Panel #3: Data and Metadata of Language Resources as Linked Data on the Web

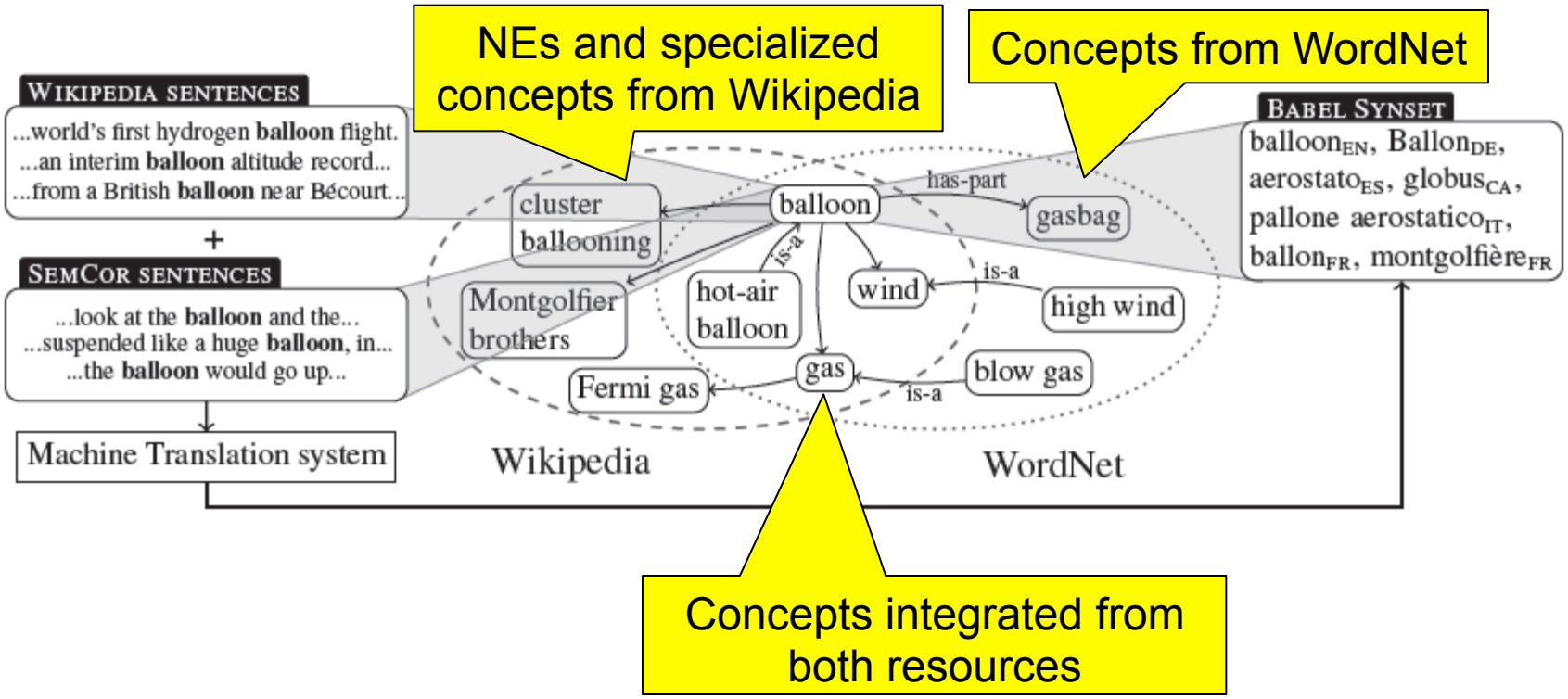
Roberto Navigli

Sapienza University of Rome

Data… what data?

BabelNet (http://babelnet.org)

- A wide-coverage multilingual semantic network and encyclopedic dictionary in 50 languages!



