Universität Bielefeld

AG Semantic Computing

Colloquium collocated with

Frederik Hogenboom's PhD defense

Talk on December 11th, 2014

»Ontology Lexicalization«

Prof. Dr. P. Cimiano

CITEC

Agenda

- Multilinguality on the Web
- Semantic Web & Linked Open Data
- Ontology Lexicalization
- Question Answering over Linked Data: the QALD Benchmark
- Conclusion

Agenda

- **Multilinguality on the Web**
- Semantic Web & Linked Open Data
- Ontology Lexicalization
- Question Answering over Linked Data: the QALD Benchmark
- Conclusion

What is the population density of Rotterdam?

What is the population density of Rotterdam?

German Wikipedia	2032 / km^2

What is the population density of Rotterdam?

German Wikipedia	2032 / km^2
English Wikipedia	2,969 / km^2

What is the population density of Rotterdam?

German Wikipedia	2032 / km^2
English Wikipedia	2,969 / km^2
Spanish Wikipedia	2,850 / km^2

Agenda

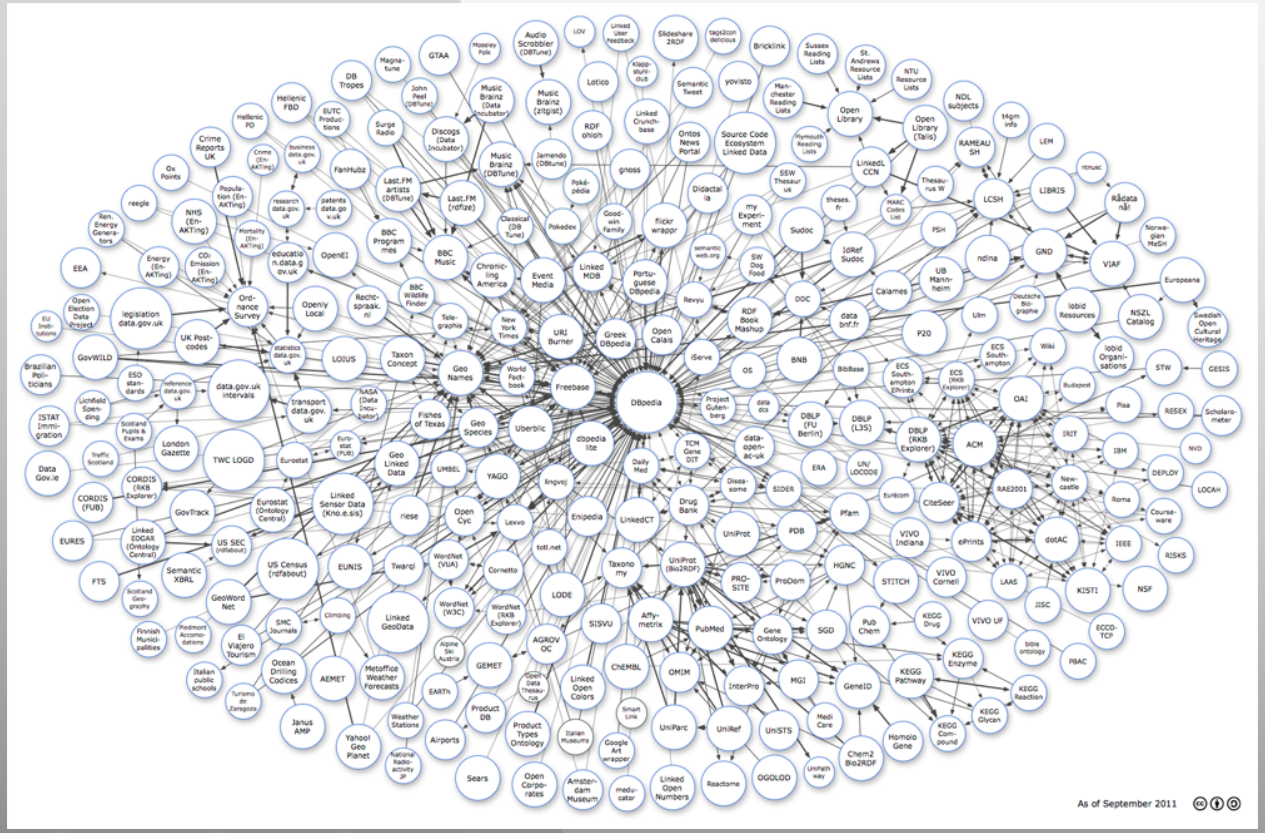
- Multilinguality on the Web
- **Semantic Web & Linked Open Data**
- Ontology Lexicalization
- Question Answering over Linked Data: the QALD Benchmark
- Conclusion

