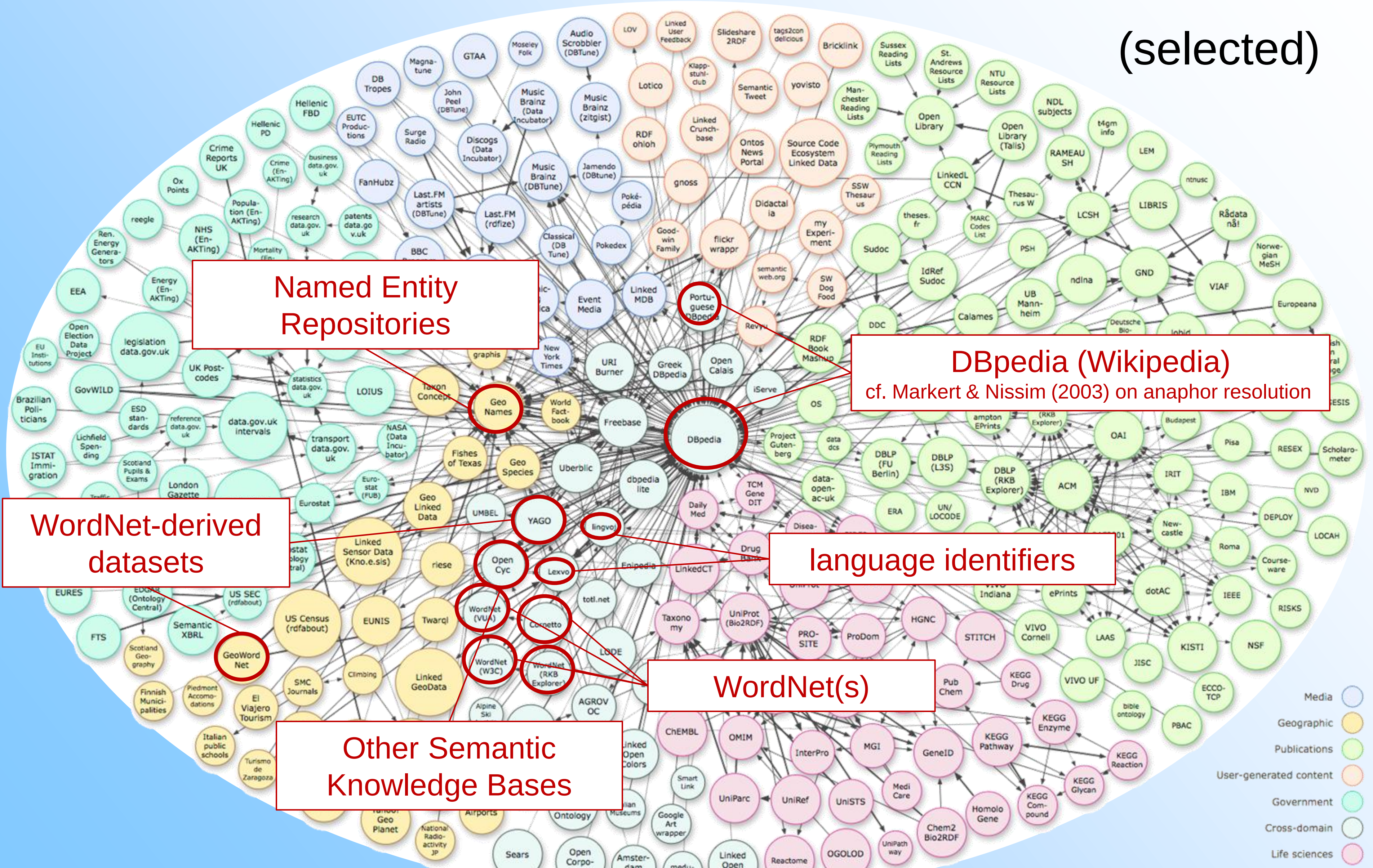


Source http://lod-cloud.net

Linguistically relevant LOD resources



(selected)



Named Entity Repositories

DBpedia (Wikipedia)
cf. Markert & Nissim (2003) on anaphor resolution

WordNet-derived datasets

language identifiers

WordNet(s)

Other Semantic Knowledge Bases

Source http://lod-cloud.net

Linked Data for Linguistics

Chiarcos, Littauer, Mendes, Moran & Nordhoff (2013)

Linked Data **for** Linguistics

- Representation and modelling
- Dynamic Import
- Structural interoperability
- Conceptual interoperability
- Federation
- Community and ecosystem

Linked Data **for** Linguistics

- Representation and modelling
- Dynamic Import
- Structural interoperability
- Conceptual interoperability
- Federation
- Community and ecosystem

Information Integration

- Structural interoperability
  - **same query language** for different data sets
- Conceptual interoperability
  - **same query** for different data sets
- Federation
  - **a single query** for different, distributed data sets

(simplified)

Community and Ecosystem

- RDF has been used in different contexts
  - Active community of users and developers
  - Rich technological infrastructure
  - Semantic Web: applied to **lexical** resources
- Also, it was applied to other linguistic resources
  - linguistic terminology (Farrar & Langendoen 2003)
  - corpora (Burchardt et al. 2005)
  - typological databases (Saulwick et al. 2005)

=> Linguistic Linked Open Data cloud (Chiarcos et al. 2012)

Linguistic Linked Open Data cloud

- a collection of linguistic resources
  - published under open licenses
  - as linked data
  - decentralized developed and maintained
  - meta data at http://datahub.io

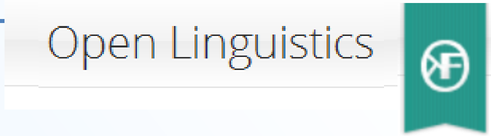
    => cloud diagram
  - developed as a community effort in the context of the Open Linguistics Working Group of the Open Knowledge Foundation

Open Knowledge Foundation (OKFN, http://okfn.org)

- non-profit organization
- founded in 2004
- promote open knowledge in all its forms
  - e.g., publication of government data (UK, US)
- provide infrastructural support for several working groups

OKFN Open Linguistics Working Group (OWLG)

- founded in Oct 2010 in Berlin, Germany
- open network of individuals interested in
  - linguistic resources and/or
  - their publication under open licenses
- multi-disciplinary
  - NLP/CL, typology/language documentation, IT, …
- infrastructure
  - mailing list, web site/blog, wiki
  - http://linguistics.okfn.org

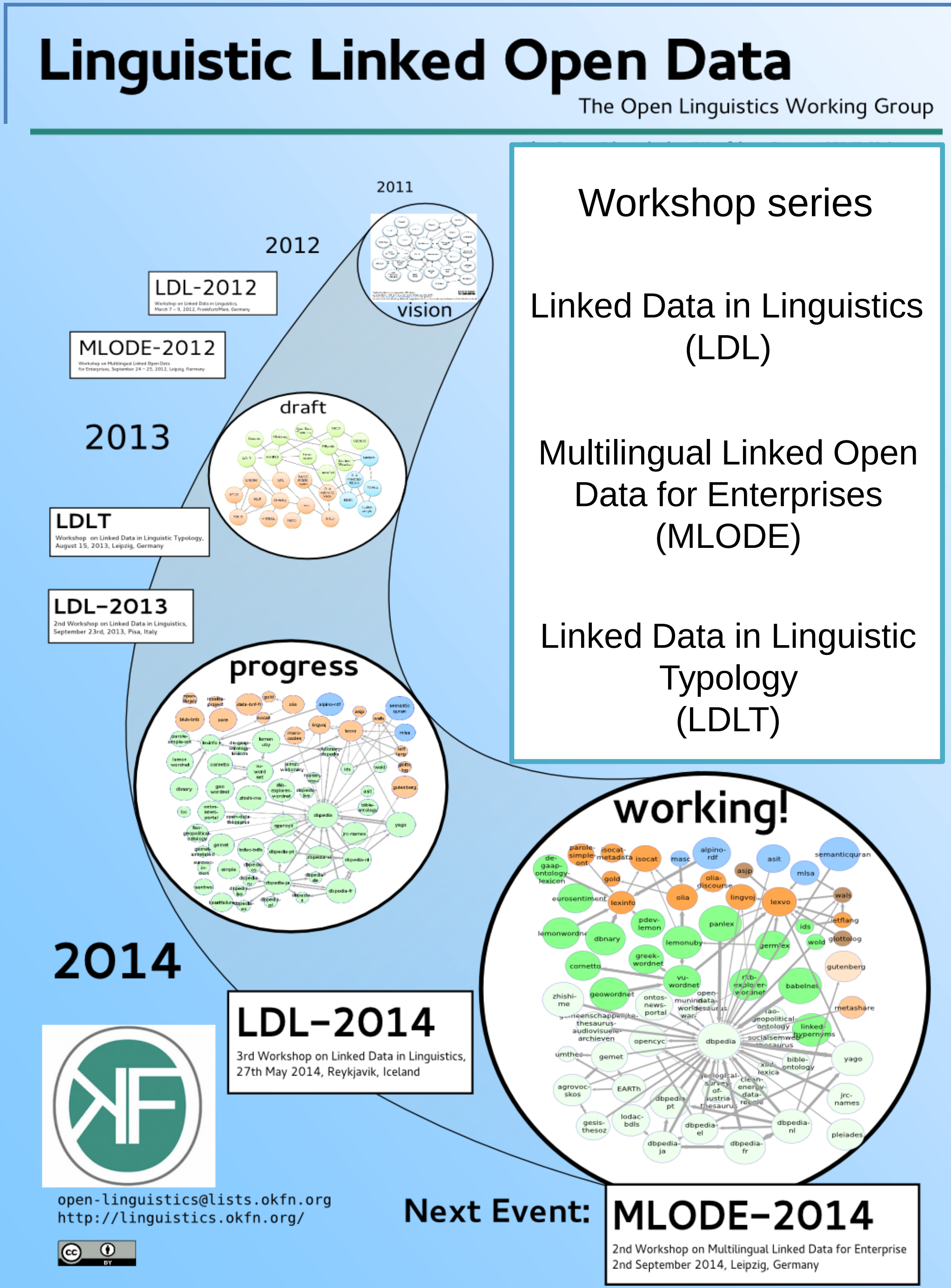


Open Linguistics

Important OWLG goals (http://linguistics.okfn.org)

1. **Promote open data** in relation to language data
2. **Facilitate communication** between researchers who use / distribute / maintain open linguistic data
3. **Mediate between providers and users** of technical infrastructures
4. Build and maintain an **index of open linguistic data sources**

OWLG activities



Linguistic Linked Open Data

The Open Linguistics Working Group

Workshop series

Linked Data in Linguistics (LDL)

Multilingual Linked Open Data for Enterprises (MLODE)

Linked Data in Linguistic Typology (LDLT)

2011

2012

LDL-2012
Workshop on Linked Data in Linguistics, March 7 – 9, 2012, Frankfurt/Main, Germany

MLODE-2012
Workshop on Multilingual Linked Open Data for Enterprises, September 24 – 25, 2012, Leipzig, Germany

2013

LDLT
Workshop on Linked Data in Linguistic Typology, August 15, 2013, Leipzig, Germany