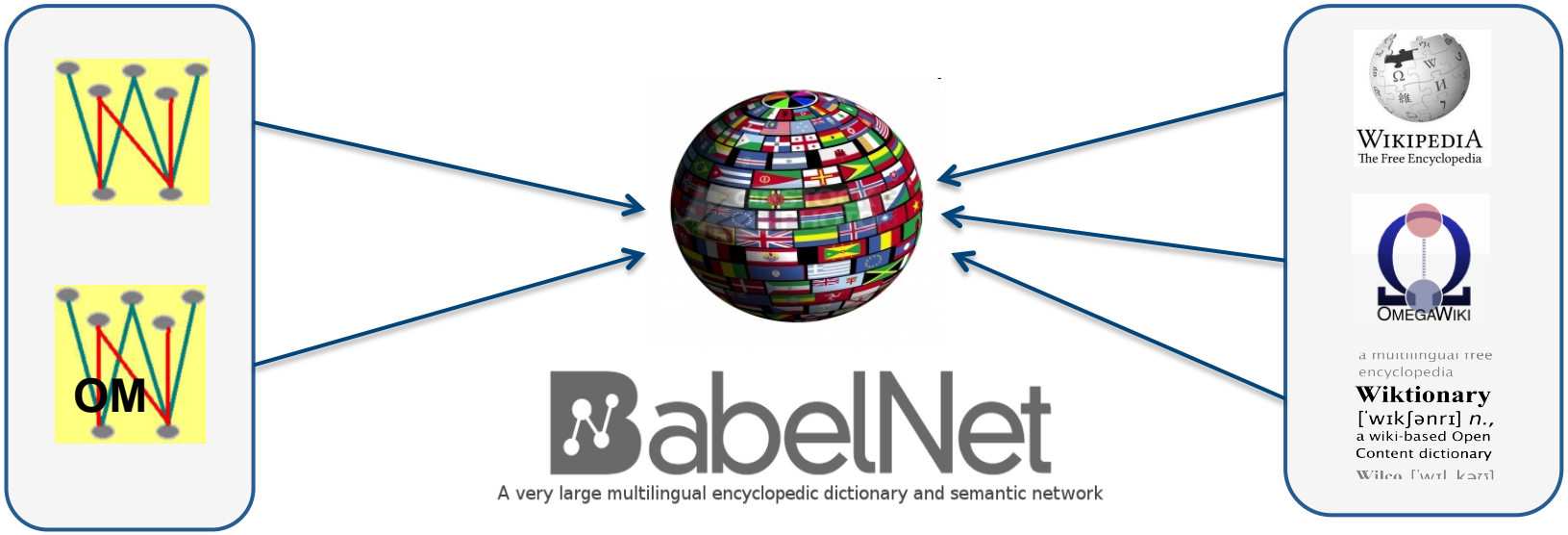
BabelNet (http://babelnet.org)

BabelNet is both a **multilingual encyclopedic dictionary**, with lexicographic and encyclopedic coverage of terms in **50 languages**, and a **semantic network** which connects concepts and named entities in a **very large network of semantic relations**, made up of **more than 9 million entries**. Read more...

Acquisition

Traditional lexical resources

Collaborative lexical resources



- fully-structured
- manually curated by experts
- available for a few languages
- difficult to maintain and update

- semi-structured
- collaboratively built by the crowd
- highly multilingual
- up-to-date