Semantic Web in a Nutshell

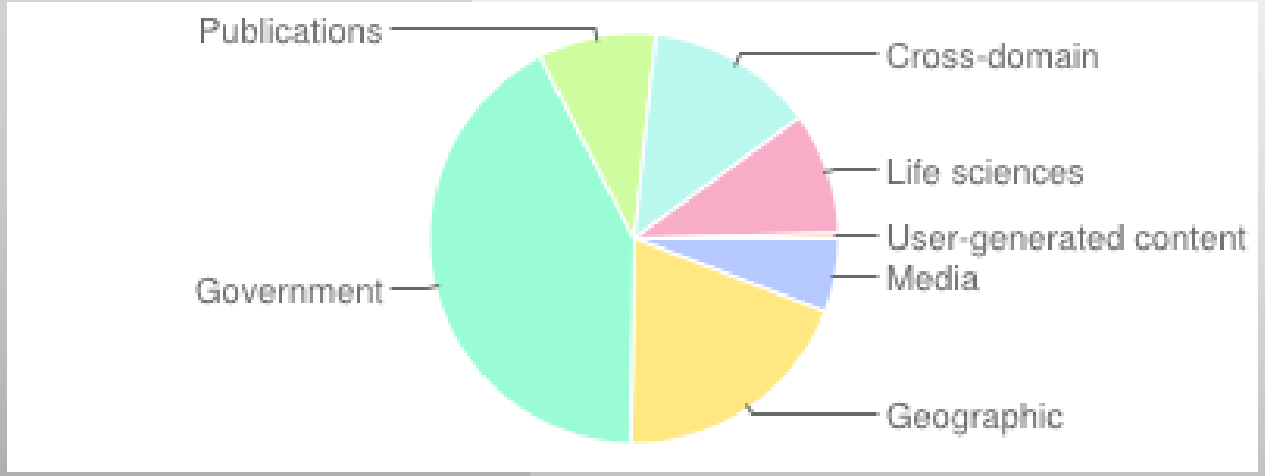
- Introduce **URIs for for all resources** described on the Web (not just documents, but people, events – e.g. a concert, music pieces, companies) – typically HTTP–URIs
- URIs are interpreted w.r.t. a model of the world
- **RDF as datamodel** to express factual knowledge in the form of triples (s, p, o) – serialized in XML or some other format
- so called **ontologies** define what the RDF means, i.e. to define the `meaning' of the symbols axiomatically
- **resources connected to each other through RDF triples,** relations uniquely identified through URIs

Linked Open Data: Size

More than 31 billion triples:



Linked Open Data: Domains



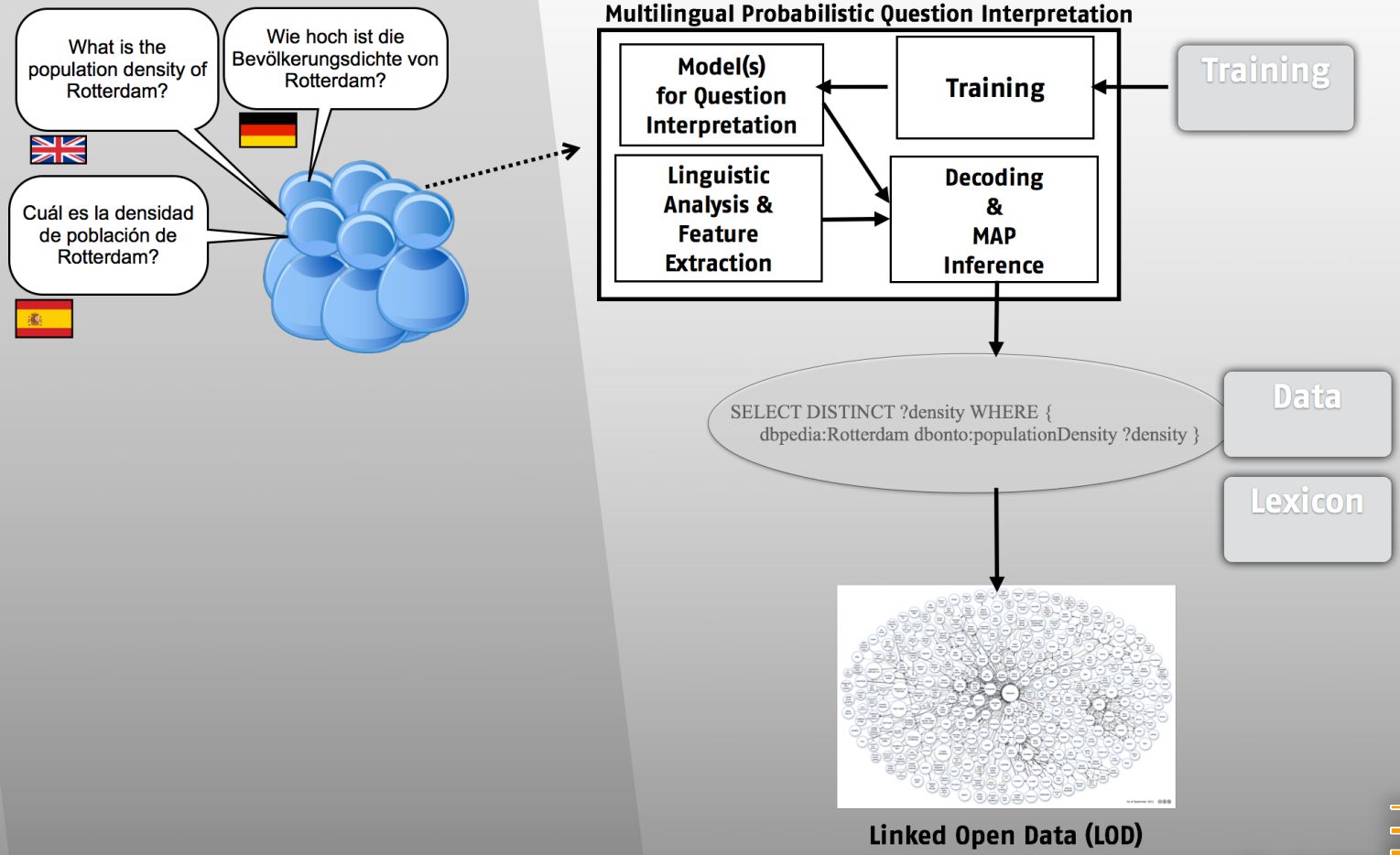
Semantic Web as an Interlingua?