LDL-2013
2nd Workshop on Linked Data in Linguistics, September 23rd, 2013, Pisa, Italy

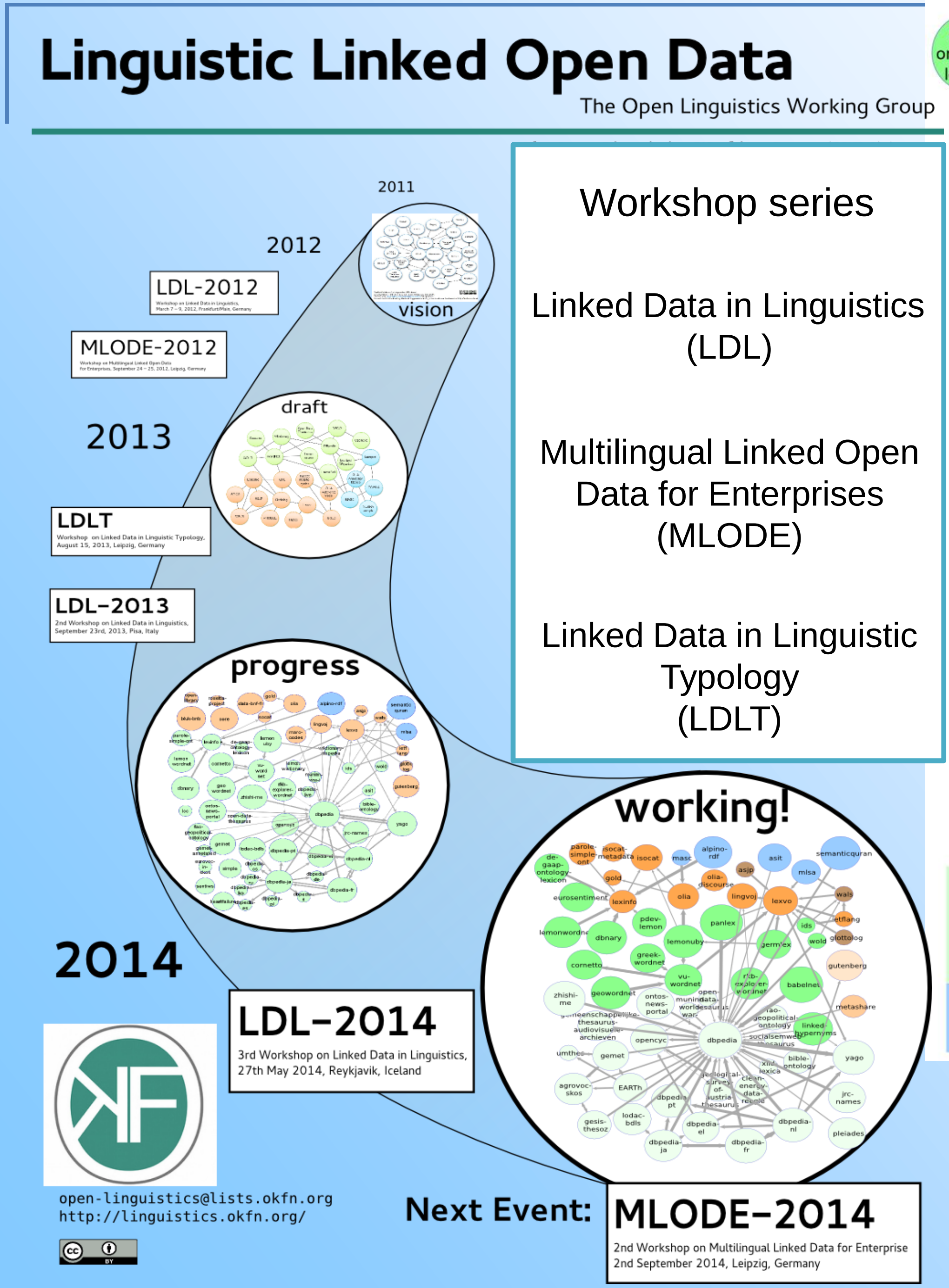
2014

LDL-2014
3rd Workshop on Linked Data in Linguistics, 27th May 2014, Reykjavik, Iceland

open-linguistics@lists.okfn.org
http://linguistics.okfn.org/

Next Event: **MLODE-2014**
2nd Workshop on Multilingual Linked Data for Enterprise
2nd September 2014, Leipzig, Germany

- point-to-point cooperations between individual members
- regular telcos/meetings
- workshops
- joint publications and presentations
- LLOD cloud development



Linguistic Linked Open Data

The Open Linguistics Working Group

2011

2012

LDL-2012
Workshop on Linked Data in Linguistics, March 7 – 9, 2012, Frankfurt/Main, Germany

MLODE-2012
Workshop on Multilingual Linked Open Data for Enterprises, September 24 – 25, 2012, Leipzig, Germany

2013

LDLT
Workshop on Linked Data in Linguistic Typology, August 15, 2013, Leipzig, Germany

LDL-2013
2nd Workshop on Linked Data in Linguistics, September 23rd, 2013, Pisa, Italy

2014

LDL-2014
3rd Workshop on Linked Data in Linguistics, 27th May 2014, Reykjavik, Iceland

Workshop series

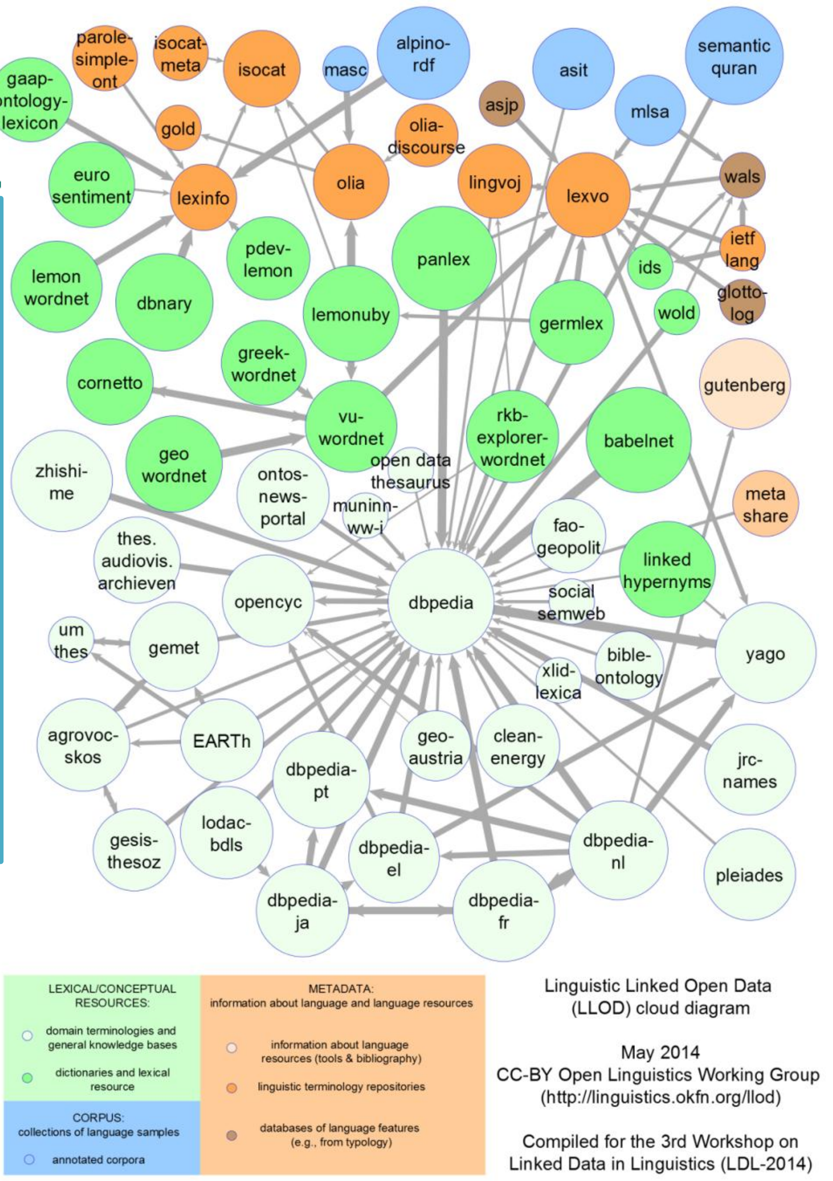
Linked Data in Linguistics (LDL)

Multilingual Linked Open Data for Enterprises (MLODE)

Linked Data in Linguistic Typology (LDLT)

open-linguistics@lists.okfn.org
http://linguistics.okfn.org/

Next Event: MLODE-2014
2nd Workshop on Multilingual Linked Data for Enterprise
2nd September 2014, Leipzig, Germany



Linguistic Linked Open Data (LLOD) cloud diagram

May 2014
CC-BY Open Linguistics Working Group
(http://linguistics.okfn.org/llod)

Compiled for the 3rd Workshop on Linked Data in Linguistics (LDL-2014)

new diagram, introduced tomorrow at LDL-2014

Building the Cloud: Examples

- Each data provider has different incentives to use Linked Data and/or RDF
- Concepts of RDF and Linked Data have been brought up to solve open problems in different subcommunities of linguistics and neighboring fields
- Examples
  - Corpora
  - Lexicons
  - Linguistic term and data bases

Building the Cloud: Examples

- Each data provider has different incentives to use Linked Data and/or RDF
- Concepts of RDF and Linked Data have been brought up to solve open problems in different subcommunities of linguistics and neighboring fields
- Examples
  - Corpora
  - Lexicons
  - Linguistic term and data bases

TODAY:
Underresourced
Languages

Case Studies

Linked Data for Underresourced Languages

Under-resourced Languages

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, …)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

Linked Data may

1. Improve conceptual and structural interoperability

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, …)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

Linked Data may

1. Improve conceptual and structural interoperability

1.a *between languages =>* Projection

- Lack of access to **languag**
  - General lack of language d
- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, …)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

Linked Data may

1. Improve conceptual and structural interoperability

2 Guide digitization efforts

- Lack of access to **langua**
  - General lack of language c
- Lack of access to **digital** language data
  - Standardized orthography & encoding (ASCII, KOI-8, SAMPA)
  - Web resources (Wikipedia, Wiktionary, …)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

Linked Data may

1. Improve conceptual and structural interoperability

2 Guide digitization efforts

3 (Partially) compensate the lack of lexical resources

- Lack of access to **language data**
  - General lack of language documentation, e.g., dictionaries
- Lack of access to **digital** la
  - Standardized orthography a
  - Web resources (Wikipedia, Wiktionary, …)
- Lack of **IT/NLP support**
  - Localized text processing software
  - Basic Language Resource Kit (http://www.blark.org/)
- Limited **interoperability** of data and tools
  - tools & annotations use different formats and conventions

Case Studies

1. Improve conceptual and structural interoperability

2 Guide digitization efforts

3 (Partially) compensate the lack of lexical resources

(A) Shared vocabularies

- lemon: lexicons
- lexvo, Glottolog: languages
- PHOIBLE: phonemes
- OLiA: annotations