Main components

Babel synset

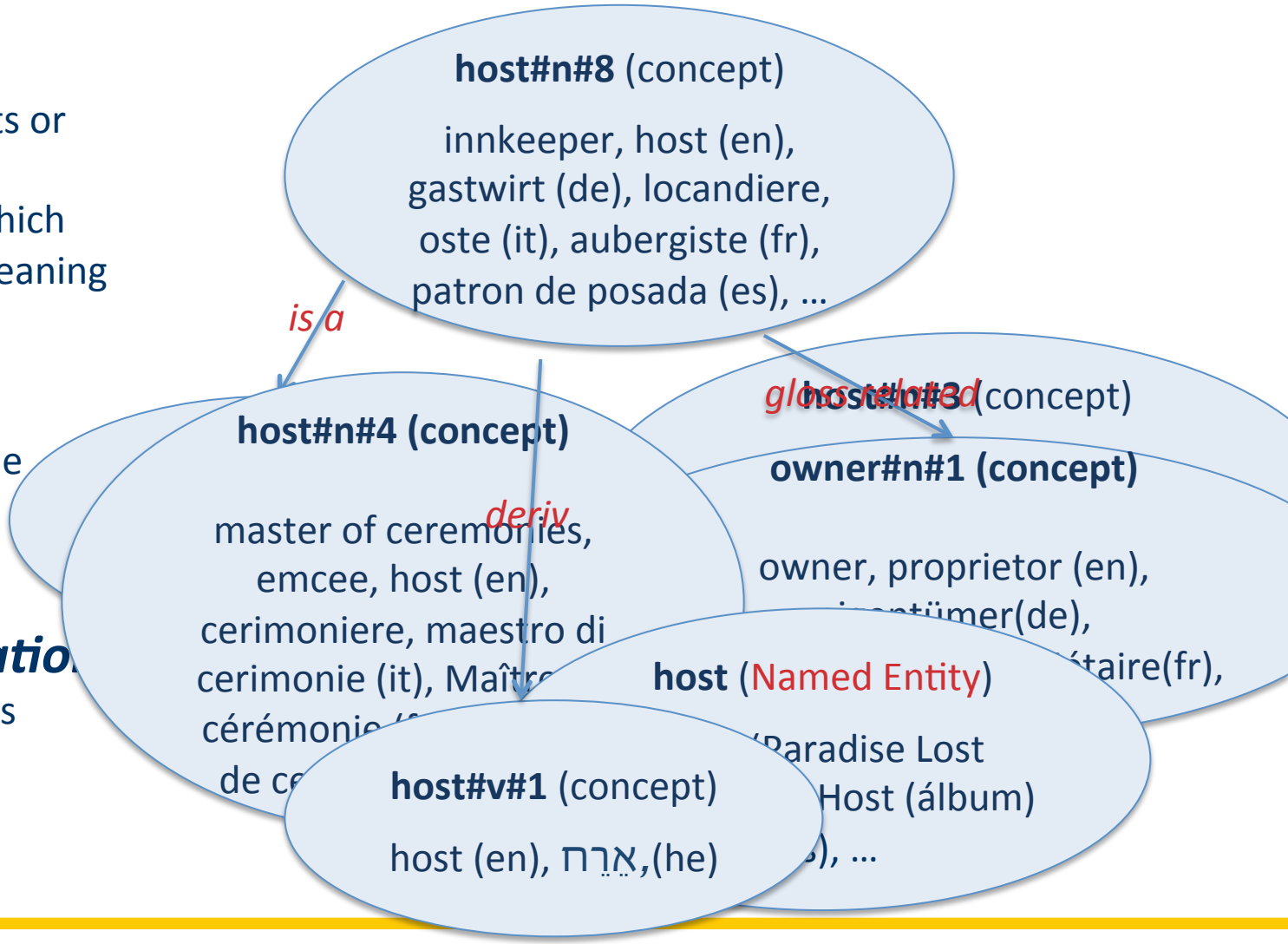
- encode concepts or named entities
- group senses which express their meaning

