Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

# Representing WordNets with OntoLex and the Global Wordnet Formats

John P. McCrae

Data Science Institute
National University of Ireland Galway

A World Leading SFI Research Centre

elexis european lexicographic infrastructure

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN Science Foundation Ireland SFI For what's next

# Outline

1. Introduction to OntoLex
2. OntoLex-Lemon Model
3. Global WordNet Grid
4. The GWN Format
5. Beyond Princeton WordNet
    a. Wikipedia linking
    b. Colloquial WordNet
    c. Open sourcing wordnet

# Introduction to OntoLex

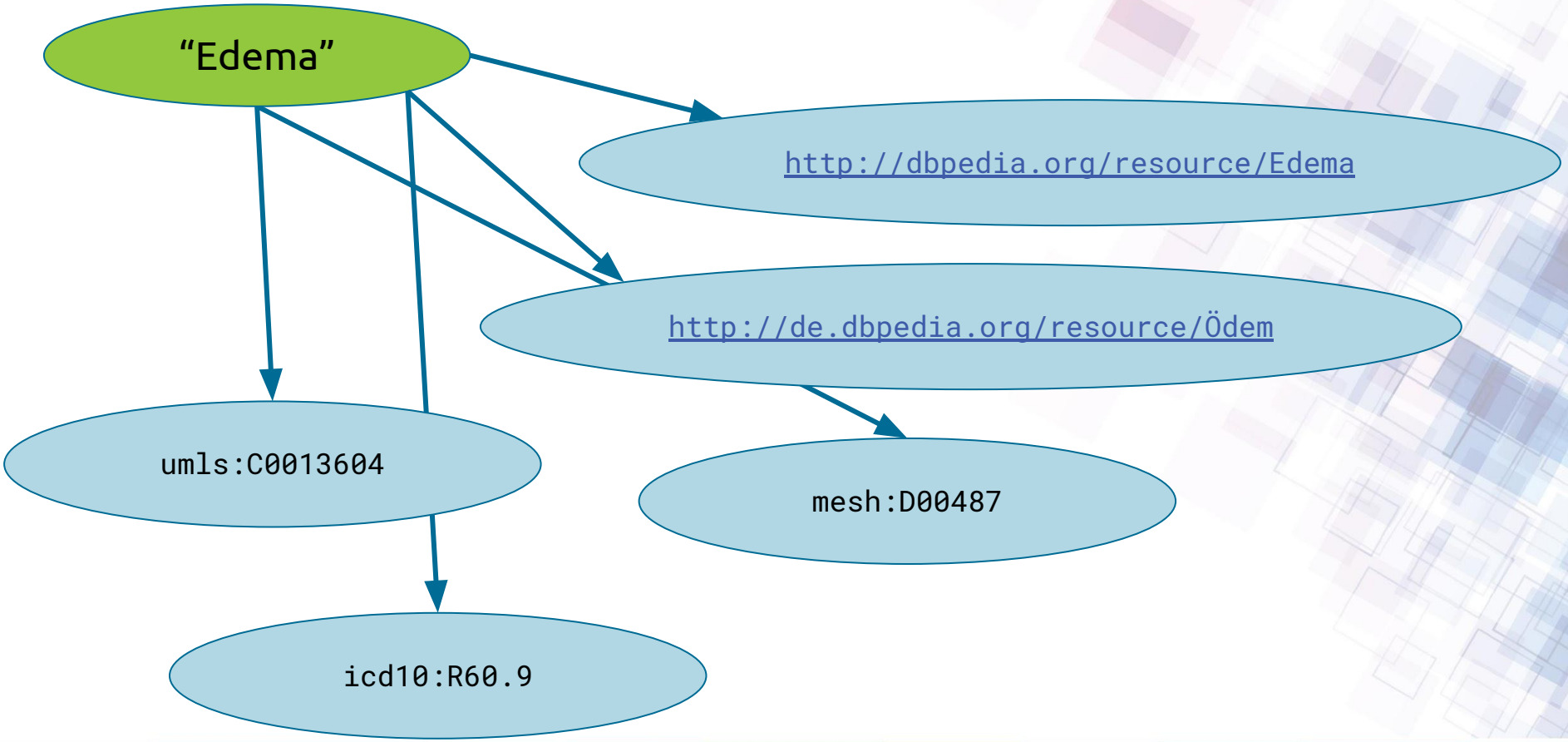
# Semantic Web

http://www.example.com/foo/ID1234

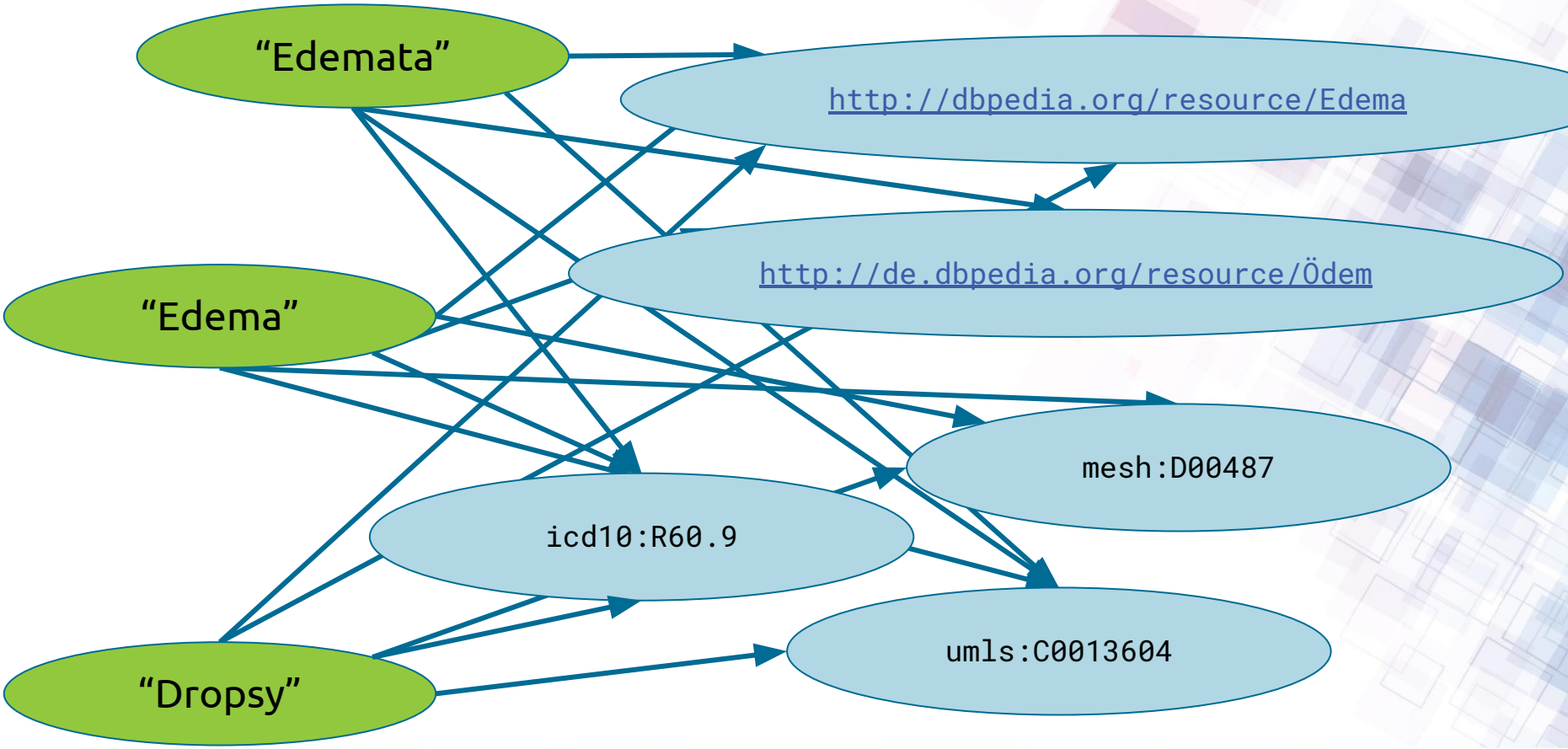
“Language”@en
“Teanga”@ga

...Cuairt Liteartha do Theangacha Mionlaigh san Eoraip...