As the linked open data cloud grows, it will bring together most of the factual knowledge that is relevant to humans, abstracting from specific human languages:

- URIs are language-independent, denote the actual entity in the real world
- facts are language-independent

Question: How will users that speak different languages access this body of language-independent knowledge?
Solution (partial): Semantic Web provides a normalizing, language-independent vocabulary!

Multilingual Question Answering



Multilingual Web Vision: Main Ingredients

- Language-independent Knowledge (Facts)
- Lexical Knowledge about how this knowledge is expressed in multiple languages:
  - Models / Vocabularies to express knowledge about lexical realization of classes, properties in particular languages
  - Methodologies for creating this knowledge
  - Techniques to ease the creation of this knowledge

Agenda

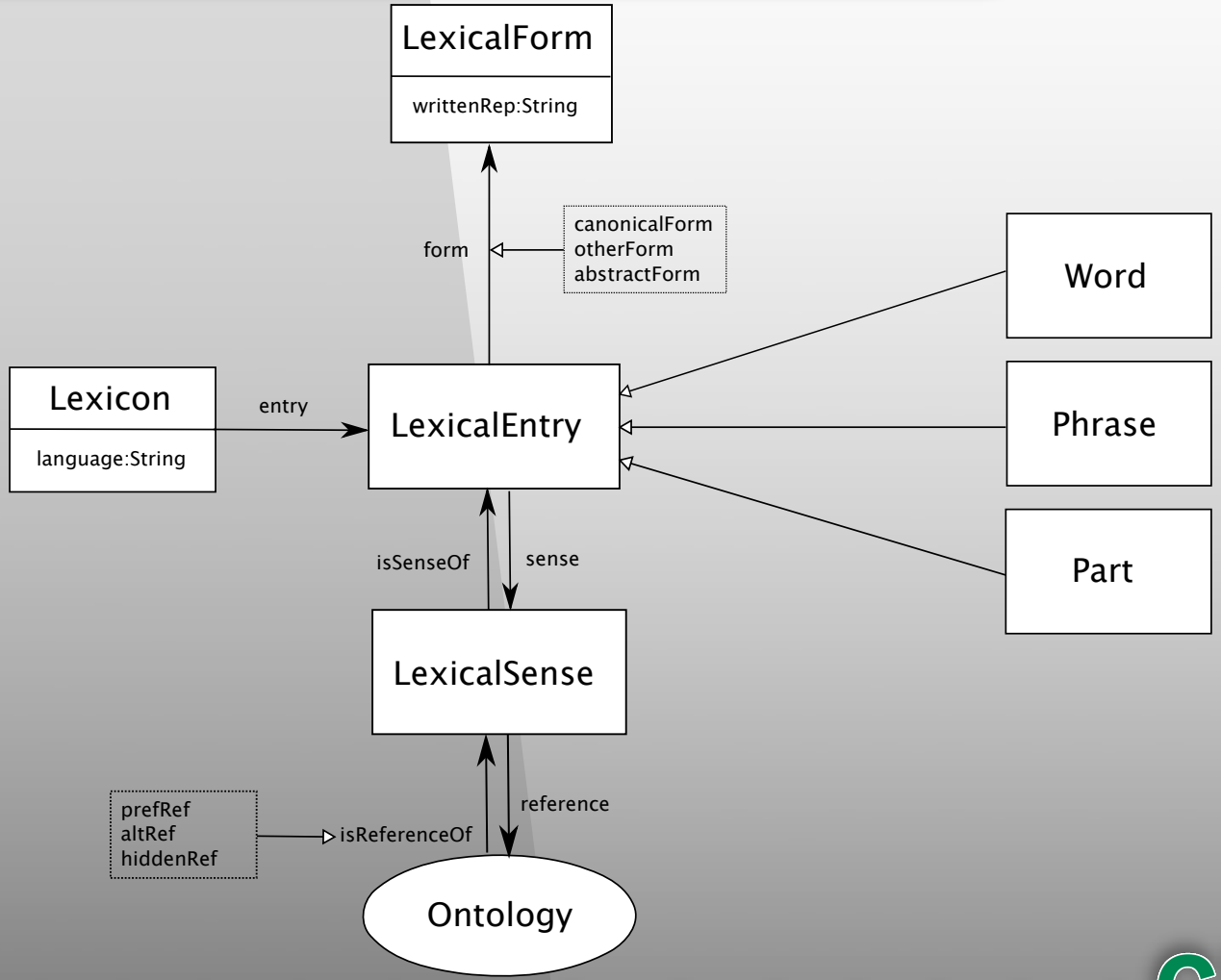
- Multilinguality on the Web
- Semantic Web & Linked Open Data
- **Ontology Lexicalization**
- Question Answering over Linked Data: the QALD Benchmark
- Conclusion