Babel senses

- terms in multiple languages

Semantic relation

- connect synsets



babelnet.org#lemonRepresentation

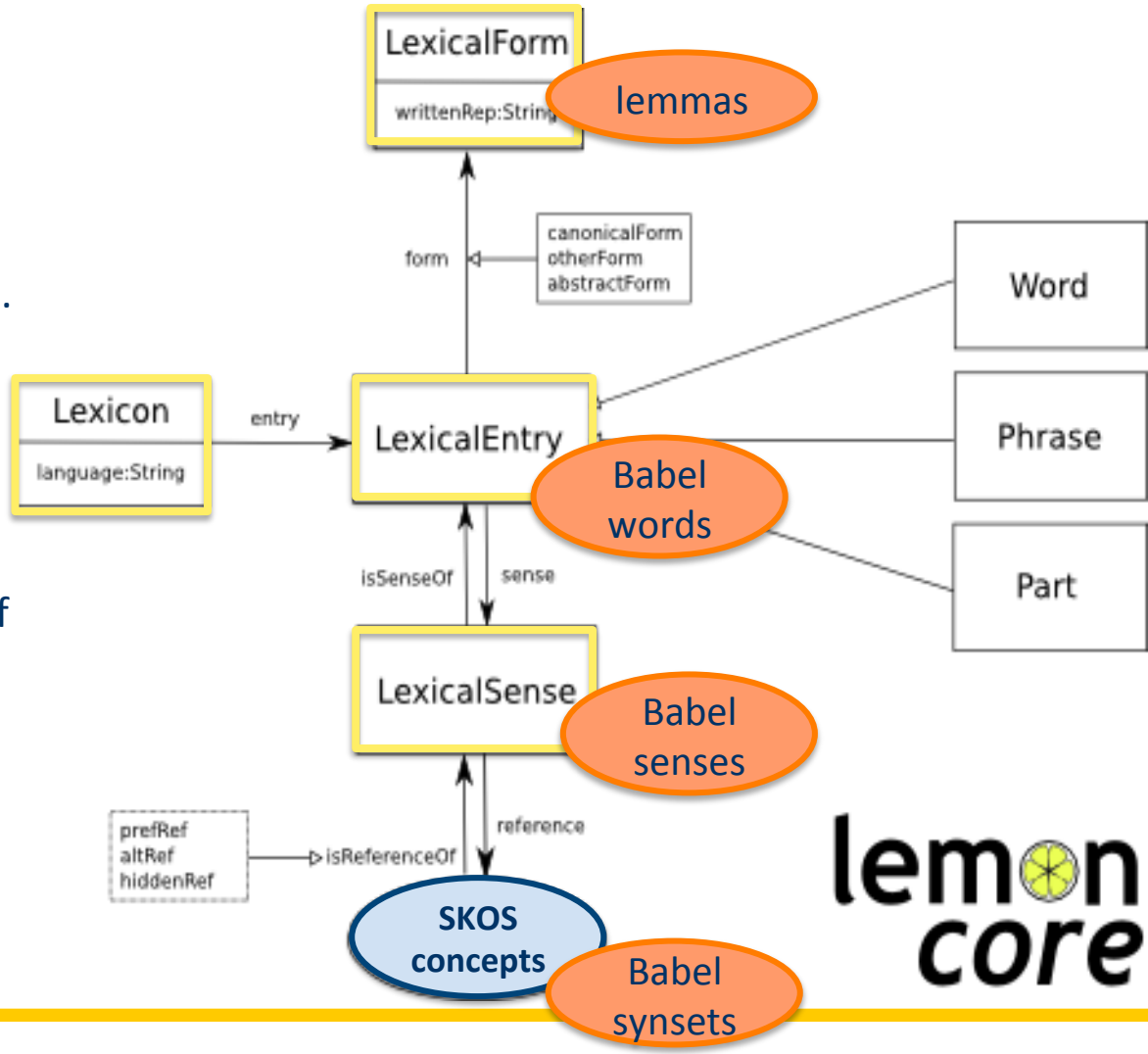
- the RDF resource consists of a set of **Lexicons**, one per language.

- Lexicons gather **Lexical Entries** which comprise the forms of an entry;
in our case: *words* of the Babel lexicon.

- **Lexical Forms** encode the surface realisation(s) of Lexical Entries;
in our case: *lemmas* of Babel words.

- **Lexical Senses** represent the usage of a word as *reference* to a specific concept;
in our case: Babel senses.