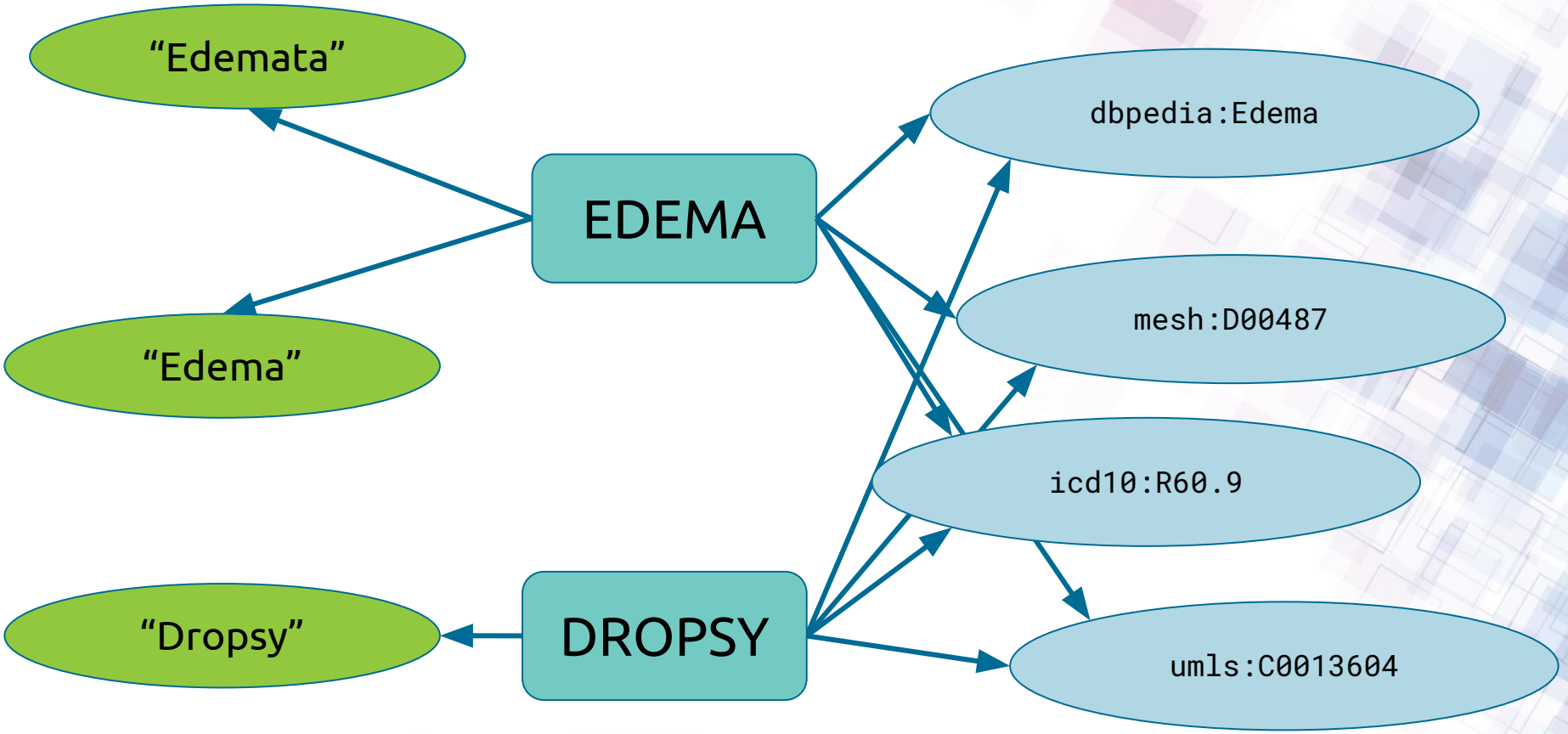
# Linked Data on the Web



# Linked Data with Language



# Lexical Entries



# The OntoLex-Lemon Model

# The OntoLex-Lemon Model

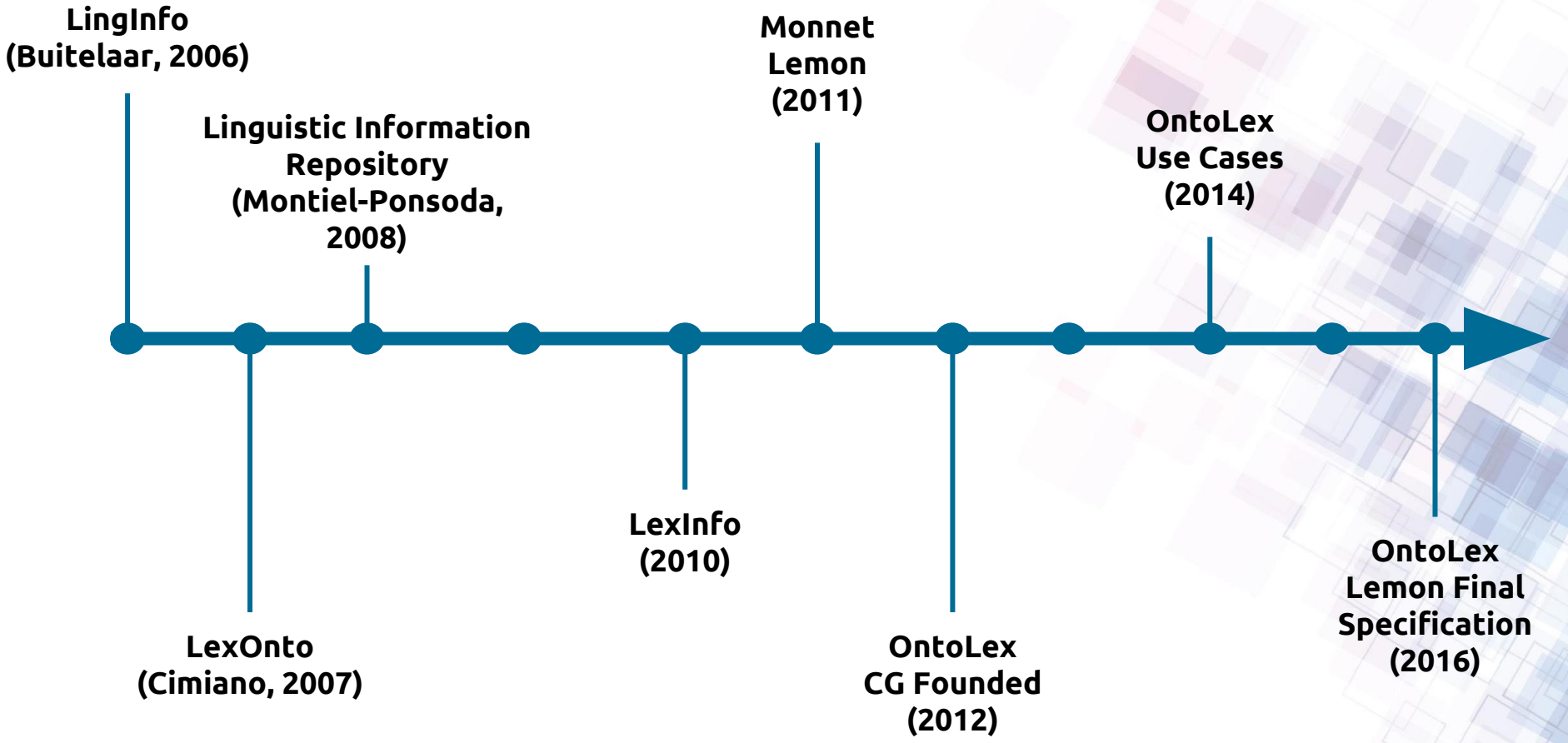
Proposed as a model for *representing lexical information relative to ontologies*

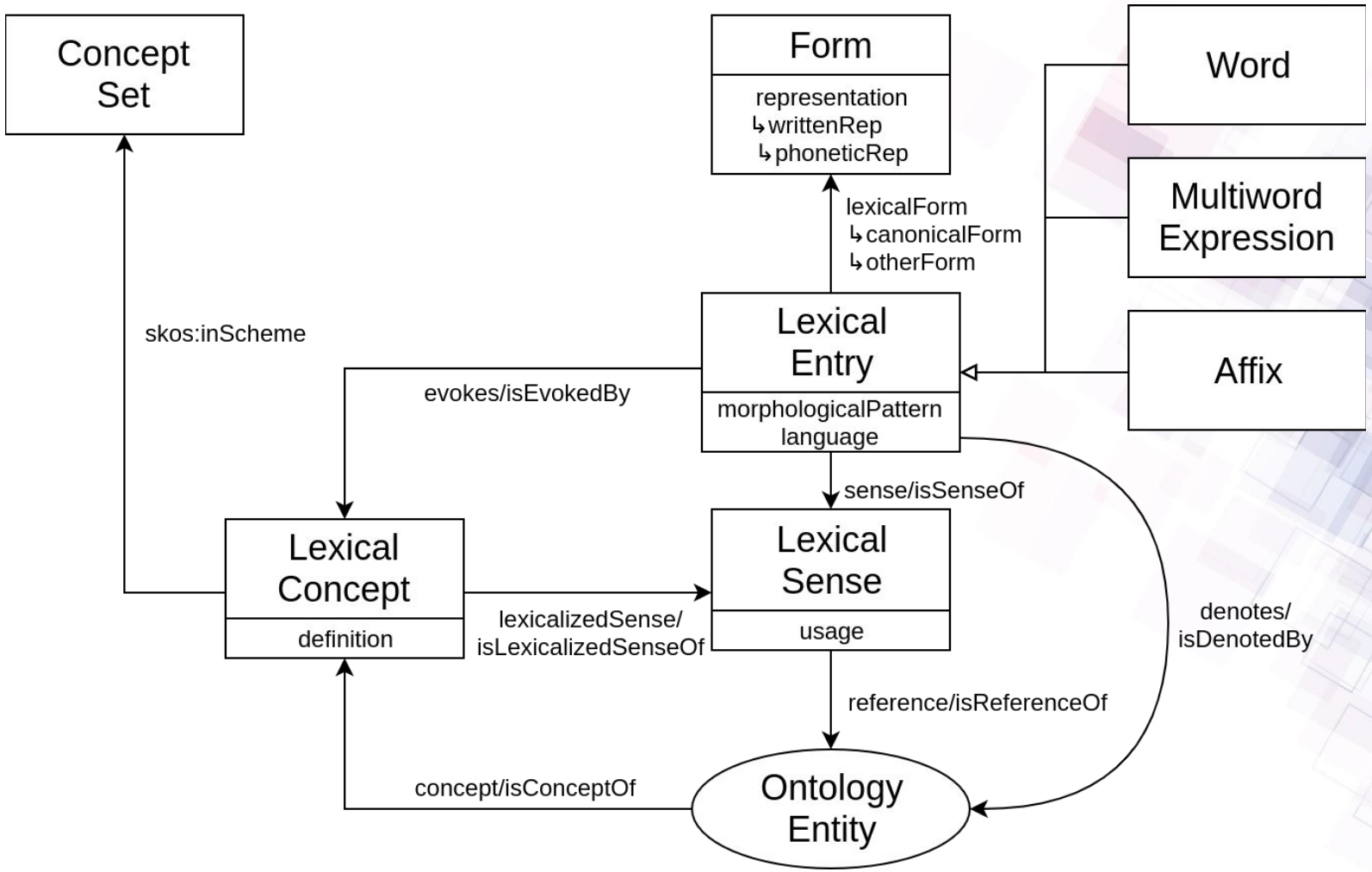
- Based on existing models
  - LMF
  - EAGLES/ISLES
  - SKOS
  - LingInfo/LexOnto/LexInfo/LIR
- Open
  - Anyone can contribute
  - All contributions are public
  - No licensing restrictions