lemon (lexicon model for ontologies)

- (meta-) model for describing ontology lexica with RDF
- enriches an ontology with linguistic information, in particular specifies how ontology concepts correspond to natural language expressions
- the meaning of lexical entries is specified by pointing to elements in the ontology (semantics by reference)
- lexicon and ontology are clearly separated

McCrae et al. (2011): Linking Lexical Resources and Ontologies on the Semantic Web with lemon. In: Proc. of ESWC 2011.

lemon core

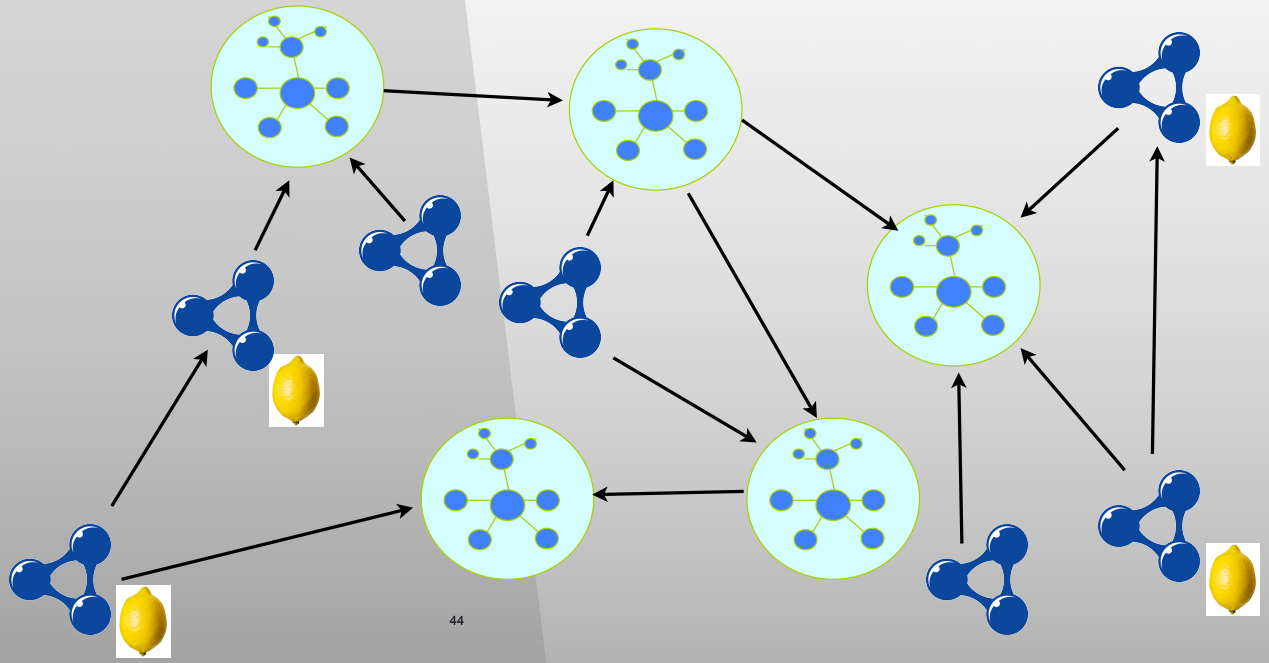

LexicalForm
writtenRep:String
form
canonicalForm
otherForm
abstractForm
Word
Lexicon
language:String
entry
LexicalEntry
Phrase
Part
isSenseOf
sense
LexicalSense
reference
prefRef
altRef
hiddenRef
isReferenceOf
Ontology

Standardization

- W3C Community Group Founded in 2012 to develop a standard model for representing information about how concepts in an ontology are expressed in different languages.
- Currently 73 participants (most of them passive listeners)
- Specification expected some time in 2014!