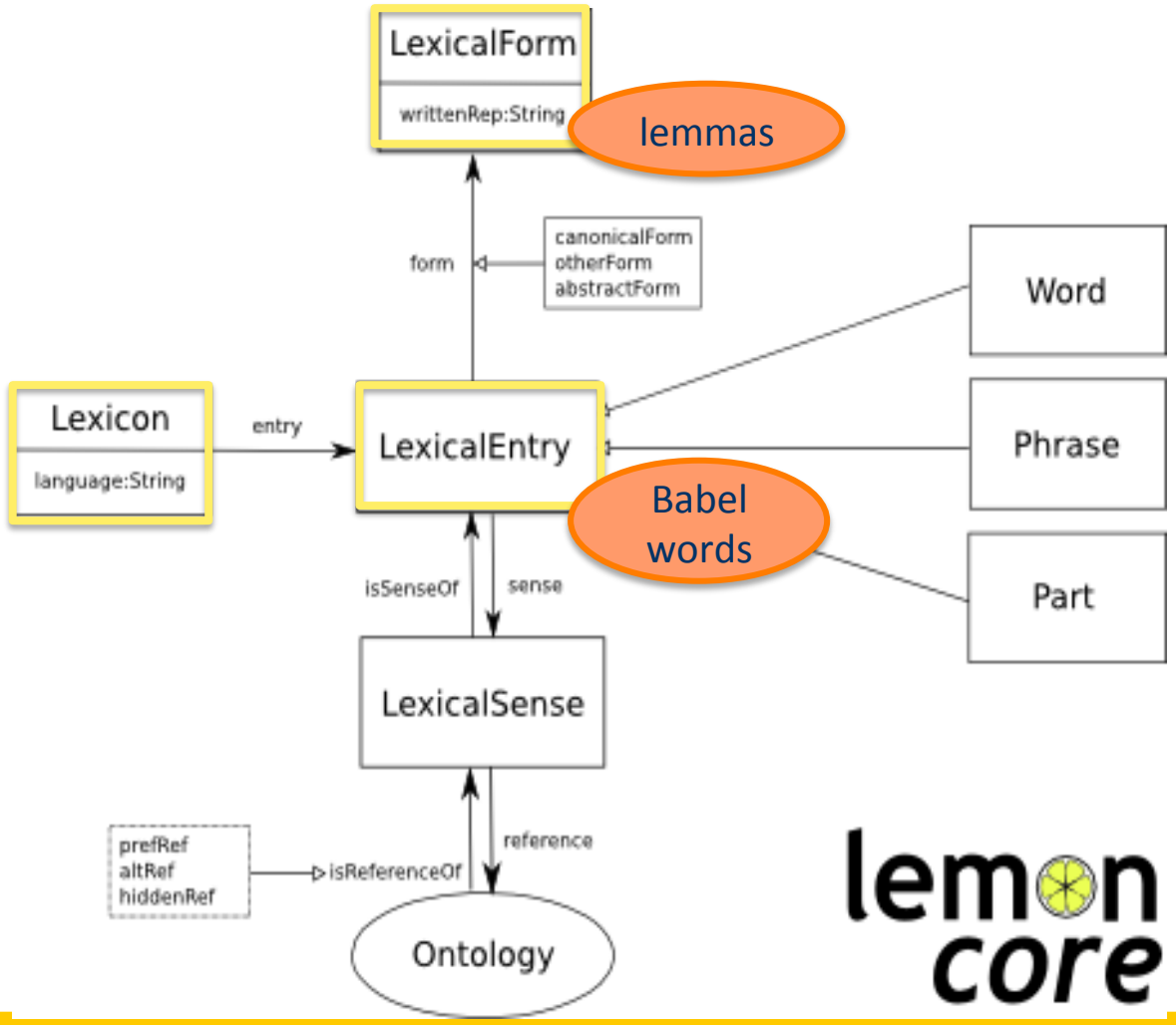
- **Skos Concepts** represent 'units of thought';
in our case: Babel synsets.



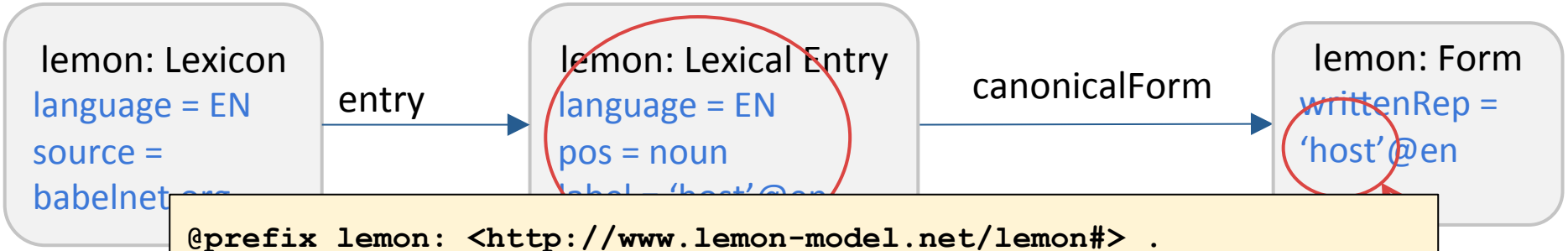
babelnet.org#vocabularies

- lemon
  - backbone of BabelNet lexical knowledge RDF representation
  - http://www.lemon-model.net/lemon#
- SKOS (Simple Knowledge Organization System)
  - skos:Concept class represents Babel synsets
  - http://www.w3.org/2004/02/skos/core#
- LexInfo 2.0
  - ontology which describes linguistic information
  - used here to represent various linguistic information
  - http://www.lexinfo.net/ontology/2.0/lexinfo#
- BabelNet:
  - domain name: http://babelnet.org/2.0/
  - vocabulary: http://babelnet.org/model/babelnet#