# OntoLex or Lemon?

The Ontolex Community Group's primary result is the Lexicon Model for Ontologies (Lemon), which consists of the following modules:

- (OntoLex) Core
  - http://www.w3.org/ns/lemon/ontolex
- Syntax and Semantics
  - http://www.w3.org/ns/lemon/synsem
- Decomposition
  - http://www.w3.org/ns/lemon/decomp
- Variation and Translation
  - http://www.w3.org/ns/lemon/vartrans
- Metadata (Lime)
  - http://www.w3.org/ns/lemon/lime


LingInfo
(Buitelaar, 2006)
Linguistic Information Repository
(Montiel-Ponsoda, 2008)
Monnet Lemon
(2011)
OntoLex Use Cases
(2014)
LexOnto
(Cimiano, 2007)
LexInfo
(2010)
OntoLex CG Founded
(2012)
OntoLex Lemon Final Specification
(2016)

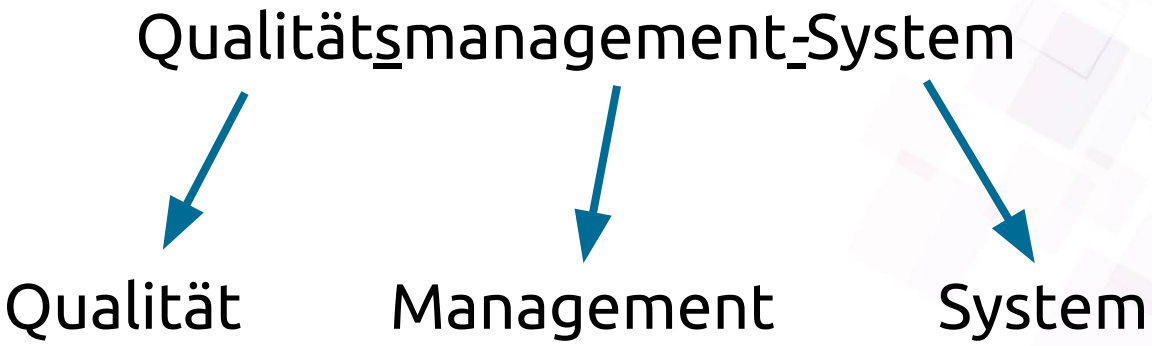

Concept Set
Form
representation
↳writtenRep
↳phoneticRep
Word
Multiword Expression
Affix
lexicalForm
↳canonicalForm
↳otherForm
skos:inScheme
Lexical Entry
morphologicalPattern
language
evokes/isEvokedBy
sense/isSenseOf
Lexical Concept
definition
Lexical Sense
usage
lexicalizedSense/
isLexicalizedSenseOf
denotes/
isDenotedBy
reference/isReferenceOf
concept/isConceptOf
Ontology Entity

# Syntax and Semantics

John knows Philipp

```
<http://john.mccr.ae> foaf:knows agsc:cimiano
```

# Decomposition



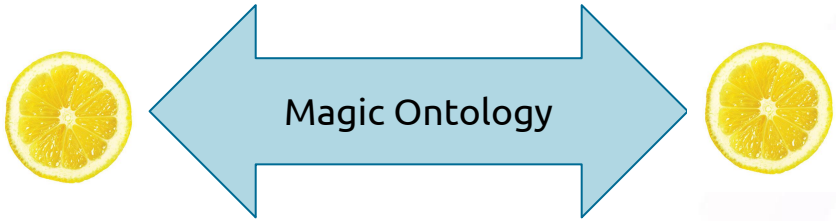
# Variation and Translation

Cultural Translation

“もち”@ja

“Japanese Rice Cake”@en

# Linguistic Metadata



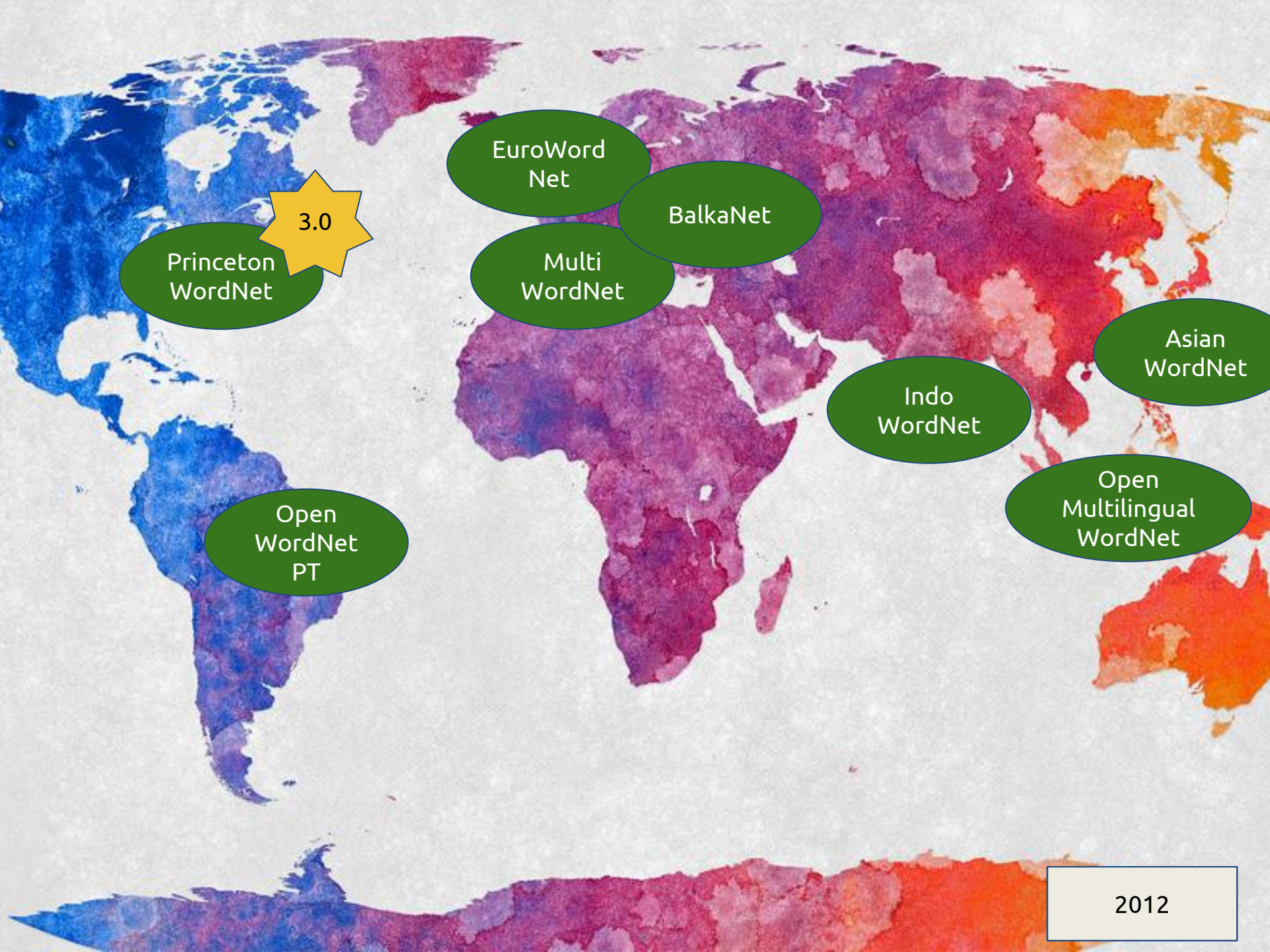

Jace the Wizard

Erhnam the Djinn

# Global WordNet Grid



Slides also thanks to Piek Vossen and Francis Bond


EuroWord Net
BalkaNet
3.0
Princeton WordNet
Multi WordNet
Asian WordNet
Indo WordNet
Open Multilingual WordNet
Open WordNet PT
2012