(B) Link and query multiple dictionaries

- QHL, PanLex, GermLex, …
- Towards a Comparative-Lexicographical Workbench

Case Studies

1. Improve conceptual and structural interoperability

2 Guide digitization efforts

(A) Shared vocabularies

- lemon: lexicons
- lexvo, Glottolog: languages
- PHOIBLE: phonemes
- OLiA: annotations

3 (Partially) compensate the lack of lexical resources

Today

(B) Link and query multiple dictionaries

- QHL, PanLex, GermLex, …
- Towards a Comparative-Lexicographical Workbench

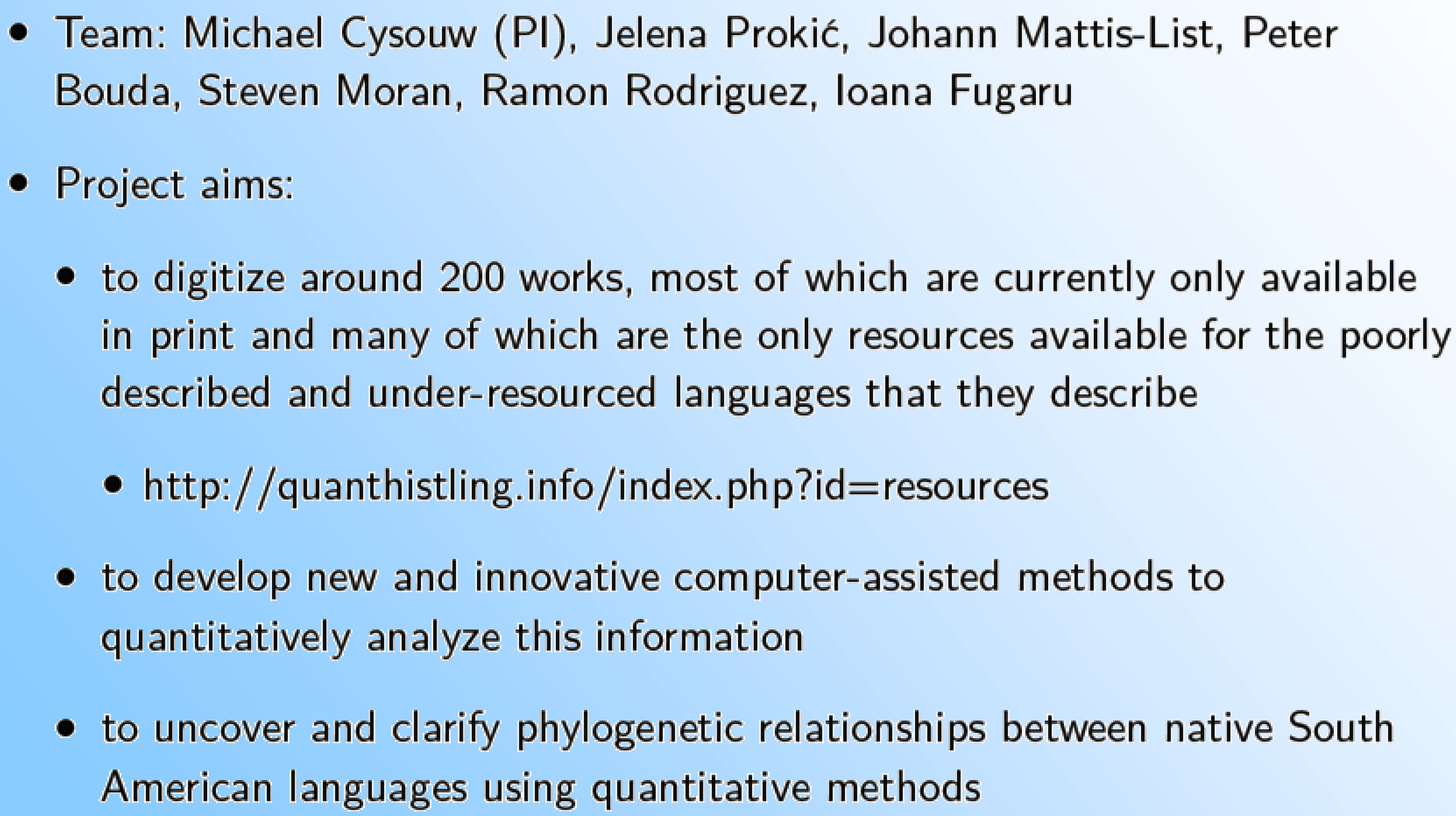
Case Studies

- Linking collections of dictionaries, e.g.,
  - PanLex (http://panlex.org/)
    - dictionaries for *all* languages in the world
  - QuantHistLing (http://quanthistling.info/)
    - South America
  - GermLex (http://datahub.io/dataset/germlex)
    - Germanic languages

Case Studies

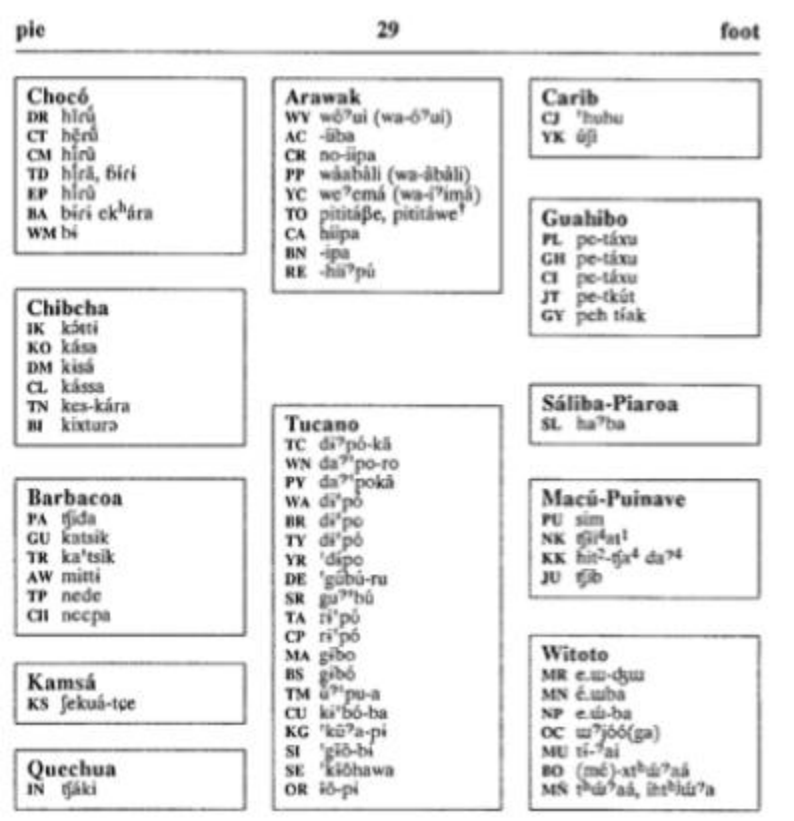
- Linking collections of dictionaries, e.g.,
  - PanLex (http://panlex.org/)
    - dictionaries for *all* languages in the world
  - QuantHistLing (http://quanthistling.info/)
    - South America (Moran and Brümmer 2013)
  - GermLex (http://datahub.io/dataset/germlex)
    - Germanic languages (tomorrow @ LDL-2014)

QuantHistLing



- Team: Michael Cysouw (PI), Jelena Prokić, Johann Mattis-List, Peter Bouda, Steven Moran, Ramon Rodriguez, Ioana Fugaru
- Project aims:
  - to digitize around 200 works, most of which are currently only available in print and many of which are the only resources available for the poorly described and under-resourced languages that they describe
    - http://quanthistling.info/index.php?id=resources
  - to develop new and innovative computer-assisted methods to quantitatively analyze this information
  - to uncover and clarify phylogenetic relationships between native South American languages using quantitative methods

QuantHistLing: Source Data



pie 29 foot

Chocó	
DR	hĩrũ
CT	hẽrũ
CM	hĩrũ
TD	hĩrã, ɓíri
EP	hĩrũ
BA	bíri ekʰára
WM	bɨ́

Chibcha	
IK	kɔ́tti
KO	kása
DM	kisá
CL	kássa
TN	kes-kára
BI	kixturə

Barbacoa	
PA	tʃida
GU	kɨtsik
TR	kaˈtsik
AW	mitti
TP	nede
CH	necpa

Kamsá	
KS	ʃekuá-tçe

Quechua	
IN	tʃáki

Arawak	
WY	wóʔui (wa-óʔui)
AC	-ĩiba
CR	no-ĩipa
PP	wáabàli (wa-ãbàli)
YC	weʔemá (wa-íʔimá)
TO	pititáβe, pititáwe†
CA	hĩipa
BN	-ipa
RE	-hĩiʔpú

Tucano	
TC	dɨʔpó-kã
WN	daʔˈpo-ro
PY	daʔˈpokã
WA	dɨˈpó
BR	dɨˈpo
TY	dɨˈpó
YR	ˈdɨpo
DE	ˈgũbũ-ru
SR	guʔˈbũ
TA	rɨˈpó
CP	rɨˈpó
MA	gɨbo
BS	gɨbó
TM	ũʔˈpu-a
CU	kɨˈbó-ba
KG	ˈkũʔa-pɨ
SI	ˈgĩõ-bɨ
SE	ˈkĩõhawa
OR	ĩõ-pɨ