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

Vision: Towards a linguistically enhanced Semantic Web



lemon lexicon for the English DBpedia ontology

Manually created lexicon for DBpedia Ontology as a proof-of-concept:

- DBpedia ontology comprises 359 classes and 1.775 properties
- First release of English lemon lexicon for DBpedia contains lexicalizations for 354 classes and 300 properties.
- It covers 98% of the classes and 20% of the properties (those with at least 10,000 triples)
- On average, 1.8 entries per ontology entity (1.3 per class and 2.4 per property).

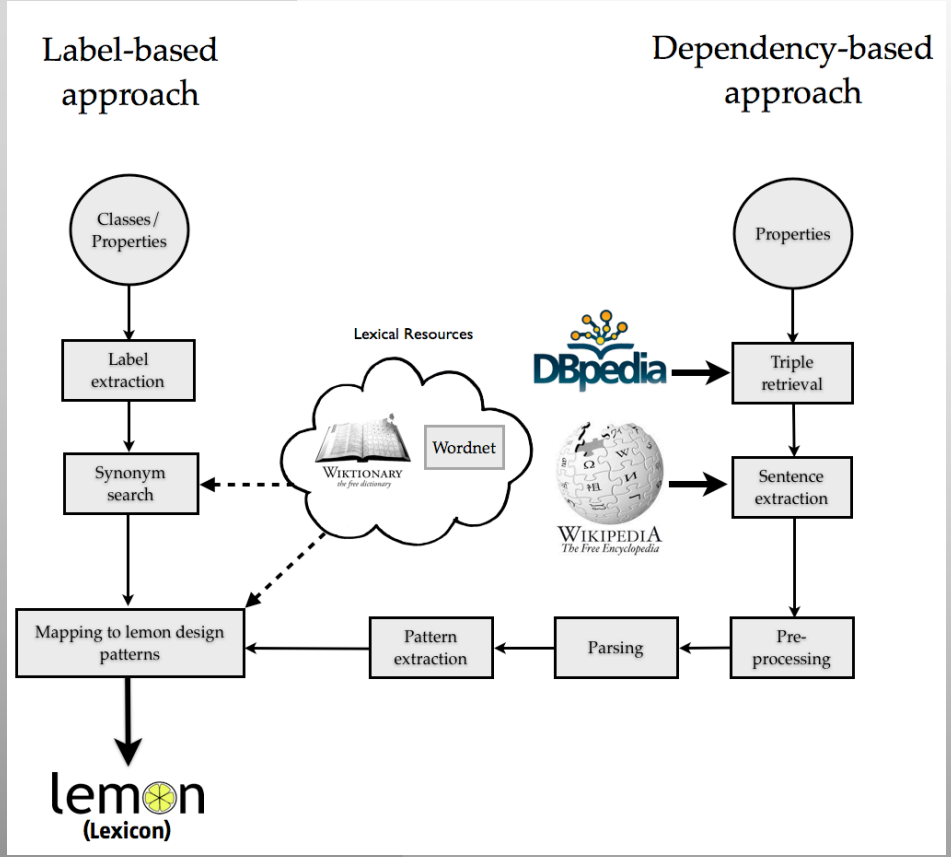
Ontology Lexicalization

- **Given:** property with RDF triples or a whole ontology
- **Find:** patterns that lexicalize these properties
- **Output:** lemon lexicon