# Problems

Different relations used

Different definitions of synonymy (tight, loose)

Different interpretations of relations

Anglo-saxon view of Princeton WordNet

Updates to Princeton break all other wordnets

Different coverage

*"25 synsets shared from 117,677 (0%)" - Open Multilingual WordNet (Bond)*

# Solution:
# CILI - Collaborative Interlingual Index

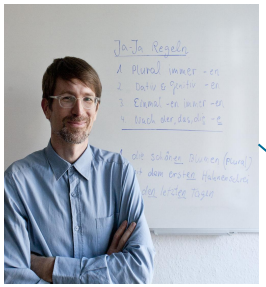
All open source wordnets linked to a single ILI

- Merge of concepts across all languages
- Princeton WordNet versions linked to ILI
- Adaptable by the open source wordnet community.
- Available as LOD and downloadable with open source license: CC-BY, CC-BY-SA.

# How to define a concept

1. A **unique**, **permanent** URI
2. A proper (!) **English** gloss
3. **Linked** to at least one other synset

# Cross-lingual Mapping



Lehrer

Teacher

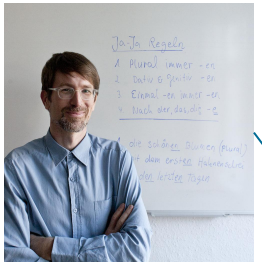
Professore


Lehrerin

Professoressa

# Interlingual Mapping

Teacher



Lehrer

Professore

i123456


Lehrerin

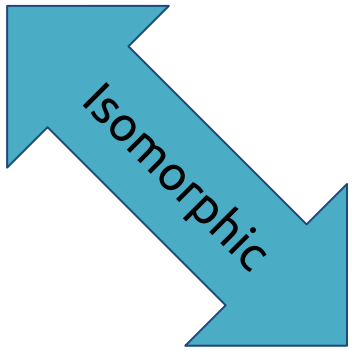
i123457

Professoressa

# WordNet LMF, JSON and RDF

WordNet LMF is 'LMF-like' XML format

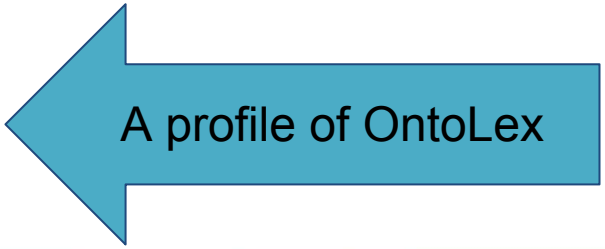
Converter/validator
http://server1.nlp.insight-centre.org:8080/gwn-converter/