babelnet.org#bn2lemon



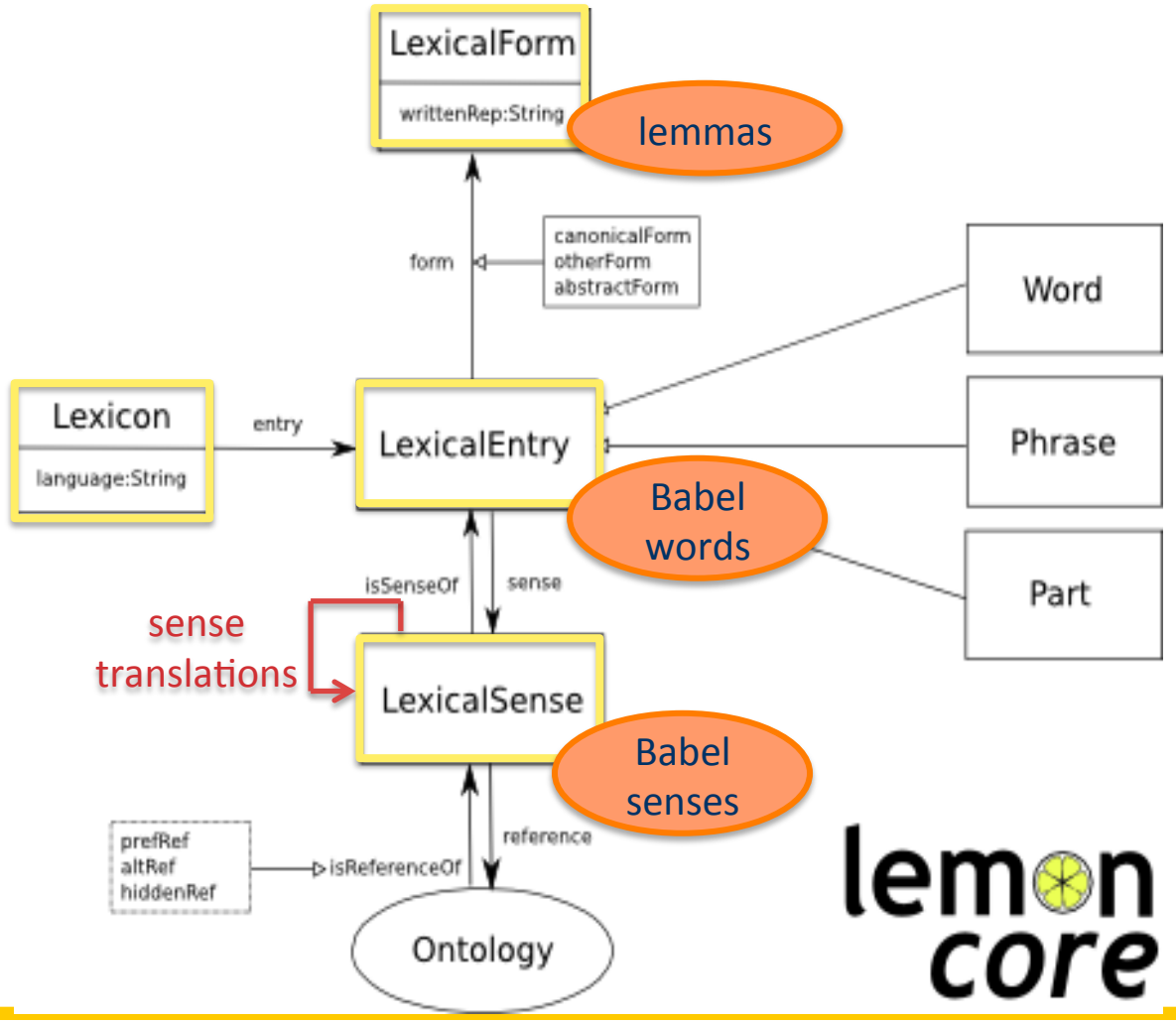
babelnet.org#representingWords




```
@prefix lemon: <http://www.lemon-model.net/lemon#> .
@prefix bn: <http://babelnet.org/2.0/> .
@prefix lexinfo:<http://www.lexinfo.net/ontology/2.0/lexinfo#> .

bn:lexicon_EN
        a                   lemon:Lexicon ;
        lemon:entry         bn:host_n_EN .

bn:host_n_EN
        a                           lemon:LexicalEntry ;
        rdfs:label                  "host"@en ;
        lemon:canonicalForm         bn:host_n_EN/canonicalForm ;
        lemon:language              "EN" ;
        lexinfo:partOfSpeech        lexinfo:noun .

bn:host_n_EN/canonicalForm
        a                       lemon:Form ;
        lemon:writtenRep        "host"@en .
```