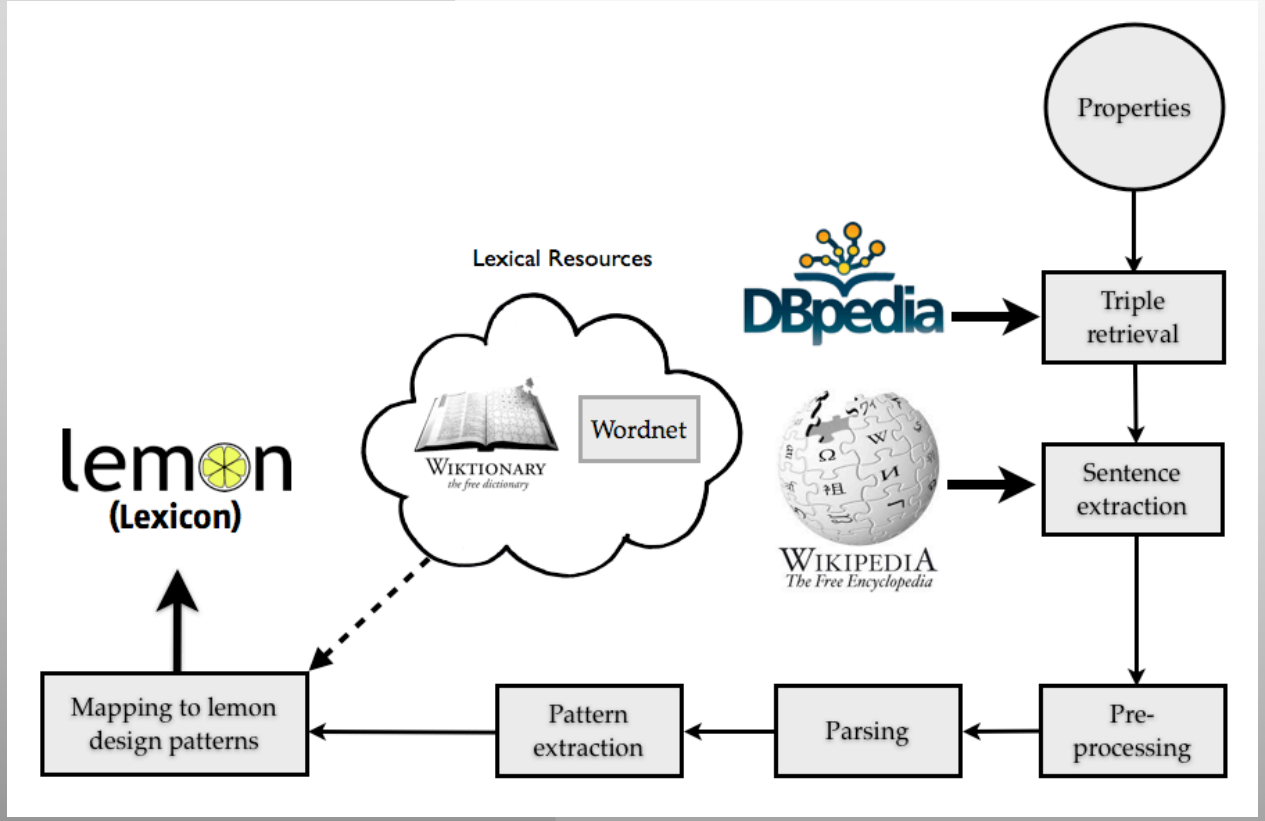
Ontology Lexicalization: ATOLL

- Input: ontology, RDF knowledge base, (domain) corpus
- Approach:
  1. find occurrences in corpus in which a given property is expressed
  2. generalize over these occurrences by extracting dependency paths
- Output: lexicon in lemon format
- Manual validation and correction of lexical entries

ATOLL: System Overview



ATOLL: Dependency-based Approach



Step 1: Triple Retrieval

For the DBpedia property `spouse`, we retrieve among many others...

```
1 <res:Barack_Obama, dbo:spouse, res:Michelle_Obama>
2 <res:Edward_VII,   dbo:spouse, res:Alexandra_of_Denmark>
3 <res:Hilda_Gadea,  dbo:spouse, res:Che_Guevara>
4 <res:Mel_Ferrer,   dbo:spouse, res:Audrey_Hepburn>
5 ...
```

Step 2: Sentence Extraction

We retrieve all sentences mentioning both entities in the same sentence:

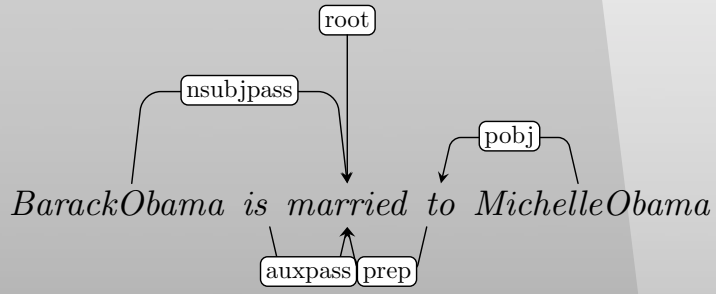
- Barack Obama is the husband of Michele Obama.
- President Barack Obama is married to First Lady Michelle Obama.