RDF/XML

Turtle

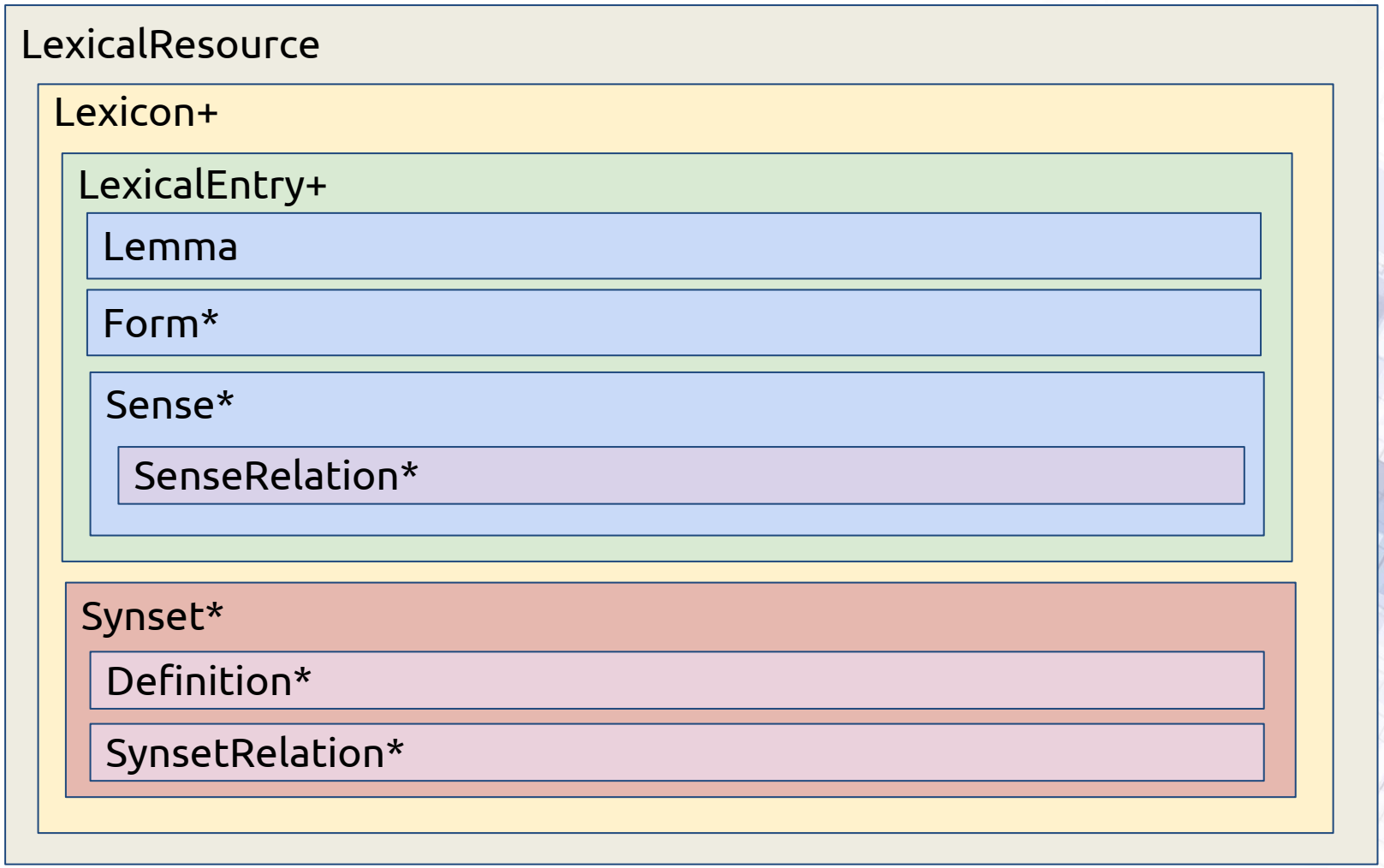
SPARQL



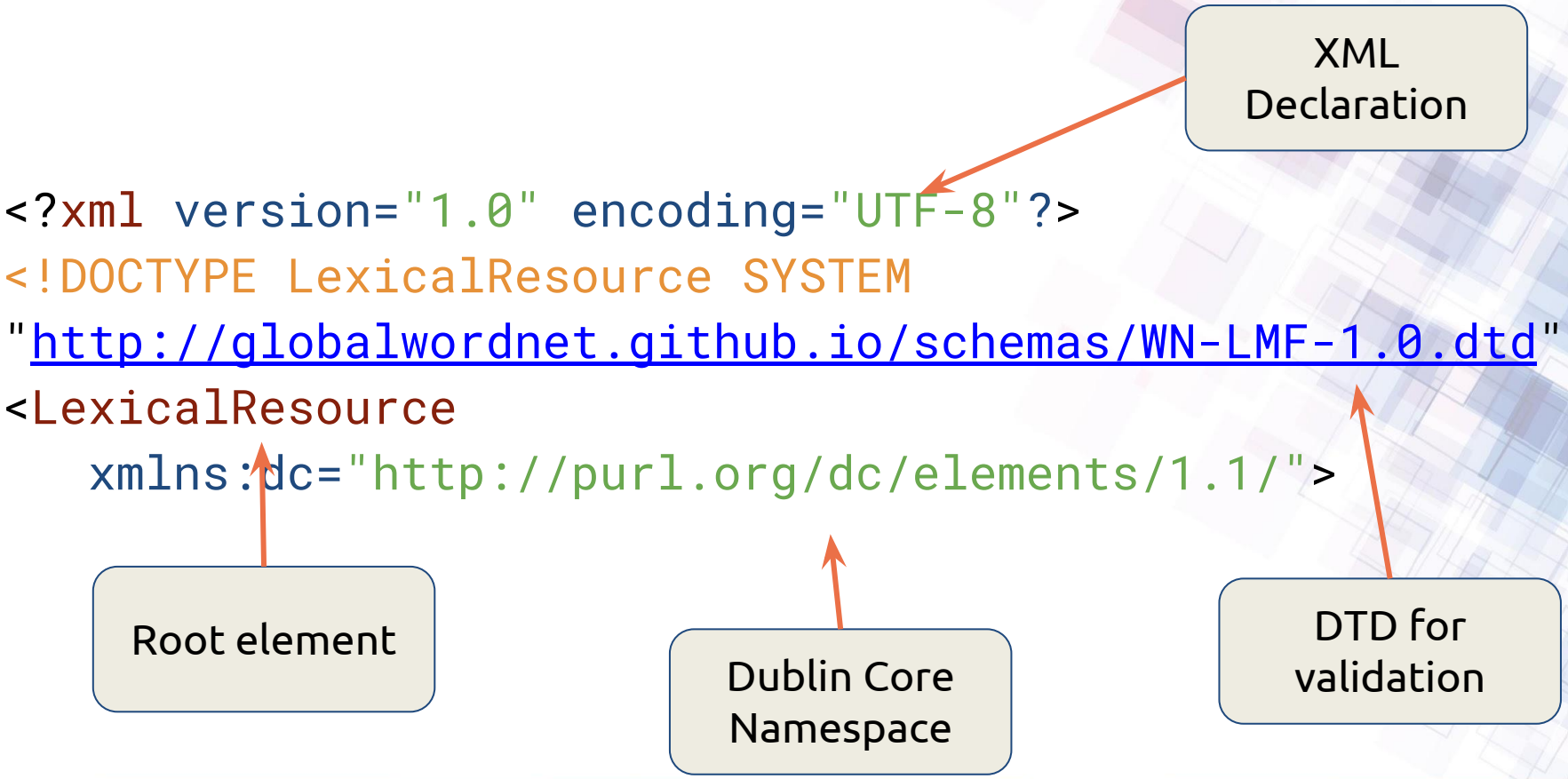

WordNet JSON is JSON-LD based representation

# The GWN WordNet Formats

# GlobalWordNet Structure



# GlobalWordNet XML: Header



# GlobalWordNet XML: The Lexicon

```
<Lexicon id="example-en"
         label="Example wordnet (English)"
         version="1.0"
         language="en"
         email="john@mccr.ae"
         citation="CILI: the Collaborative ..."
    license="https://creativecommons.org/publicdomain/zero/1.0/"
         url="http://globalwordnet.github.io/schemas/"
         dc:publisher="Global Wordnet Association">
```

ID, label, version

Language (ISO-639)

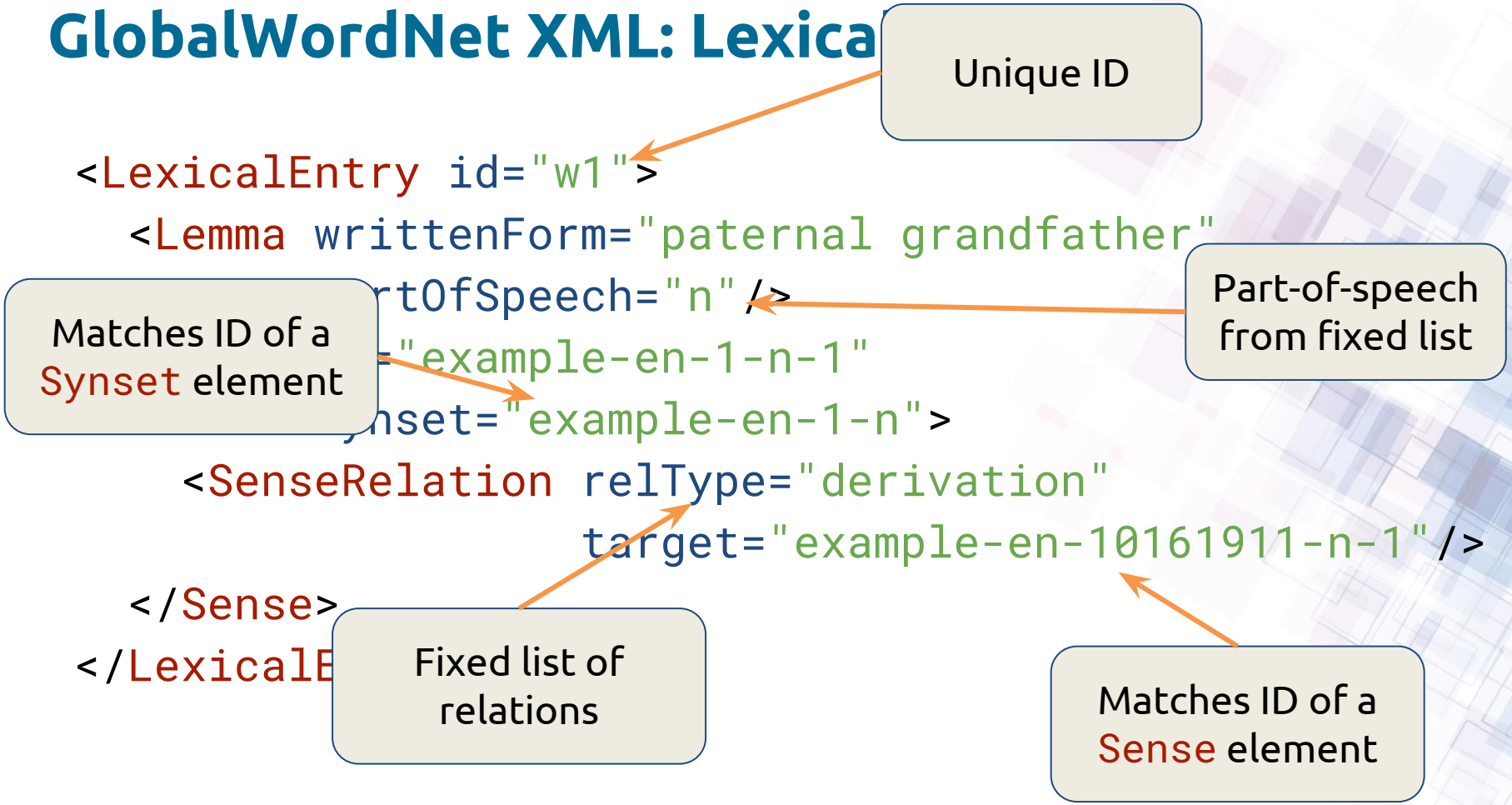
Author/ citation

Dublin Core properties

Homepage

License

# GlobalWordNet XML: Lexica



# GlobalWordNet XML: Synsets

```
<Synset id="example-en-10161911-n"
        ili="i90287"
        partOfSpeech="n">
  <Definition language="en">
    the father of your father or mother
  </Definition>
  <SynsetRelation relType="hypernym"
                  target="example-en-10162692-n"/>
</Synset>
```

Interlingual Identifier

(Optional) Language

Like sense relations but applies to all synset members

Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

# Beyond Princeton WordNet: Wikipedia Linking

A World Leading SFI Research Centre

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN Science Foundation Ireland SFI For what's next

# Lexical vs. Encyclopedic

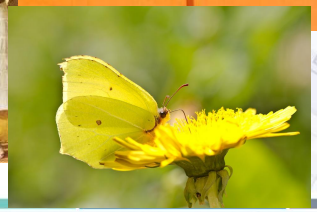
Yellow (in a dictionary)

- Is a verb, noun and adjective
- Secondary synonyms: cowardly, warning (especially in soccer)

Yellow (in an encyclopedia)

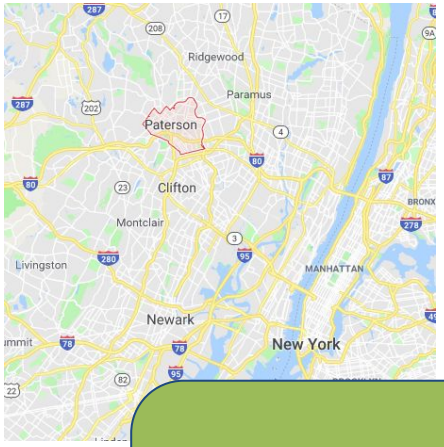
- A colour
- 2 books
- 8 Films or TV shows
- 4 songs
- A butterfly



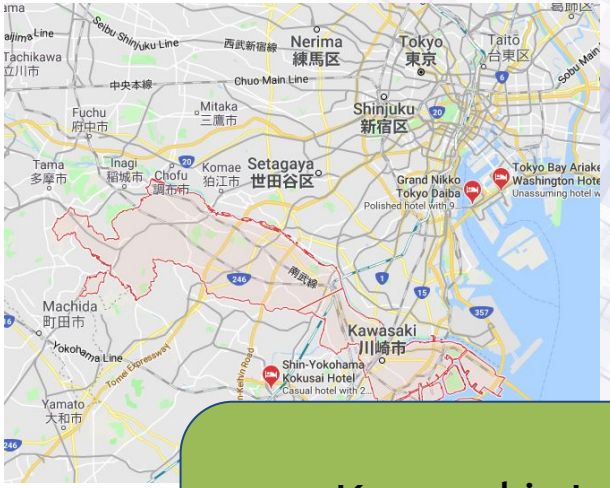


# Princeton WordNet

- Is a lexical resource with some encyclopedic information
- This information is quite biased to Anglo-Saxon, American and even North Eastern US context.