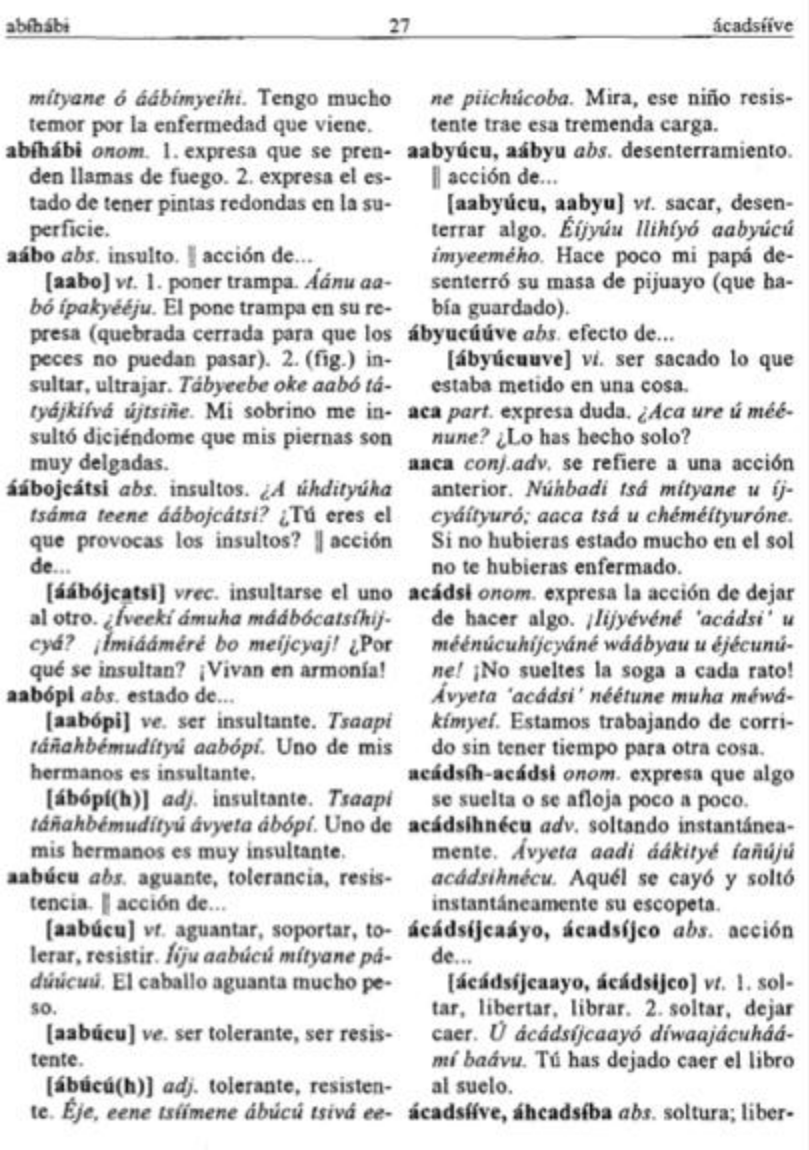
Carib	
CJ	ˈhuhu
YK	úʃi

Guahibo	
PL	pe-táxu
GH	pe-táxu
CI	pe-táxu
JT	pe-tkút
GY	peh tíak

Sáliba-Piaroa	
SL	haʔba

Macú-Puinave	
PU	sim
NK	tʃi^{4}at^1
KK	hit^2-tʃa^4 daʔ^4
JU	tʃɔ̃b

Witoto	
MR	e.ɯ-dʒɯ
MN	é.ɯba
NP	e.ɯ́-ba
OC	ɯʔjóó(ga)
MU	tí-ʔai
BO	(mé)-xtʰɯ́ʔaá
MÑ	tʰɯ́ʔaá, ìxtʰɯ́ɯ́ʔa



abíhábi 27 ácadsíive

mítyane ó áábimyeíhi. Tengo mucho temor por la enfermedad que viene.

abíhábi *onom.* 1. expresa que se prenden llamas de fuego. 2. expresa el estado de tener pintas redondas en la superficie.

aábo *abs.* insulto. ‖ acción de...

[aabo] *vt.* 1. poner trampa. *Áánu aabó ípakyééju.* El pone trampa en su represa (quebrada cerrada para que los peces no puedan pasar). 2. (fig.) insultar, ultrajar. *Tábyeebe oke aabó tátyájkiívá újtsiñe.* Mi sobrino me insultó diciéndome que mis piernas son muy delgadas.

áábojcátsi *abs.* insultos. *¿A úhdityúha tsáma teene áábojcátsi?* ¿Tú eres el que provocas los insultos? ‖ acción de...

[áábójcatsi] *vrec.* insultarse el uno al otro. *¿Íveekí ámuha mááb ócatsíhijcyá? ¡Ímiáámêrê bo meíjcyaj!* ¿Por qué se insultan? ¡Vivan en armonía!

aabópi *abs.* estado de...

[aabópi] *ve.* ser insultante. *Tsaapi tááñahbémudítyú aabópí.* Uno de mis hermanos es insultante.

[ábópí(h)] *adj.* insultante. *Tsaapi tááñahbémudítyú ávyeta ábópí.* Uno de mis hermanos es muy insultante.

aabúcu *abs.* aguante, tolerancia, resistencia. ‖ acción de...

[aabúcu] *vt.* aguantar, soportar, tolerar, resistir. *Íiju aabúcú mítyane pádúúcuú.* El caballo aguanta mucho peso.

[aabúcu] *ve.* ser tolerante, ser resistente.

[ábúcú(h)] *adj.* tolerante, resistente. *Éje, eene tsíímene ábúcú tsivá eene piichúcoba.* Mira, ese niño resistente trae esa tremenda carga.

aabyúcu, aábyu *abs.* desenterramiento. ‖ acción de...

[aabyúcu, aabyu] *vt.* sacar, desenterrar algo. *Éíjyúu llihíyó aabyúcú ímyeeméhо.* Hace poco mi papá desenterró su masa de pijuayo (que había guardado).

ábyucúúve *abs.* efecto de...

[ábyúcuuve] *vi.* ser sacado lo que estaba metido en una cosa.

aca *part.* expresa duda. *¿Aca ure ú méénune?* ¿Lo has hecho solo?

aaca *conj.adv.* se refiere a una acción anterior. *Núhbadi tsá mítyane u íjcyáítyuró; aaca tsá u chéméítyuróne.* Si no hubieras estado mucho en el sol no te hubieras enfermado.

acádsi *onom.* expresa la acción de dejar de hacer algo. *¡Iíjyévéné 'acádsi' u méénúcuhíjcyáné wáábyau u éjécunúne!* ¡No sueltes la soga a cada rato! *Ávyeta 'acádsi' néétune muha méwákímyeí.* Estamos trabajando de corrido sin tener tiempo para otra cosa.

acádsíh-acádsi *onom.* expresa que algo se suelta o se afloja poco a poco.

acádsihnécu *adv.* soltando instantáneamente. *Ávyeta aadi áákityé íañújú acádsihnécu.* Aquél se cayó y soltó instantáneamente su escopeta.

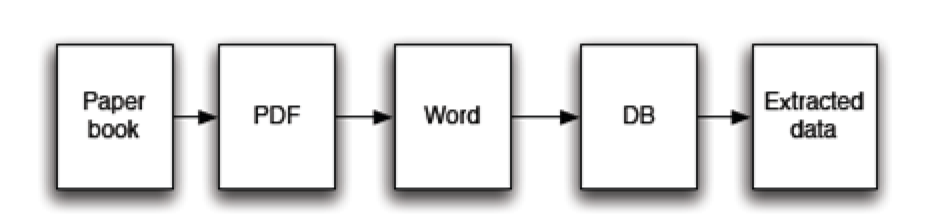
ácádsíjcaáyo, ácadsíjco *abs.* acción de...

[ácádsíjcaayo, ácádsijco] *vt.* 1. soltar, libertar, librar. 2. soltar, dejar caer. *Ú ácádsíjcaayó díwaajácuhááմí baávu.* Tú has dejado caer el libro al suelo.

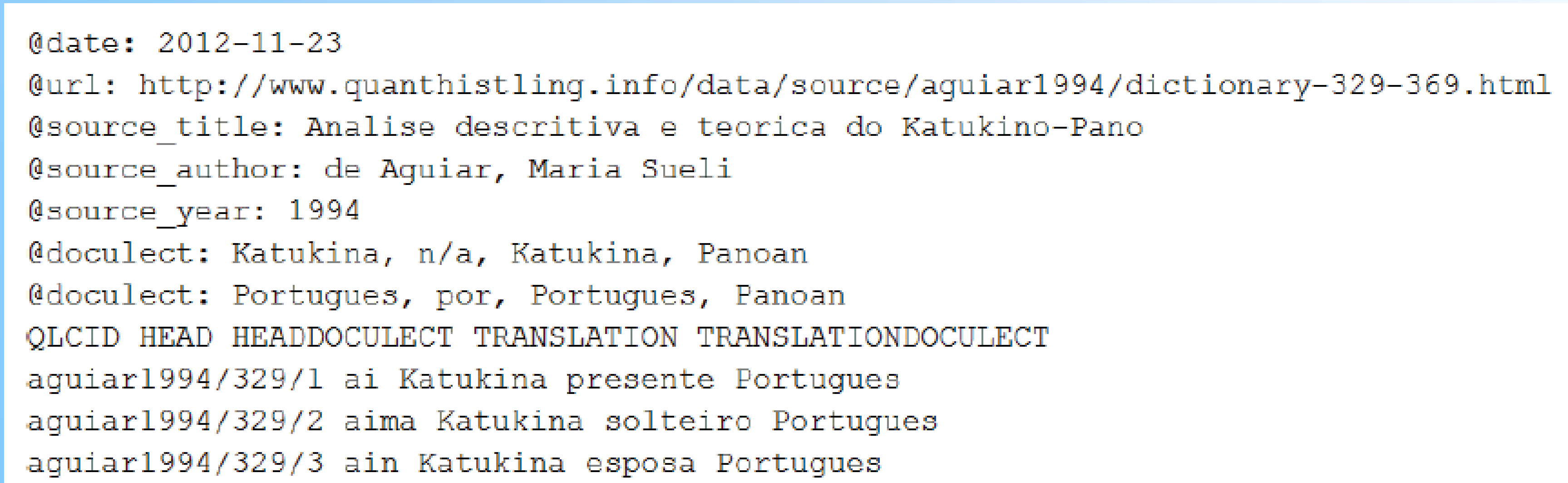
ácadsíive, áhcadsíba *abs.* soltura; liber-

QuantHistLing: Extraction

- Digitization pipeline (prepares the data for analysis)
    - http://quanthistling.info/data/
- We digitize the whole resource
- 80 dictionaries down, 120 to go...



- Simple data output format that contains metadata (prefixed with “@”) and tab-delimited lexical output



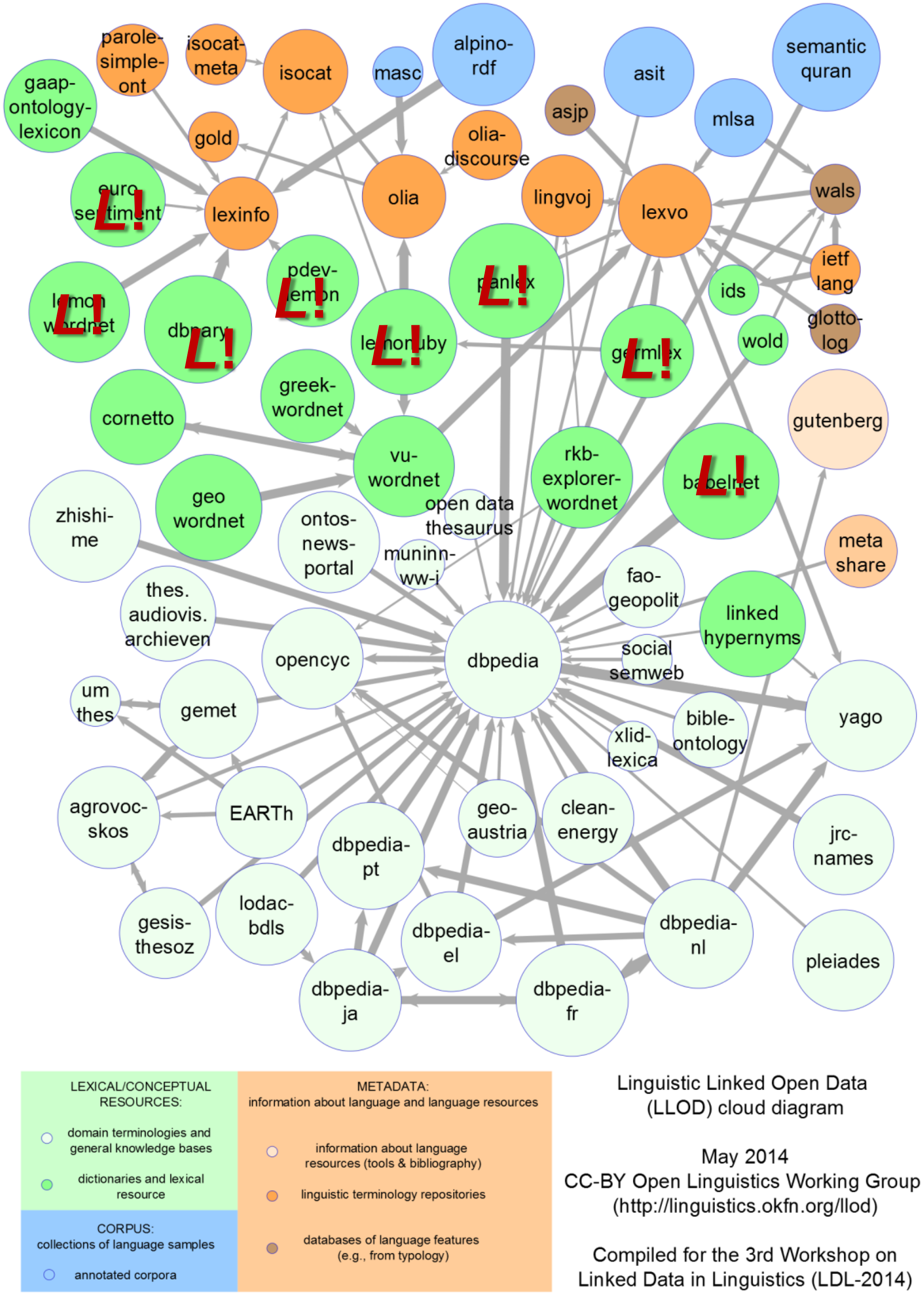
```
@date: 2012-11-23
@url: http://www.quanthistling.info/data/source/aguiar1994/dictionary-329-369.html
@source_title: Analise descritiva e teorica do Katukino-Pano
@source_author: de Aguiar, Maria Sueli
@source_year: 1994
@doculect: Katukina, n/a, Katukina, Panoan
@doculect: Portugues, por, Portugues, Panoan
QLCID HEAD HEADDOCULECT TRANSLATION TRANSLATIONDOCULECT
aguiar1994/329/1 ai Katukina presente Portugues
aguiar1994/329/2 aima Katukina solteiro Portugues
aguiar1994/329/3 ain Katukina esposa Portugues
```