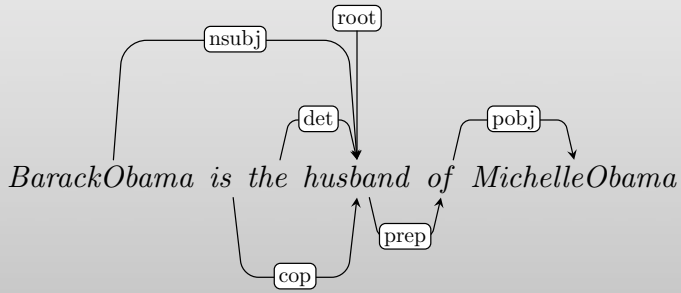
Synonyms extracted from anchor texts

Barack Obama	Term Frequency	Michelle Obama	Term Frequency
Barack Obama	16809	Michelle Obama	866
Obama	957	Michelle Robinson	14
Barack H. Obama	63	First Lady Michelle Obama	9
Barak Obama	60	Michelle LaVaughn Robinson Obama	4
Barack Obama's	47	Michele Obama	4
Barrack Obama	41	Michelle Robinson Obama	3
Barack	31	Melvinia Shields	2
Barack_Obama	22	Mrs. Obama	1
Obama, Barack	13	Michelle obama	1
Sen. Barack Obama	10	Michelle_Obama	1

Step 3: Dependency Parsing

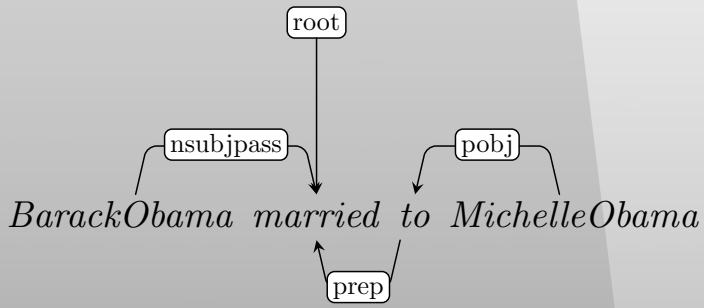
We use the Maltparser to parse all sentences:

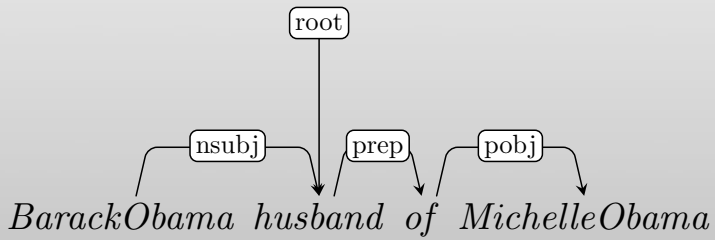




Step 4: Pattern Extraction

Extract shortest path in the dependency tree:





Mapping to lemon

```
:husband a lemon:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  lemon:canonicalForm [ lemon:writtenRep "husband"@en ] ;
  lemon:synBehavior [ rdf:type lexinfo:NounPPFrame ;
                      lexinfo:copulativeArg       :x_subj ;
                      lexinfo:prepositionalObject :y_pobj ] ;
  lemon:sense [ lemon:reference
                <http://dbpedia.org/ontology/spouse>;
                lemon:subjOfProp :x_subj ;
                lemon:objOfProp  :y_pobj ] .

:y_pobj lemon:marker [ lemon:canonicalForm
                       [ lemon:writtenRep "of"@en ] ] .
```

Mapping to lemon using design patterns and a domain-specific language

```
RelationalNoun("husband",dbpedia:spouse,
     propSubj = PossessiveAdjunct,
     propObj  = CopulativeArg),

StateVerb("marry",dbpedia:spouse),
propSubj = lex:partner1 as Subject,
propObj  = lex:partner2 as PrepositionalObject("to")),
```

Evaluation on DBpedia

- DBpedia 3.8
- Wikipedia corpus (EN)
- Training (threshold fixing): 20 classes and 60 properties
- Test: 354 classes, 300-60=240 properties

Evaluation: DBpedia

Evaluation measures:

- Precision and Recall:
  - Precision: How many of the automatically generated lexical entries (at lemma level) are also in the gold standard lexicon?
  - Recall: How many of the lexical entries in the gold standard lexicon are also in the automatically generated lexicon?
- Frame Accuracy:
  - Are the subcategorization frame and its arguments correct?
  - Have these syntactic arguments been mapped correctly to the semantic arguments (domain and range of a property)?

Example for spouse

Obvious ones:

X married/marries Y
X is married to Y
X is the husband of Y
X is the wife of Y
X is the better half of Y

Spurious ones:

Y is the mate of Y
X is the partner of Y
X met Y

Evaluation: DBpedia test

Evaluation on 240 properties and 354 classed from DBpedia for top-5 lemon entries

Precision	Recall	F-Measure	Frame Accuracy
43%	67%	53%	93%

Far from perfect, but appropriate for a semi-automatic mode!