Paterson, NJ, USA
Pop: 147,000
In Princeton WordNet




Kawasaki, Japan
Pop: 1,500,000
Not In Princeton WordNet

# Wikipedia

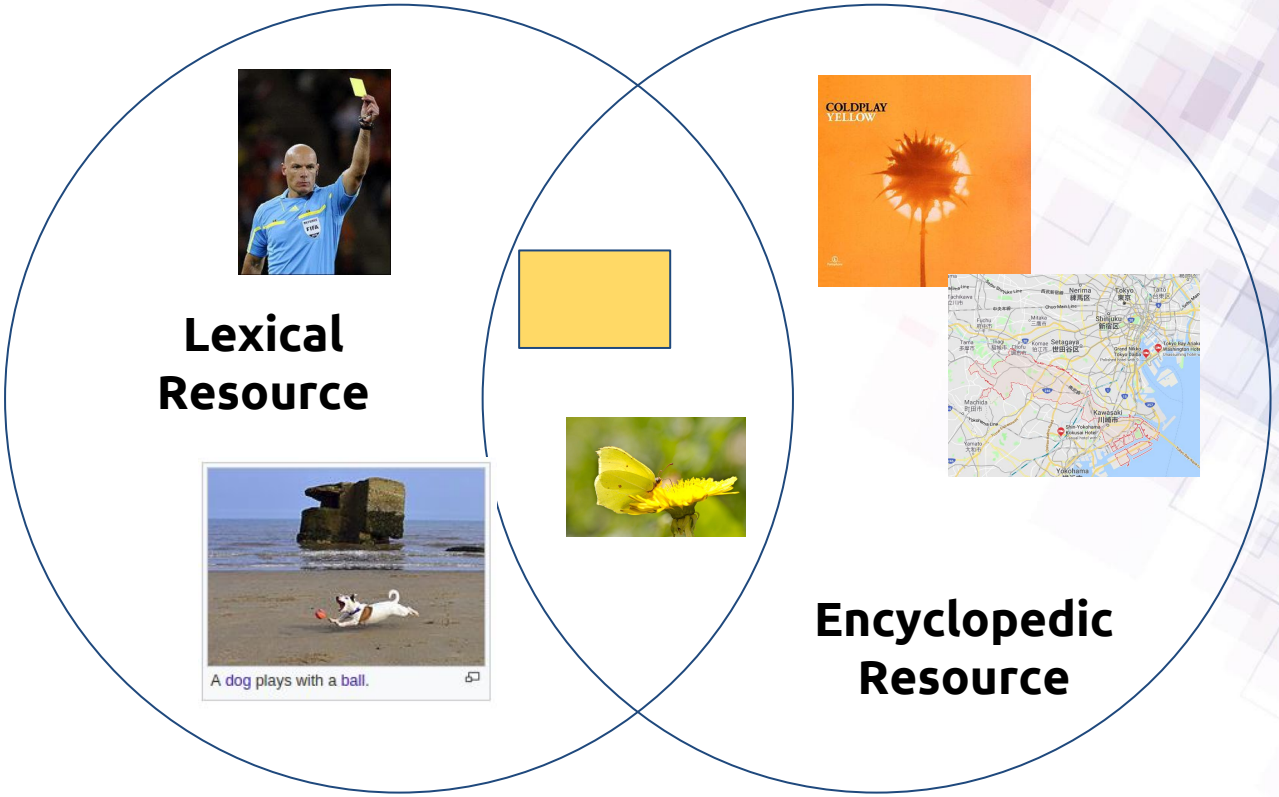
- Wikipedia is open and most-widely used encyclopedia
- Many lexical concepts are included, e.g.,
  - Play (activity)
    - https://en.wikipedia.org/wiki/Play_(activity)

A dog plays with a ball.

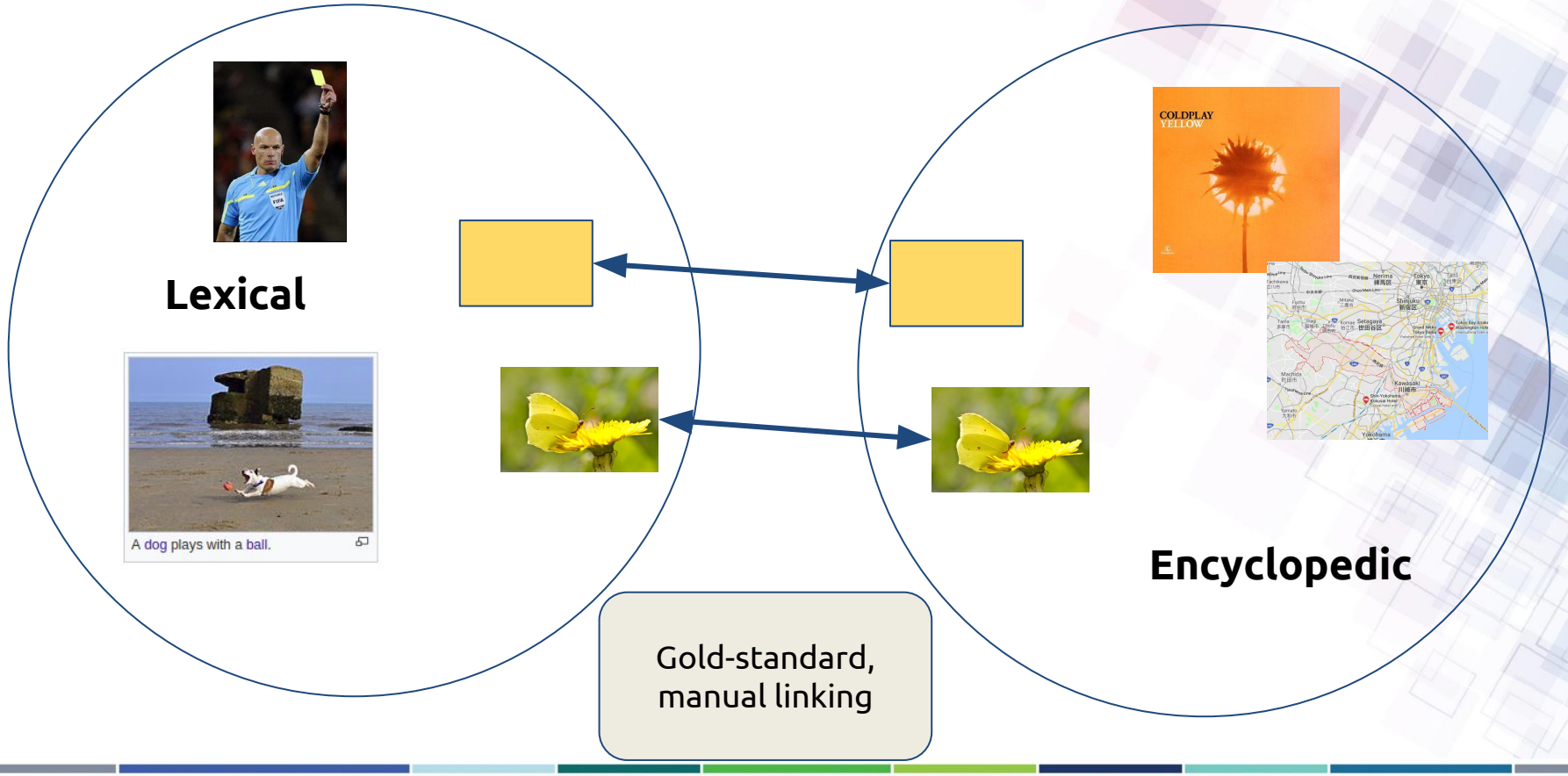
Cocker spaniel playing with a monkey doll