QuantHistLing: From Data to Database using Linked Data and *lemon*

- We convert the QLC data into Linked Data that conforms to the Lemon model with a simple Python script
- Lemon is an ontological model for modeling lexicons and machine-readable dictionaries for linking to the Semantic Web and the Linked Data cloud
  - http://lemon-model.net/
- Lemon developers also active in the W3C Ontology-Lexica Community Group
  - Goal is to "develop models for the representation of lexica (and machine readable dictionaries) relative to ontologies"
  - http://www.w3.org/community/ontolex/

Why *lemon*:

(Relatively) widely used & actively maintained

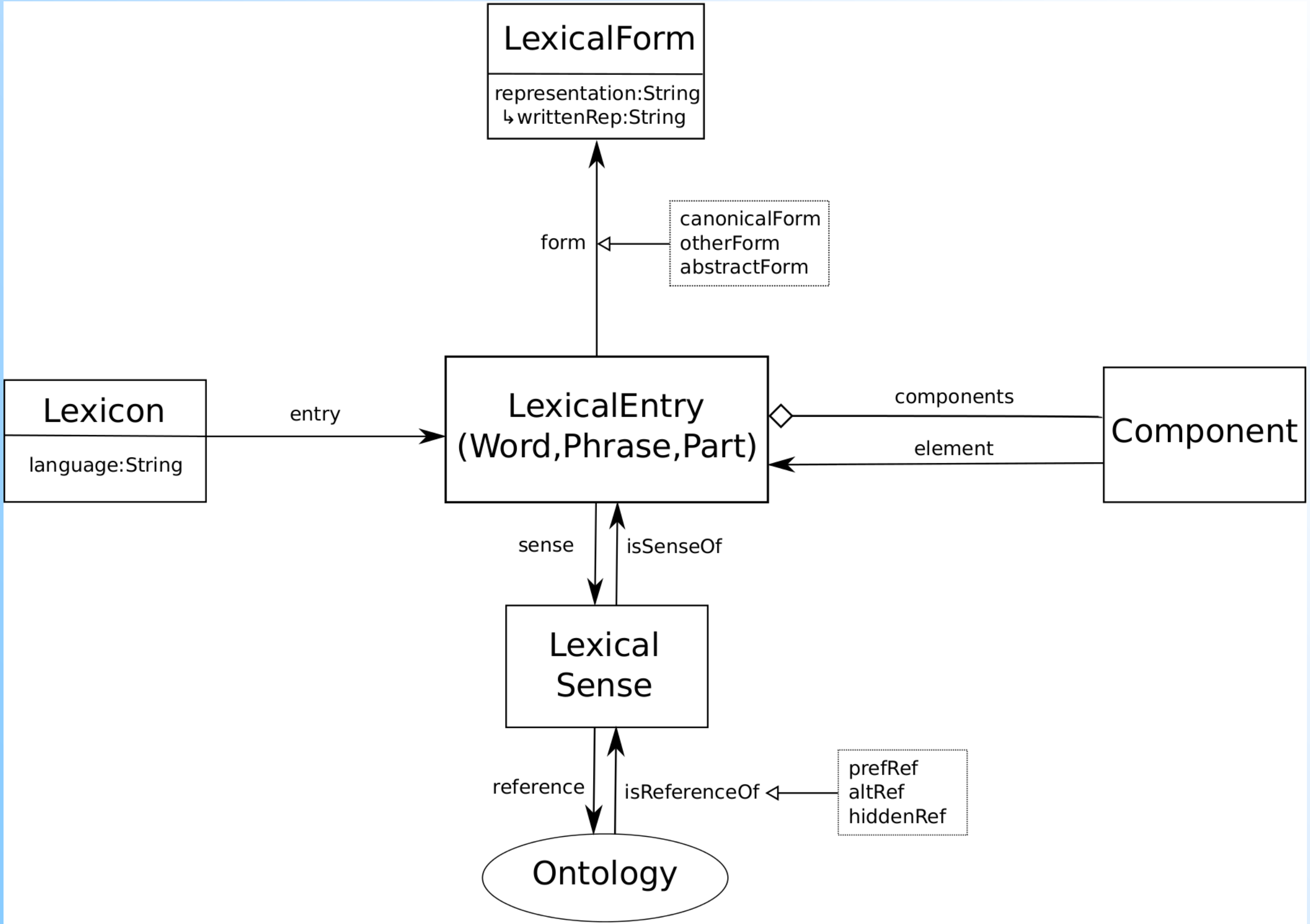


Linguistic Linked Open Data (LLOD) cloud diagram

May 2014
CC-BY Open Linguistics Working Group
(http://linguistics.okfn.org/llod)

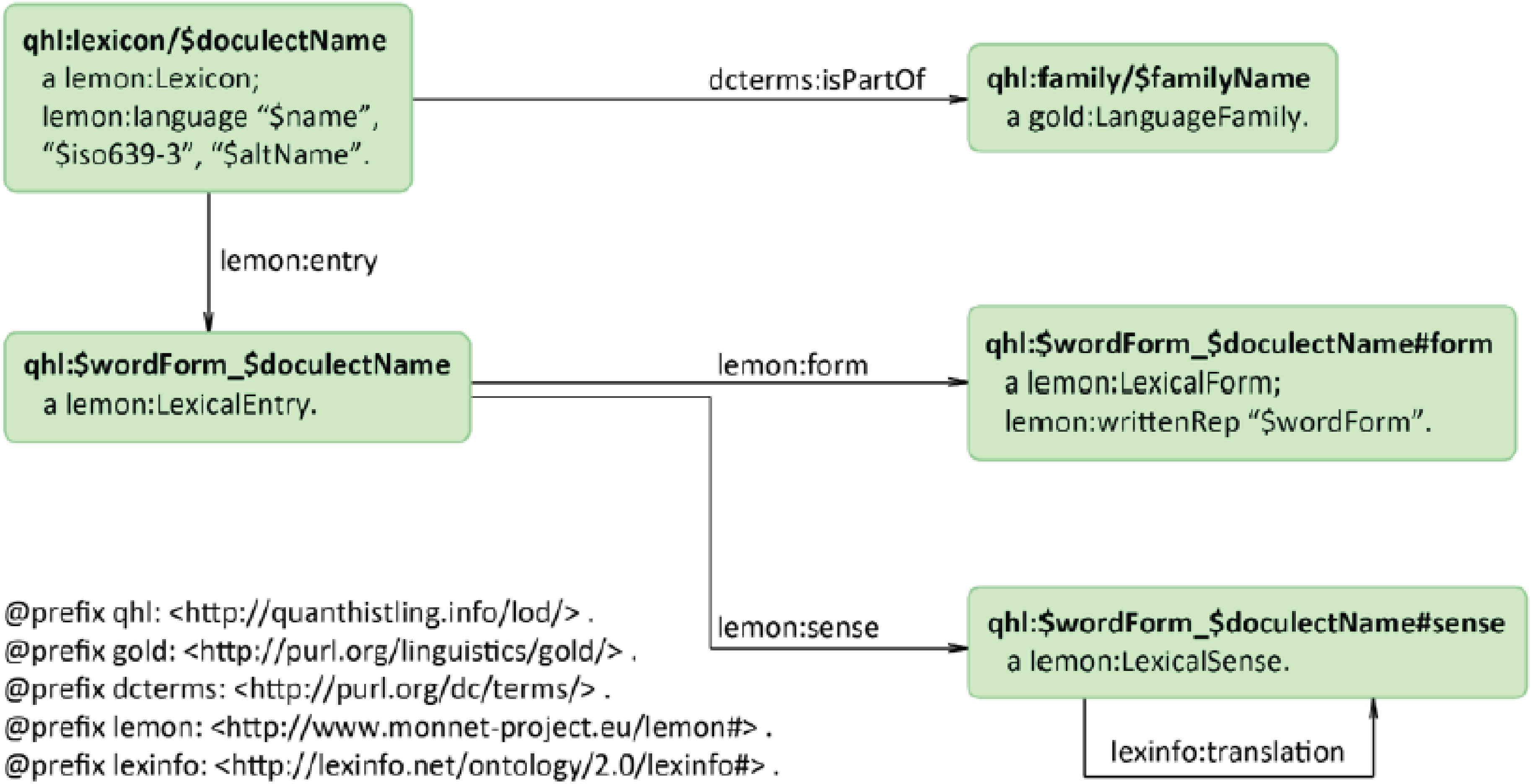
Compiled for the 3rd Workshop on Linked Data in Linguistics (LDL-2014)

lemon Core


LexicalForm
representation:String
↳writtenRep:String
form
canonicalForm
otherForm
abstractForm
Lexicon
language:String
entry
LexicalEntry
(Word,Phrase,Part)
components
element
Component
sense
isSenseOf
Lexical
Sense
reference
isReferenceOf
prefRef
altRef
hiddenRef
Ontology

QuantHistLing: *lemon* Sample

- We convert the QLC data into Linked Data that conforms to the Lemon model with a simple Python script



qhl:lexicon/$doculectName
a lemon:Lexicon;
lemon:language “$name”,
“$iso639-3”, “$altName”.