lemma

babelnet.org#bn2lemon



Summarizing our experience (1)

- URL design
  - alternative: IRI vs. URI vs. both
  - choice: IRI for readability, URI for higher "computational compliance"
- do we adopt Lemon as is or do we go for the main classes only?
  - finding the best tradeoff between richness and simplicity
  - we went for quite simple at the beginning of the project (leaving the Lexicons for example), before moving towards a fine-grained and almost comprehensive representation of BN information.
- RDF BabelNet design, several problems:
  - how to model links to Wikipedia and DBpedia information: rdf:seeAlso vs. owl:sameAs vs. skos:exactMatch

Summarizing our experience (2)

- How to model BN textual definitions:
  - hard with skos, therefore creation of a in-house/dedicated class.
- Representation of classes of senses (=synsets): skos:concept or owl:class?
  - Adoption of skos:concept, thanks to its definition, and because of its use to model similar objects in other RDF resources (e.g. WordNet)
- Metadata encoding with dc terms, synset representation with SKOS, linguistic properties with LexInfo, etc.
- The choice of the models and the definition of properties got refined as the conversion work went ahead

Issues with Data and Potential of LLD

- Having too many data makes it hard to explore and connect them
  - Plus: redundancy could be a problem, but also a resource
- But: links help considerably!
- Linked Data help collect information from different resources at the same time on the basis of connections between knowledge resources

Metadata… which metadata?