People having fun

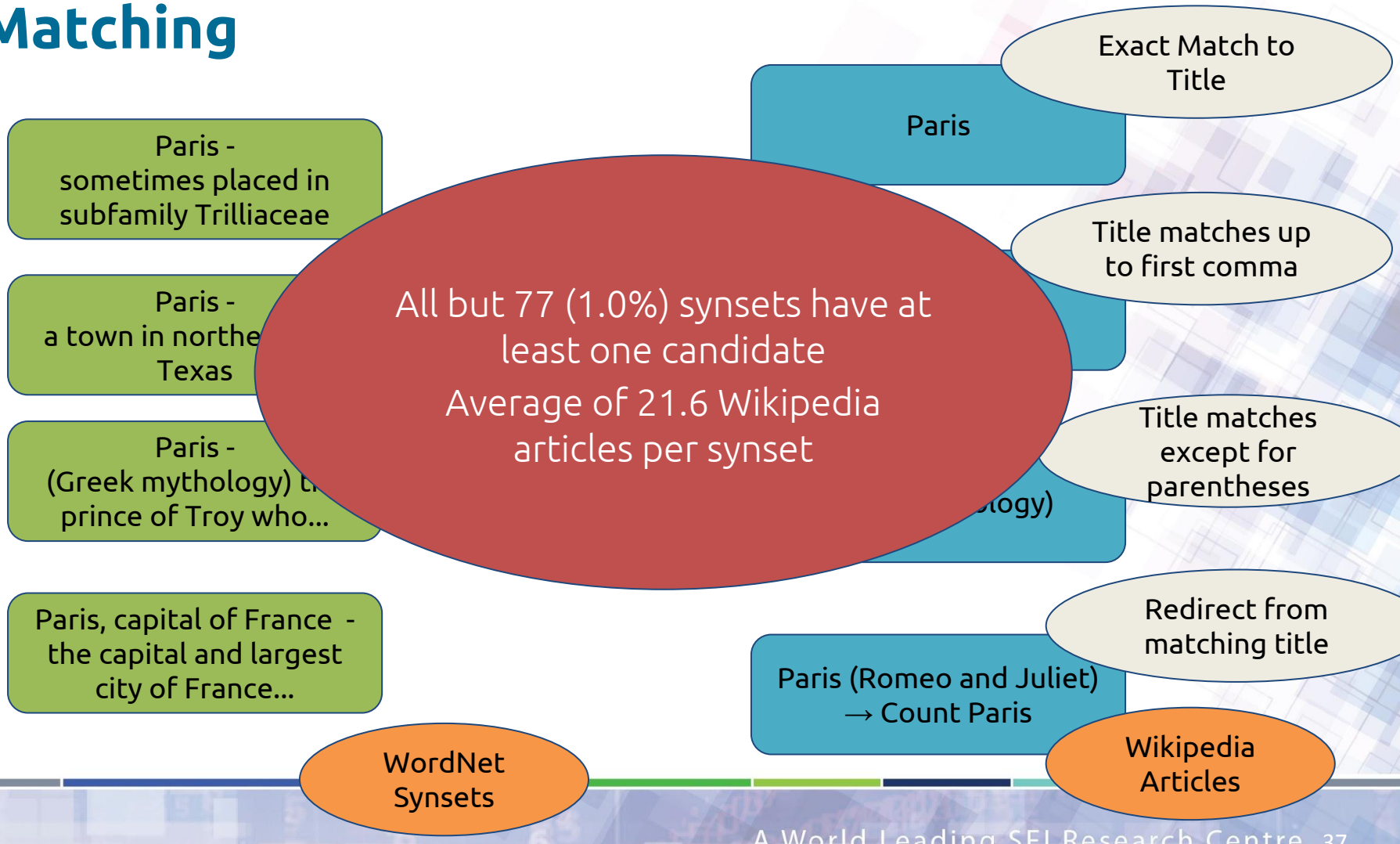
# Overlap between lexical and encyclopedic resources