dcterms:isPartOf

qhl:family/$familyName
a gold:LanguageFamily.

lemon:entry

qhl:$wordForm_$doculectName
a lemon:LexicalEntry.

lemon:form

qhl:$wordForm_$doculectName#form
a lemon:LexicalForm;
lemon:writtenRep “$wordForm”.

lemon:sense

qhl:$wordForm_$doculectName#sense
a lemon:LexicalSense.

lexinfo:translation

```
@prefix qhl: <http://quanthistling.info/lod/> .
@prefix gold: <http://purl.org/linguistics/gold/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix lexinfo: <http://lexinfo.net/ontology/2.0/lexinfo#> .
```

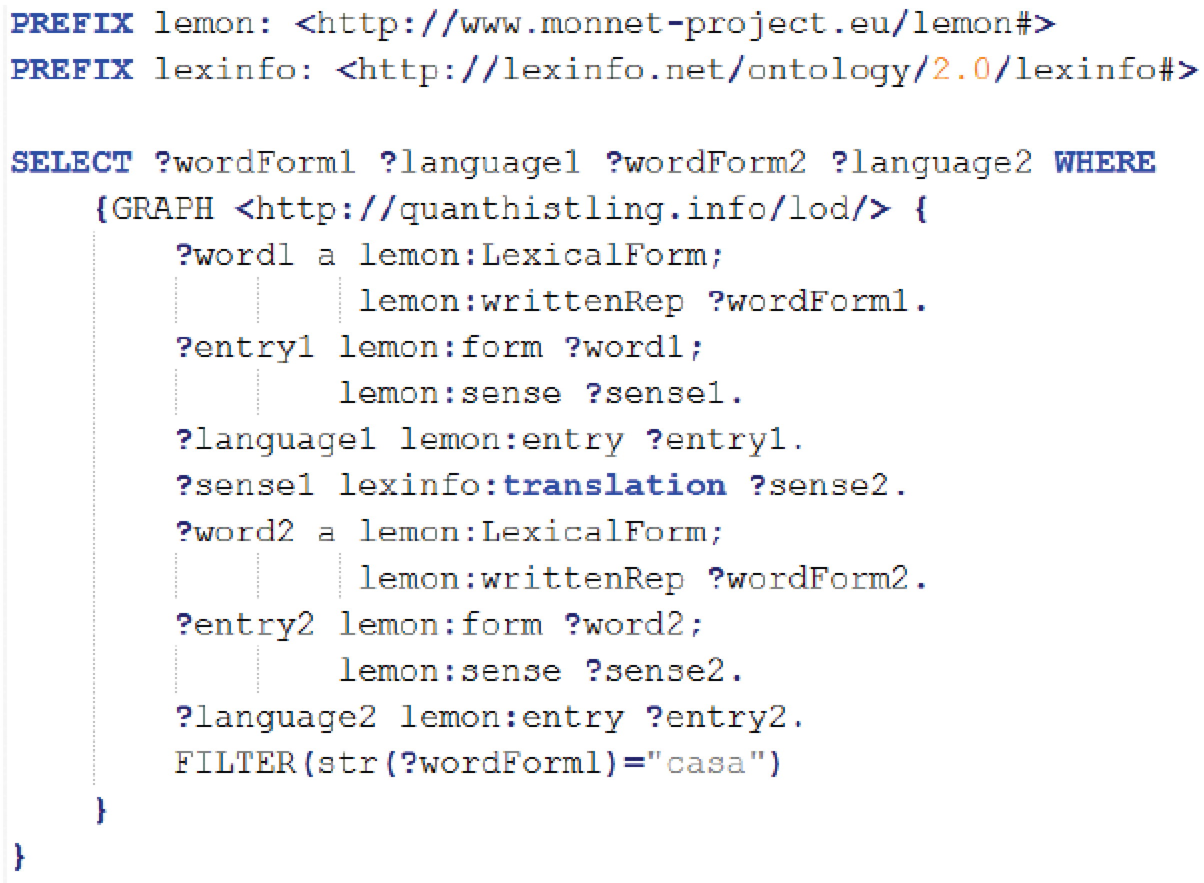
QuantHistLing: Search

- As a first step, we have converted the QHL data into RDF and it is available online through a SPARQL endpoint
  - http://linked-data.org/sparql/ (preliminary)
  - http://linked-data.org/datasets/ (data dump)
- Querying the combined dictionaries and lexicons is straightforward
  - Return all triples:

    -
      ```
      SELECT * WHERE
          {GRAPH <http://quanthistling.info/lod/>
              {?s ?p ?o}
      }
      ```
  - Returns over 3.8 million triples

QuantHistLing: Search

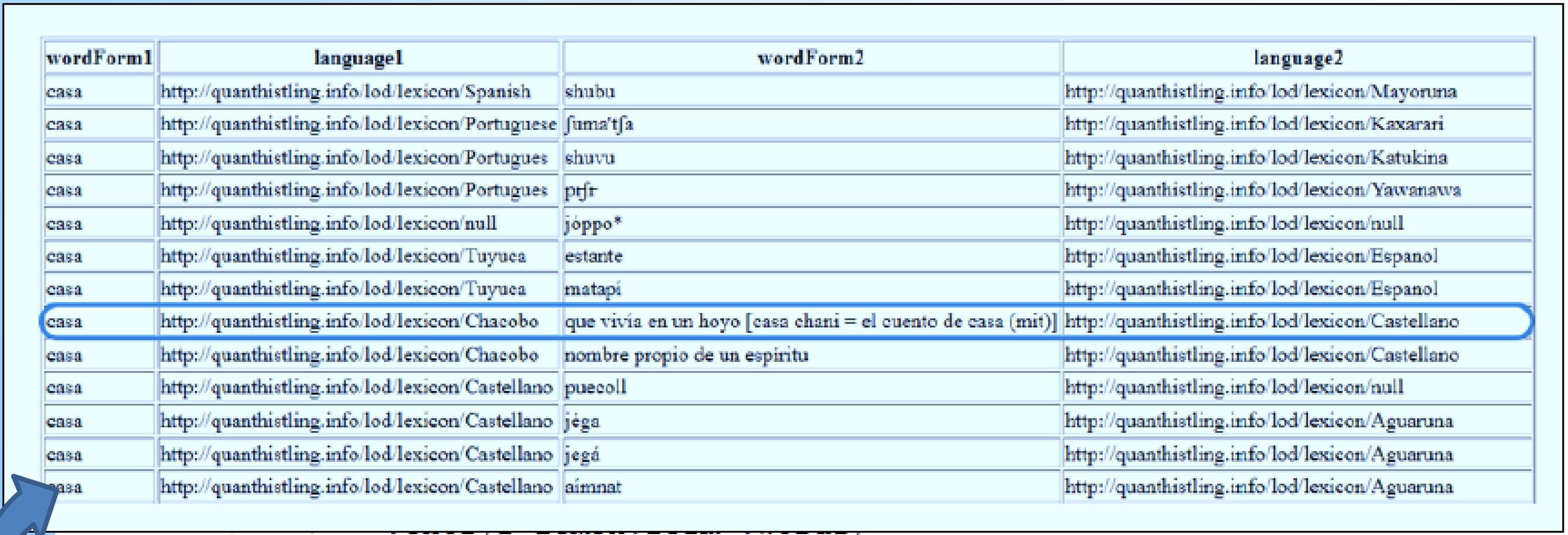
- Pairs of languages in the translation graph that contain written forms for the lexical sense “casa”



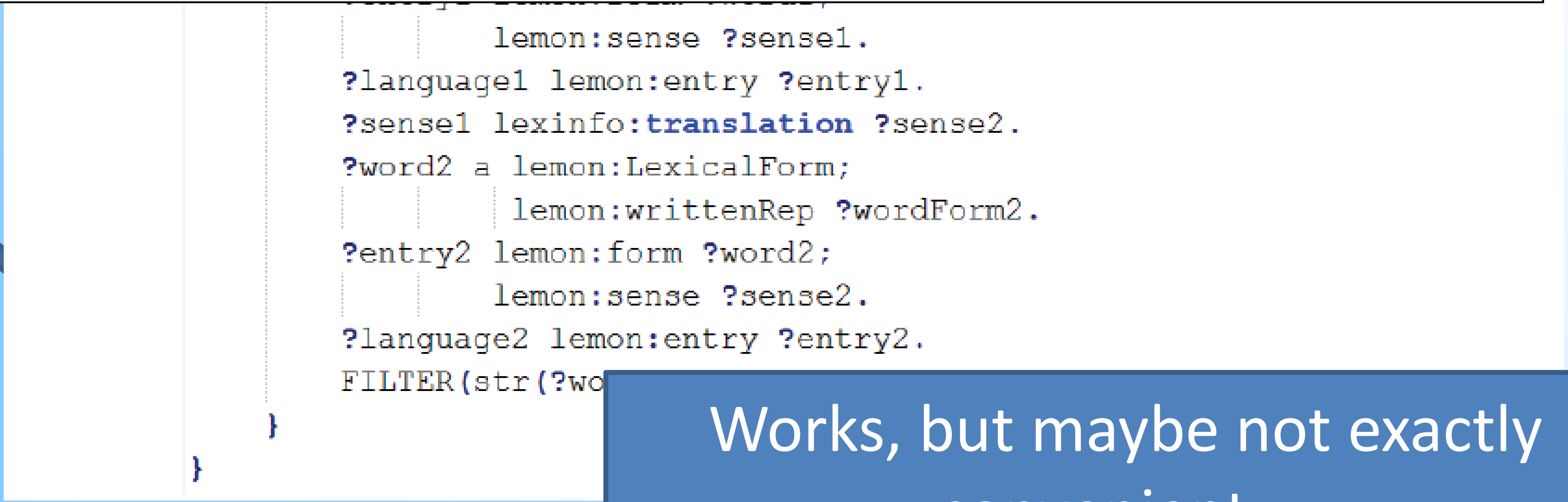
```
PREFIX lemon: <http://www.monnet-project.eu/lemon#>
PREFIX lexinfo: <http://lexinfo.net/ontology/2.0/lexinfo#>

SELECT ?wordForm1 ?language1 ?wordForm2 ?language2 WHERE
    {GRAPH <http://quanthistling.info/lod/> {
        ?word1 a lemon:LexicalForm;
                 lemon:writtenRep ?wordForm1.
        ?entry1 lemon:form ?word1;
                lemon:sense ?sense1.
        ?language1 lemon:entry ?entry1.
        ?sense1 lexinfo:translation ?sense2.
        ?word2 a lemon:LexicalForm;
                 lemon:writtenRep ?wordForm2.
        ?entry2 lemon:form ?word2;
                lemon:sense ?sense2.
        ?language2 lemon:entry ?entry2.
        FILTER(str(?wordForm1)="casa")
    }
}
```