Agenda

- Multilinguality on the Web
- Semantic Web & Linked Open Data
- Ontology Lexicalization
- **Question Answering over Linked Data: the QALD Benchmark**
- Conclusion

Pushing the field: benchmarking

- Benchmarks crucial to push the field further and create incentives for researchers to participate
- Quantitative comparison of different approaches to the same task
- So far: no shared challenge on Question Answering over Linked Data

Different Editions of the QALD Benchmarking Challenge

Four editions of QALD benchmarking campaign:

- QALD-1 (2011), collocated with the Workshop on Question Answering over Linked Data at ESWC, Crete.
- QALD-2 (2012), collocated with the Workshop on Interacting with Linked Data at ESWC, Crete.
- QALD-3 (2013), collocated with the Cross-Language Evaluation Forum (CLEF), Valencia, next week
- QALD-4 (2014), collocated with the Cross-language Evaluation Forum (CLEF), just accepted

TREC-style Methodology

- **Task:** Input: NL Question, Output: SPARQL Query / List of URIs / List of Strings
- **Datsets:** DBpedia, MusicBrainz
- **Gold standard:** a set of NL queries annotated with SPARQL queries, hand-crafted by students not involved in R&D
- **Evaluation:** Precision, Recall, F-Measure, averaged over all queries
- **Infrastructure:** SPARQL endpoint for datasets

Datasets

- **DBpedia:** around 3.5 Million entities extracted from DBpedia, 280 million RDF triples (Version 3.6) and 370 million RDF triples (Version 3.7)
- **MusicBrainz:** a collaborative open- content music database consisting of 15 Mio. RDF triples describing artists, their albums, tracks, when bands where created and split

Training Test

- QALD-1
  - Train: 50+50 (DBpedia / MusicBrainz)
  - Test: 50+50 (DBpedia / MusicBrainz)
- QALD-2
  - Train: 100+100 (DBpedia / MusicBrainz)
  - Test: 100+50 (DBpedia / MusicBrainz)
- QALD-3
  - Train: 100+100 (DBpedia / MusicBrainz), in 6(!) languages: English, Spanish, German, Italian, French, and Dutch
  - Test: 100+100+50 (DBpedia / ESpedia / MusicBrainz)

No. Participants

- QALD-1: 2 (DBpedia) + 2 (MusicBrainz)
- QALD-2: 4 (DBpedia), F-Measures: 0.38 - 0.46
- QALD-3: 6 (DBpedia) + 2 (MusicBrainz), F-Measures: 0.17 - 0.51 (Outlier: 0.90)
- QALD-4: $>$ 10 (???)

Results (QALD-2)

	Precision	Recall	F-measure
SemSeK	0.44	0.48	0.46 (0.36)
Alexandria	0.43	0.46	0.45 (0.11)
MHE	0.36	0.4	0.38 (0.37)
QAKiS	0.39	0.37	0.38 (0.13)

Questions answered by all systems (QALD-3)

- What is the capital of Canada?
- Who is the governor of Wyoming?
- What is the birth name of Angela Merkel?
- How many employees does Google have?

Questions answered by no system (QALD-3)

- Give me all members of Prodigy.
- Does the new Battlestar Galactica series have more episodes than the old one?
- Show me all songs from Bruce Springsteen released between 1980 and 1990.
- Give me all B-sides of the Ramones.

Phenomena/Challenges

- Lexical Gap (59% – 65%): Who is Barack Obama married to? (Property: spouse)
- Lexical Ambiguities (25%): Who was the wife of Lincoln?
- Linguistically Light Expressions (10%–20%): Give me all movies with Tom Cruise (Property: starring)
- Complex Queries (24%)
  - Aggregation: How many bands broke up in 2010?
  - Superlatives: What is the highest mountain?
  - Comparisons: Who recorded more singles than Madonna?
  - Temporal: Who was born on the same day as Frank Sinatra?

Conclusion

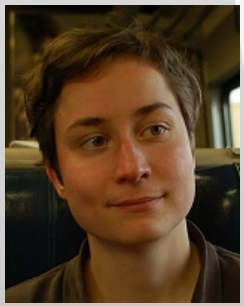
- Vision of a Multilingual Semantic Web
- Linked Data is key, but needs to be enriched with linguistic information
- Querying it in natural language is a challenge: lexical gap, complex queries, aggregation
- Querying it in multiple languages is an even greater challenge
- Divide-And-Conquer: Collaborative approach with semi-automatic support, involve people that care about their ontology and ask them to do some manual work.

Outlook

- Support Multiple Languages
- Addressing Property Promiscuity (e.g. team)
- Leverage the Wisdom and Resources of the Masses: Crowdsourcing
- Proof-of-Concept for Multilingual QA

Acknowledgements

Christina Unger:



John McCrae:

Sebastian Walter:

monnet

PortDial

CITEC

Co-organizers of the QALD Challenges

- Prof. Dr. Enrico Motta (Knowledge Media Institute, Open University)
- Dr. Paul Buitelaar (DERI Galway)
- Dr. Richard Cyganiak
- Dr. Vanessa Lopez (IBM Dublin)
- Dr. Axel Ngonga-Ngonga (University of Leipzig)
- Dr. Elena Cabrio (INRIA)

And all people that have provided suggestions, data, queries, etc.