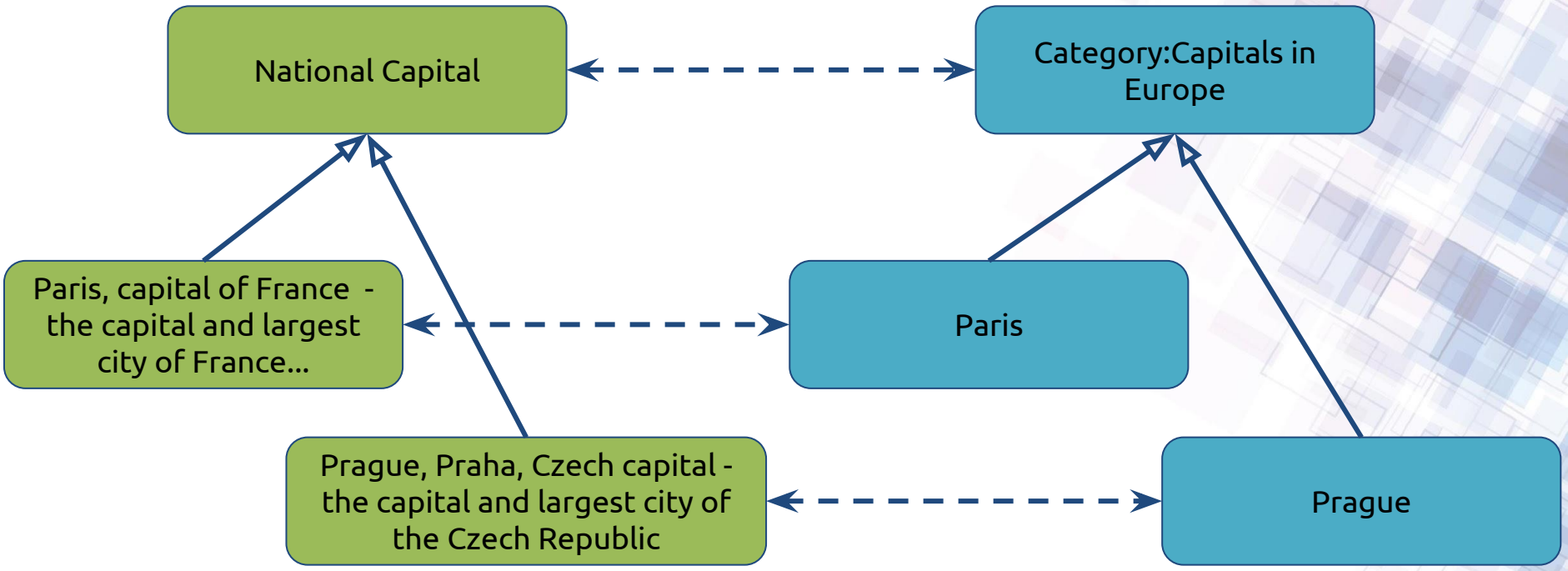
# Linking between resource types



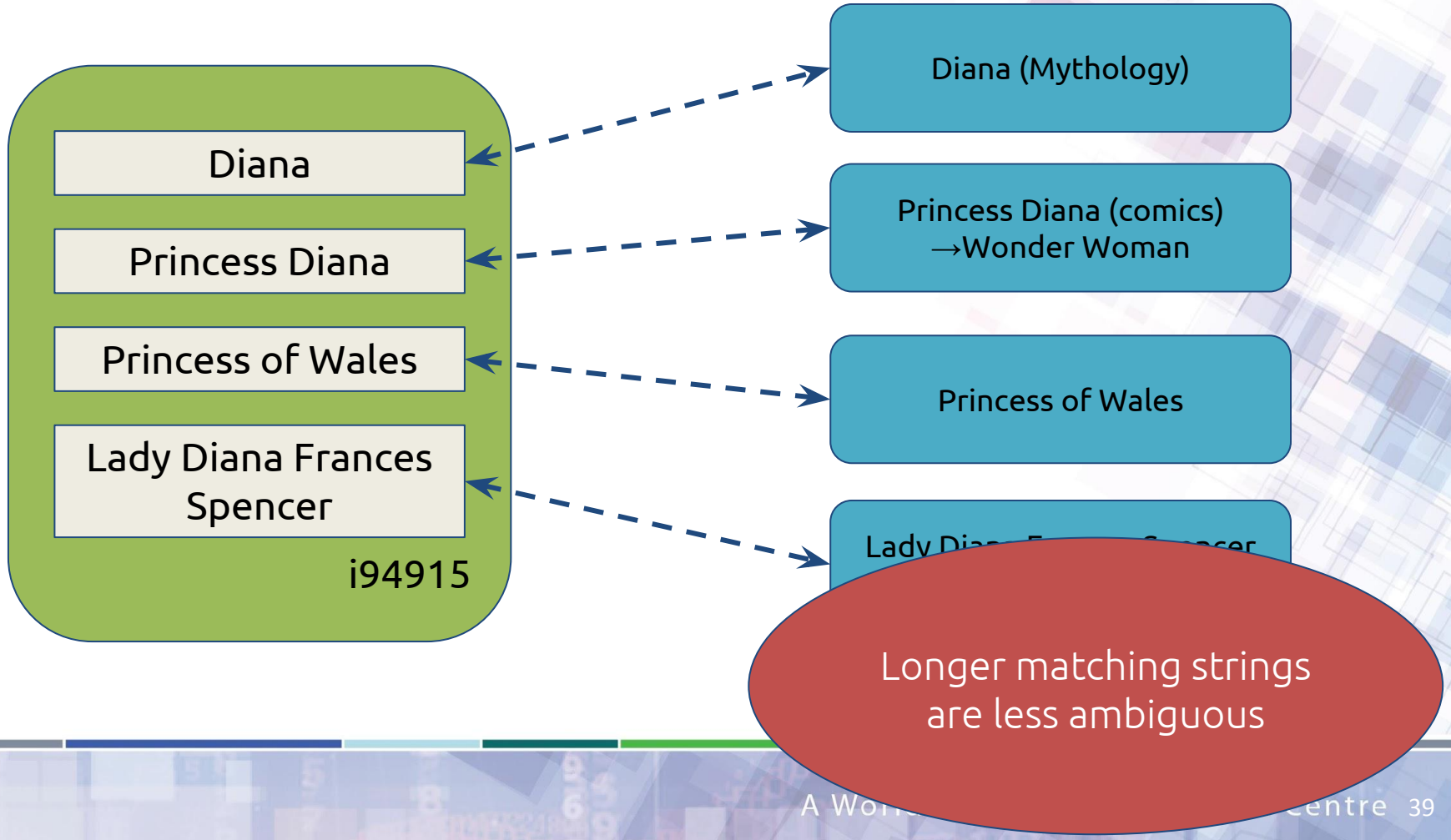
# Matching


Exact Match to Title
Paris
Paris - sometimes placed in subfamily Trilliaceae
Title matches up to first comma
Paris - a town in northe Texas
All but 77 (1.0%) synsets have at least one candidate
Average of 21.6 Wikipedia articles per synset
Title matches except for parentheses
Paris - (Greek mythology) t prince of Troy who...
ology)
Redirect from matching title
Paris, capital of France - the capital and largest city of France...
Paris (Romeo and Juliet) → Count Paris
WordNet Synsets
Wikipedia Articles

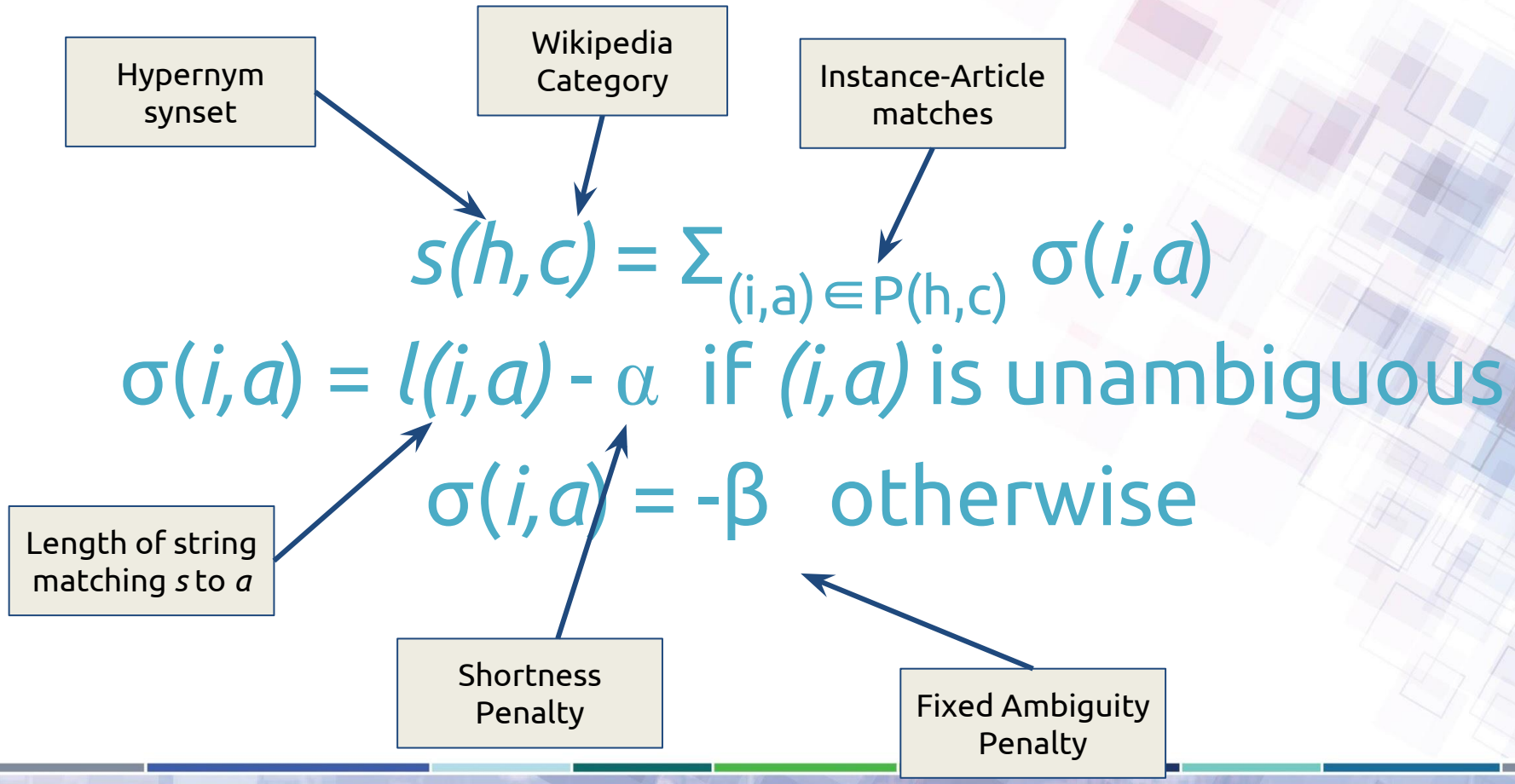
# Category Matches (based on Suchanek's YAGO)



# Length based matches



# Ranking Category matches



# WordNet-Wikipedia Mapping

Manually checked linking of 7,687 synsets to Wikipedia available at

https://github.com/jmccrae/wn-wiki-instances

# Beyond Princeton WordNet: Colloquial WordNet

# Princeton WordNet Release Schedule

Princeton WordNet is infrequently released

- Version 2.1: Mar 2005
- Version 3.0: Dec 2006
- Version 3.1: Nov 2012 (fewer synsets)

# English 12 years ago

Tweeting

Spreading


Exit

# English Now

Tweeting

Manspreading

Brexit

# Edit Entry

Lemma

delegitimize

Example

Trump and his cronies attempts to delegitimize the media should concern everyone .

Now will you admit goal has been to delegitimize 1st Black president ?

Spicer delegitimizes the press . But can you delegitimize 2 . 5million protesters ?

prior to taking office Trump has managed to delegitimize his own presidency far more than he ever

how the very media trump is attempting to delegitimize gave trump all the coverage in the world

RT On the radio ABC News tries to delegitimize Our President even after he says in two

RT The democrats continually try to delegitimize President Trump but they never fail to end

yes they want to delegitimize Palestinians and actually stole their heritage identity to

shows are fun and all but they ultimately delegitimize dance as a career

start planting forged emails as an attempt to delegitimize the leaks

Confidence

Very Strong | **Strong** | Medium | Weak | Skip

Status

**General** | Novel | Vulgar | Abbreviation | Misspelling | Inflected Form | Name | Not Lexical | Error

## Senses

Sense 1 ×

Part of Speech

Noun | **Verb** | Adjective | Adverb | Other

Synonym

# Lexicographers Wanted

http://colloqwn.linguistic-lod.org/

Empowering Citizens. Smarter Societies.

Insight
Centre for Data Analytics

# Beyond Princeton Wordnet: English WordNet

A World Leading SFI Research Centre

DCU NUI Galway OÉ Gaillimh UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh UCD DUBLIN Science Foundation Ireland SFI For what's next

# English WordNet

*"The English WordNet as individual files in GWN-LMF format. This Wordnet is an attempt to make it easier to provide feedback in the form of patches to the Princeton Wordnet. It is undergoing a first trial in Summer 2018."*

Aiming for a 3.2 release this year

## Added one simple typo patch.

Browse files

master (#4)

**daikinomura** committed 3 days ago  1 parent c0e1571  commit d7863562b3c0822eb7b75e81a40218121e1e29a6

Showing **1 changed file** with **1 addition** and **1 deletion**.

Unified | Split

2 src/wn31-noun.person.xml  View

```
@@ -116666,7 +116666,7 @@
116666 116666       </Synset>
116667 116667       <!-- house painter -->
116668 116668       <Synset id="ewn-10208798-n" ili="i90596" partOfSpeech="n" dc:subject="noun.person">
116669         -       <Definition>a painter of houses a similar buildings</Definition>
       116669   +       <Definition>a painter of houses or similar buildings</Definition>
116670 116670         <SynsetRelation relType="hypernym" target="ewn-10413608-n"/> <!-- painter -->
116671 116671       </Synset>
116672 116672       <!-- resident physician, house physician, resident -->
```

# Summary

# Summary

- OntoLex was a model designed for the Semantic Web
  - Has found usage in a wider lexicographic context
- One usage is in the Global WordNet Grid
  - Part of defining globally unique interlingual identifiers for all concepts
- The XML/JSON format of this is a good general model for wordnets
- Enables new projects
  - Wikipedia linking
  - New WordNets extending the Princeton WordNet
  - Open source projects based on Princeton WordNet