QuantHistLing: Search



wordForm1	language1	wordForm2	language2
casa	http://quanthistling.info/lod/lexicon/Spanish	shubu	http://quanthistling.info/lod/lexicon/Mayoruna
casa	http://quanthistling.info/lod/lexicon/Portuguese	ʃuma'tʃa	http://quanthistling.info/lod/lexicon/Kaxarari
casa	http://quanthistling.info/lod/lexicon/Portugues	shuvu	http://quanthistling.info/lod/lexicon/Katukina
casa	http://quanthistling.info/lod/lexicon/Portugues	prʃr	http://quanthistling.info/lod/lexicon/Yawanawa
casa	http://quanthistling.info/lod/lexicon/null	jóppo*	http://quanthistling.info/lod/lexicon/null
casa	http://quanthistling.info/lod/lexicon/Tuyuca	estante	http://quanthistling.info/lod/lexicon/Espanol
casa	http://quanthistling.info/lod/lexicon/Tuyuca	matapí	http://quanthistling.info/lod/lexicon/Espanol
casa	http://quanthistling.info/lod/lexicon/Chacobo	que vivia en un hoyo [casa chani = el cuento de casa (mit)]	http://quanthistling.info/lod/lexicon/Castellano
casa	http://quanthistling.info/lod/lexicon/Chacobo	nombre propio de un espiritu	http://quanthistling.info/lod/lexicon/Castellano
casa	http://quanthistling.info/lod/lexicon/Castellano	puecoll	http://quanthistling.info/lod/lexicon/null
casa	http://quanthistling.info/lod/lexicon/Castellano	jéga	http://quanthistling.info/lod/lexicon/Aguaruna
casa	http://quanthistling.info/lod/lexicon/Castellano	jegá	http://quanthistling.info/lod/lexicon/Aguaruna
casa	http://quanthistling.info/lod/lexicon/Castellano	aímnat	http://quanthistling.info/lod/lexicon/Aguaruna



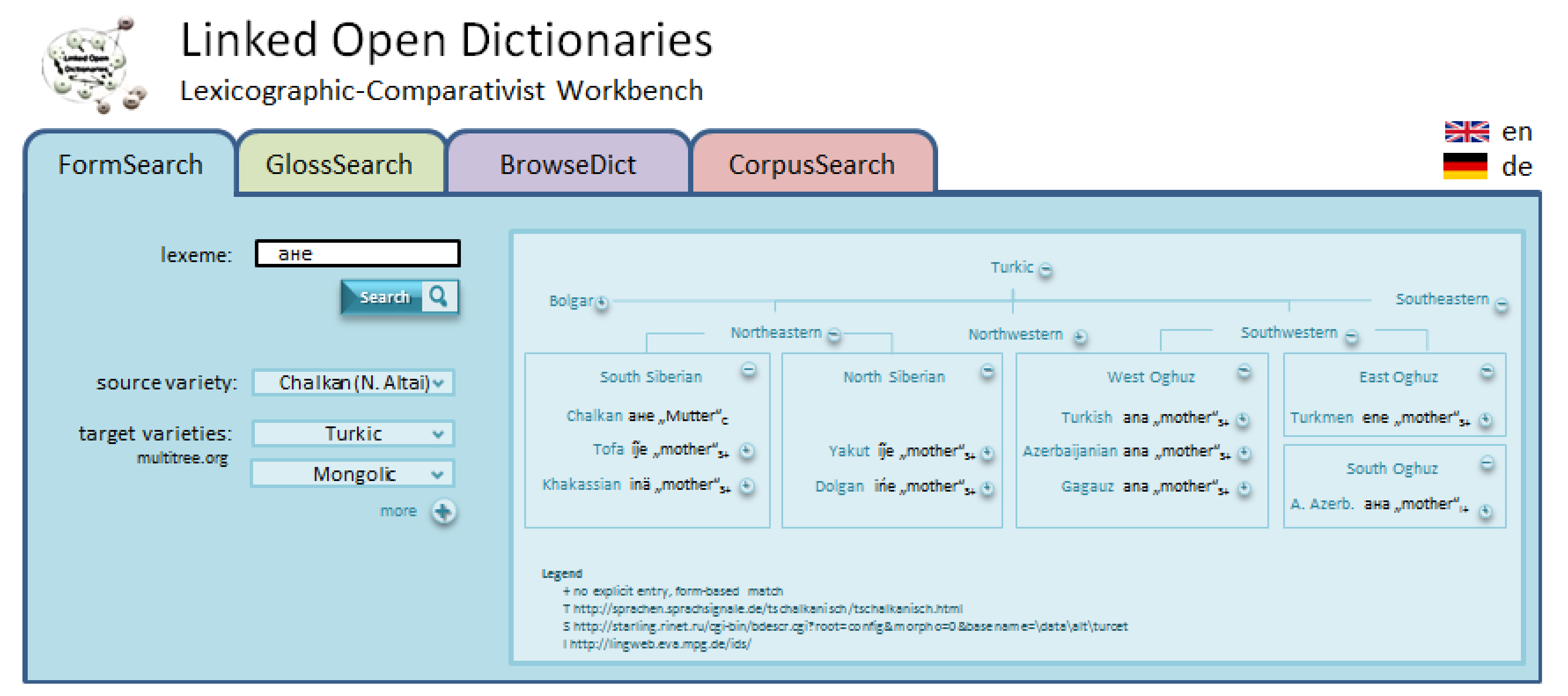
```
                lemon:sense ?sense1.
        ?language1 lemon:entry ?entry1.
        ?sense1 lexinfo:translation ?sense2.
        ?word2 a lemon:LexicalForm;
                 lemon:writtenRep ?wordForm2.
        ?entry2 lemon:form ?word2;
                lemon:sense ?sense2.
        ?language2 lemon:entry ?entry2.
        FILTER(str(?wo
    }
}
```

Works, but maybe not exactly convenient …

Linked Open Dictionaries (LiODi) Towards a Workbench

- Scenario: Language contact studies
  - query for a lexeme across multiple dictionaries
    - filter for source and target languages and language families
  - query across *diverse* resources available in the LLOD cloud
    - glosses to be linked to existing *lemon* resources, e.g., DBnary, WordNet
- Currently in preparation
  - Chiarcos, C. (in prep.), *Linked Open Dictionaries. Towards a Workbench for Comparative Lexicography*
  - Early implementation efforts in Frankfurt

Linked Open Dictionaries (LiODi) Towards a Workbench



Linked Open Dictionaries

Lexicographic-Comparativist Workbench

lexeme: ане

source variety: Chalkan (N. Altai)

target varieties: Turkic

multitree.org

Mongolic

Chalkan ане „Mutter“$_{C}$

Tofa ĭje „mother“$_{S+}$

Khakassian inä „mother“$_{S+}$

Yakut ĭje „mother“$_{S+}$

Dolgan ińe „mother“$_{S+}$

Turkish ana „mother“$_{S+}$

Azerbaijanian ana „mother“$_{S+}$

Gagauz ana „mother“$_{S+}$

Turkmen ene „mother“$_{S+}$

A. Azerb. ана „mother“$_{I+}$

Legend

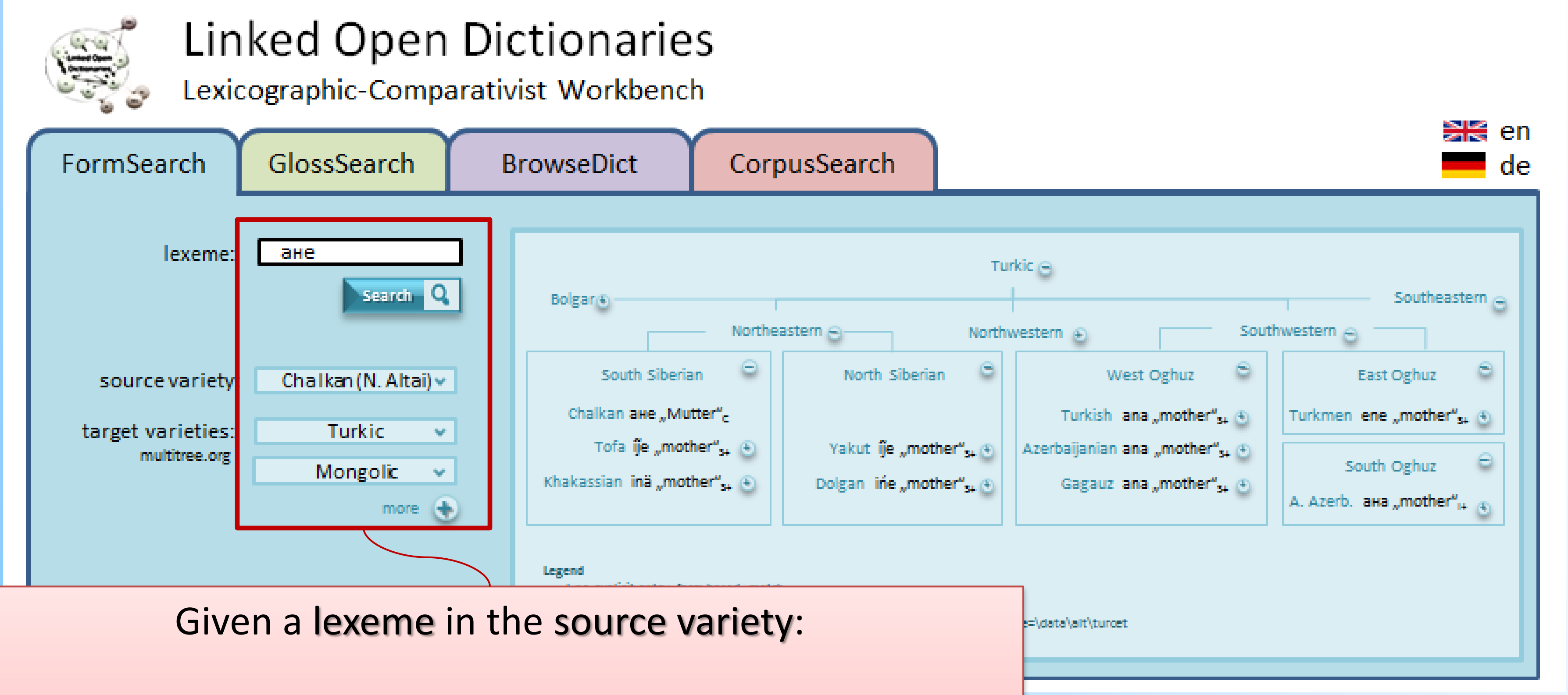
+ no explicit entry, form-based match

T http://sprachen.sprachsignale.de/tschalkanisch/tschalkanisch.html

S http://starling.rinet.ru/cgi-bin/bdescr.cgi?root=config&morpho=0&basename=\data\alt\turcet

I http://lingweb.eva.mpg.de/ids/

Linked Open Dictionaries (LiODi) Towards a Workbench



Linked Open Dictionaries

Lexicographic-Comparativist Workbench

lexeme: ане

source variety: Chalkan (N. Altai)

target varieties: Turkic, Mongolic

multitree.org

Chalkan ане „Mutter"$_{C}$

Tofa ĳe „mother"$_{S+}$

Khakassian inä „mother"$_{S+}$

Yakut ĳe „mother"$_{S+}$

Dolgan ińe „mother"$_{S+}$

Turkish ana „mother"$_{S+}$

Azerbaijanian ana „mother"$_{S+}$

Gagauz ana „mother"$_{S+}$

Turkmen ene „mother"$_{S+}$

A. Azerb. ана „mother"$_{I+}$

=\data\alt\turcet

Given a **lexeme** in the **source variety**:

Retrieve

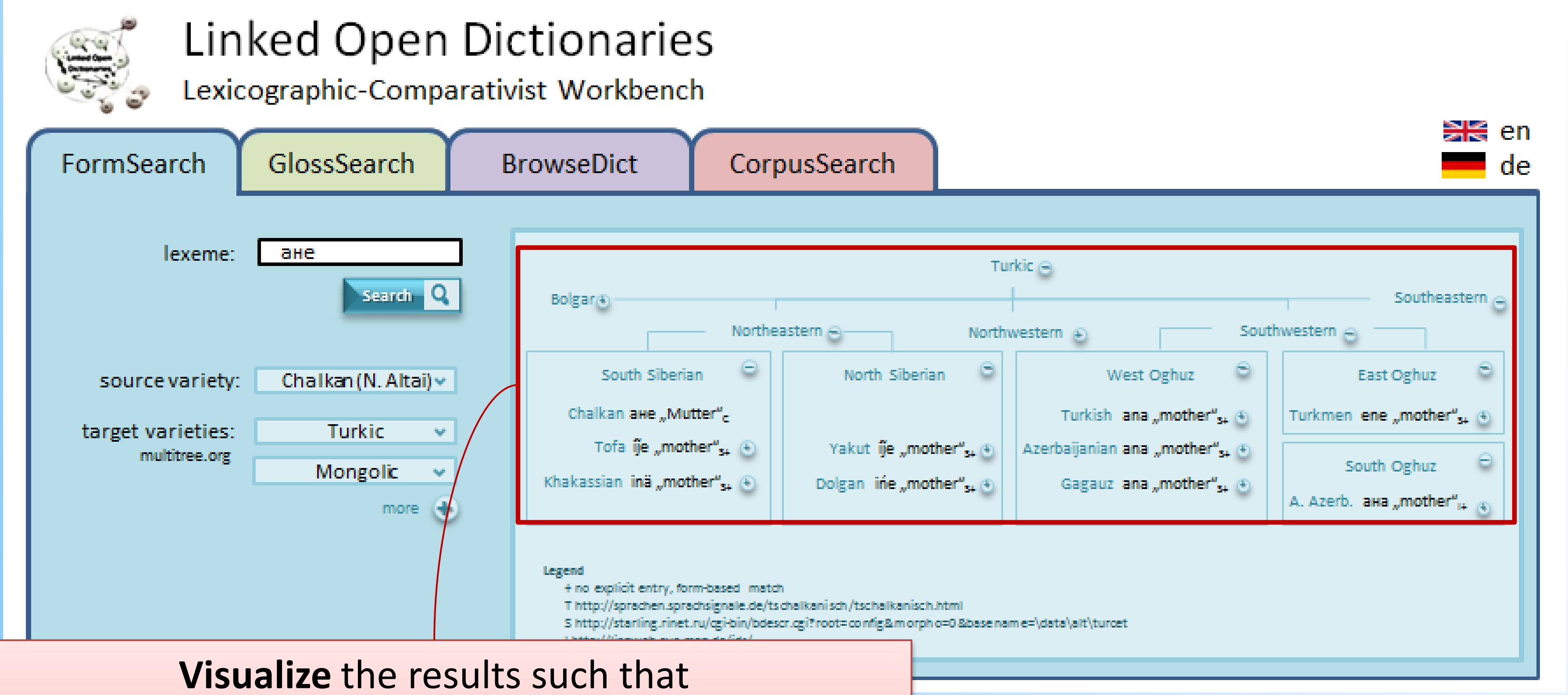
(a) all direct matches from the **target varieties**, and

(b) every other word from the target varieties that is either

(b.1) linked with a result from (a), or

(b.2) has the same gloss as a result from (a)

Linked Open Dictionaries (LiODi)
Towards a Workbench



Linked Open Dictionaries

Lexicographic-Comparativist Workbench

lexeme: ане

source variety: Chalkan (N. Altai)

target varieties: Turkic
multitree.org
Mongolic

Chalkan анe „Mutter“$_{C}$

Tofa i̯je „mother“$_{S+}$

Khakassian inä „mother“$_{S+}$

Yakut i̯je „mother“$_{S+}$

Dolgan ińe „mother“$_{S+}$

Turkish ana „mother“$_{S+}$

Azerbaijanian ana „mother“$_{S+}$

Gagauz ana „mother“$_{S+}$

Turkmen ene „mother“$_{S+}$

A. Azerb. ана „mother“$_{I+}$

Legend

+ no explicit entry, form-based match

T http://sprachen.sprachsignale.de/tschalkanisch/tschalkanisch.html

S http://starling.rinet.ru/cgi-bin/bdescr.cgi?root=config&morpho=0&basename=\data\alt\turcet

Visualize the results such that

(a) lemma and gloss are shown,

(b) matches are grouped according to some (externally provided) pylogenetic tree, and

(c) the path of dictionaries consulted is shown

What‘s in for underresourced languages ?

- Language documentation
  - Material collected on field trips is usually *afterwards* analysed, e.g., using annotation tools like ELAN or Toolbox
  - For the analysis of difficult words, it may not be possible to get in contact with native speakers
  - A distributional analysis of the word form and its meaning in related or neighboring varieties may help to disambiguate

  => partially compensates the lack of lexical resources

But wait!

- If a single query is to be applied on different resources, then relying on *lemon* is not enough
  - *lemon* provides data structures, **but**
    - for content and metadata, it relies on external vocabularies
- Interoperability depends on a *bundle* of vocabularies
  - WordNet, DBpedia, *any* ontology (lexical senses)
  - lexvo (language identifiers)
  - glottolog (languoid identifiers *from linguistic typology*)
  - PHOIBLE (phoneme inventories and phonological structures)
  - OLiA (annotations)
  - ISOcat (resource metadata)
  - GOLD (grammatical concepts)

Discussion

Problems and Questions

Summary

- Linked Data
  - General introduction
  - Benefits for linguist(ic)s
- Linguistic Linked Open Data
  - Community activities
- Use cases
  - Querying multiple dictionaries, filter and visualize by structured language metadata
    - Independently developed resources, shared vocabularies

Problems and Questions, and what to do about them

- RDF is misunderstood
  - RDF/XML is hard too read and process
  - As an alternative format, Turtle may be a compromise
- SPARQL is complicated
  - but not meant to be used by linguists in the field – it can nevertheless be used to develop tools for them
- Federation is a great concept, but causes too much traffic
  - Maintain your own sync‘ed copy of relevant external resources

Problems and Questions, and what to do about them

- *lemon* is neither developed for nor by linguists
  - but a vocabulary under development, so giving linguists a voice may be an option
- How can I publish my data as Linked Data ?
  - Ask, e.g, on the OWLG mailing list. Most likely, someone may help, and maybe, this will be a linguist, as well.
- Who could host my data?
  - That‘s a problem we can only solve as a community. If you write your next proposal, think of an end point for your data and help others to host (some of) their data.

Problems and Questions, and what to do about them

- How do I get into the LLOD cloud (diagram)?
  - Convert your data to RDF and put it under an open license
  - Create an entry at datahub.io
    - provide URL of a data dump or a SPARQL end point
  - Tag it as „linguistic“
  - Specify „triples“ and „links:xy“ (for datahub dataset xy).
  - Join the mailing list and wait for the next diagram generation announcement to make sure all went well.
  - Make sure your URLs are alive.

Problems and Questions, and what to do about them

- I encountered technical issues with datahub.io
  - Possible. It is not a perfect solution, and some colleagues are working on an alternative, but for the moment, we have to rely on it.
- Can I actually *do* anything with the LLOD cloud?
  - No, the diagram is merely a snapshot of the datahub.io metadata. It helps you to discover datasets and their dependencies.
  - But it tells you where to retrieve data dumps for local use or how to call SPARQL end points

Thank you !

Special thanks to

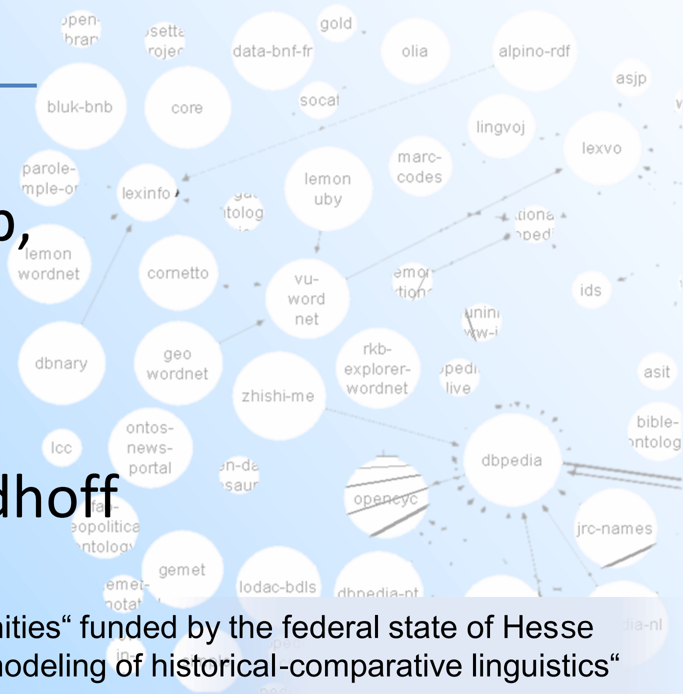
Laurette Pretorius & Claudia Soria,

The Open Linguistics Working Group,

Martin Brümmer, John McCrae,

Robert Forkel, Martin Haspelmath,

Sebastian Hellmann, Sebastian Nordhoff



Christian Chiarcos‘ work was partially supported by the LOEWE cluster „Digital Humanities“ funded by the federal state of Hesse
Steven Moran‘s work was supported by the ERC starting grant 240816 „Quantitative modeling of historical-comparative linguistics“

Sources

- Nontechnical Introduction
  - Chiarcos, C., Hellmann, S., Nordhoff S. (2012), *Introduction to Linked Data in Linguistics 2012*, presented at LDL-2012, March 2012, Frankfurt/M., Germany
- PHOIBLE example
  - Moran, S. 2012. Using Linked Data to Create a Typological Knowledge Base. In Chiarcos, C., Nordhoff, S., and Hellmann, S. (eds), *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.

Sources

- Linked Data for Linguistics
  - Chiarcos, C., Moran, S., Mendes, P., Nordhoff, S., Littauer, R. (2013). Building a Linked Open Data Cloud of Linguistic Resources. In Gurevych, I. and Kim, J. (eds), *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Springer, Berlin, Heidelberg.
- Case Studies: QuantHistLing
  - Moran, S., Brümmer, M. (2013), Lemon-aid: using Lemon to aid quantitative historical linguistic analysis. In Chiarcos et al. (eds.), *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013)*, Pisa, Italy, Sep 2013

Sources

- Case Studies: *lemon* Core Model
  - McCrae, J. (2014), *Ontology-lexica with lemon*, part of the LREC-2014 tutorial on Linked Data for Language Technologies (T10)
- Case Studies: Comp-Lex Workbench
  - Chiarcos, C. (in prep.), *Linked Open Dictionaries. Towards a Workbench for Comparative Lexicography*