babelnet.org#representingSenses

lemon: Lexical Entry

language = EN
pos = noun
label = 'host'@en
sense = host_sense1

lemon:sense

lemon: Lexical Sense

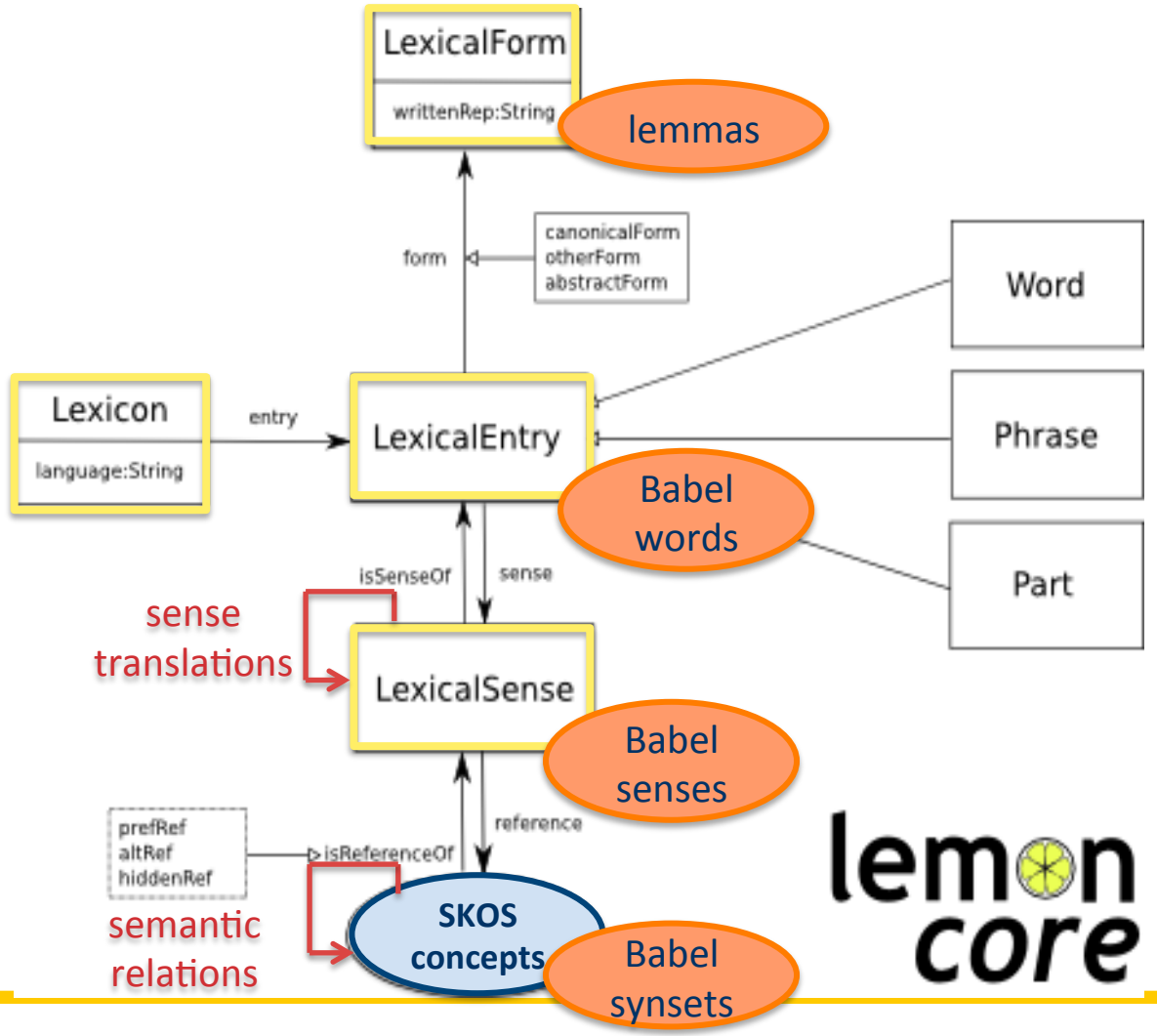
lemon:reference = host_synset1
dc:source = wordnet
dc:license = wordnet_license
bn-lemon:byRedirection = true/false
bn-lemon:byTrans = true/false
lexinfo:translations = other_sense(s)

```
bn:host_n_EN
        a                   lemon:LexicalEntry ;
        …                   …
        lemon:sense         bn:host_EN/s00012029n .

http://babelnet.org/2.0/host_EN/s00012029n
        a                       lemon:LexicalSense ;
        dc:source               http://wordnet.princeton.edu/
        dc:license              http://wordnet.princeton.edu/wordnet/license/
        lemon:reference         bn:s00012029n
        lexinfo:translation     bn:gastwirt_DE/s00012029
```

Provenance

License information about each sense

babelnet.org#bn2lemon



babelnet.org#representingSynsets

lemon:LexicalSense

lemon:reference = host_synset1

dc:source = sense_source

```
http://babelnet.org/2.0/host_EN/s00012029n
        a                   lemon:LexicalSense ;
        …                   …
        lemon:reference     bn:s00012029n

http://babelnet.org/2.0/s00012029n
        a                   skos:Concept
        bn-definition       http://babelnet.org/2.0/s00012029n_Gloss1_EN
        dc:license          <http://creativecommons.org/licenses/by-nc-sa/3.0>

http://babelnet.org/2.0/s00012029n
        a                   bn-lemon:BabelGloss
        bn-lemon:gloss      'the owner or manager of an i
        lemon:language      EN
        dc:source           http://wordnet.princeton.edu/
        dc:license          http://wordnet.princeton.edu/wordnet/license
```

Provenance

License information about each sense

babelnet.org#semanticRelations

- skos:narrower/broader for hypo and hypernyms
- lexinfo relations (meronym, holonym, derivedForm, etc.)
- skos:related for unspecified relations

```
http://babelnet.org/2.0/s00012029n
        a                       skos:Concept
        …
        skos:broader            bn:s00061046n
        lexinfo:derivedForm     bn:s00089484v
        skos:related            bn:s00007078n
```

On the architecture

- …should make it easy to collect lexical data of interest from as many LD as possible
  - This will require NLP services such as term/LD extraction and Babelfy-like entity linking and disambiguation
- …with clear constraints on metadata and controlled access to data
  - See Victor Rodríguez Doncel's work on *Linguistic Linked Licensed Data*