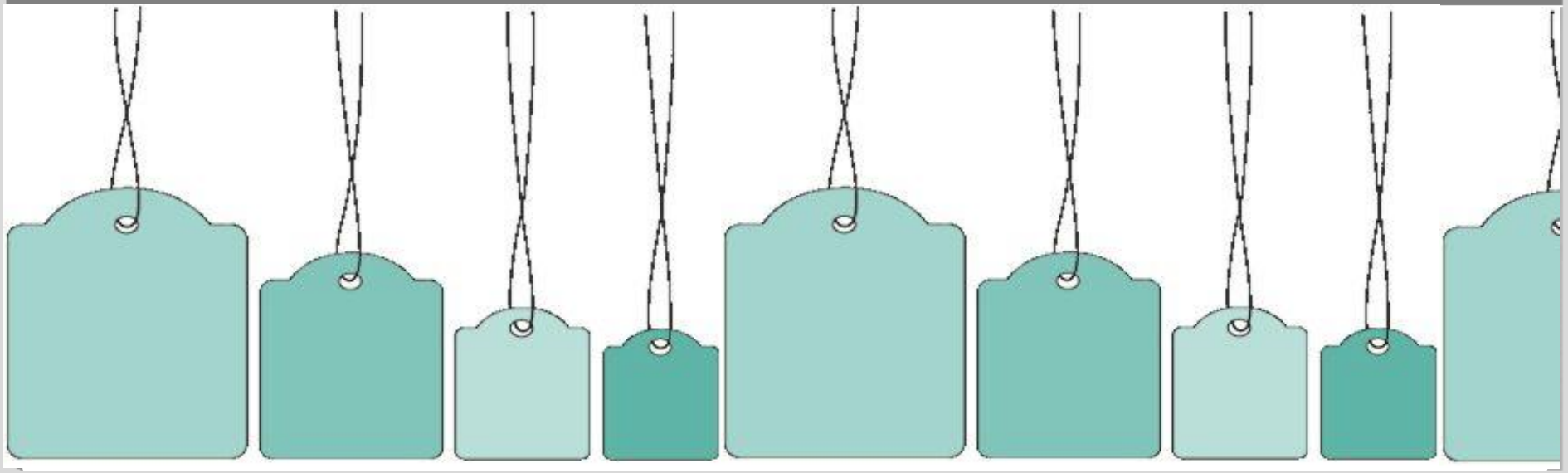
Karlsruhe Institute of Technology

RENDER
Reflecting Knowledge Diversity

Labels in the Web of Data

Basil Ell, Denny Vrandečić, and Elena Simperl
10th International Semantic Web Conference, Bonn
26 October 2011

INSTITUTE FOR APPLIED INFORMATICS AND FORMAL DESCRIPTION METHODS



Main message in a nutshell

- Labels necessary but often missing (62%) or problematic
- Findings relevant for linked data publishers & consumers
- Relevant for front-end tools (linked data browsers, semantic web search engines)

Outline

- Motivation
- Related Work
- Challenges
- Labeling properties
- Metrics
- Results
- Guidelines
- Conclusions

MOTIVATION

Motivation

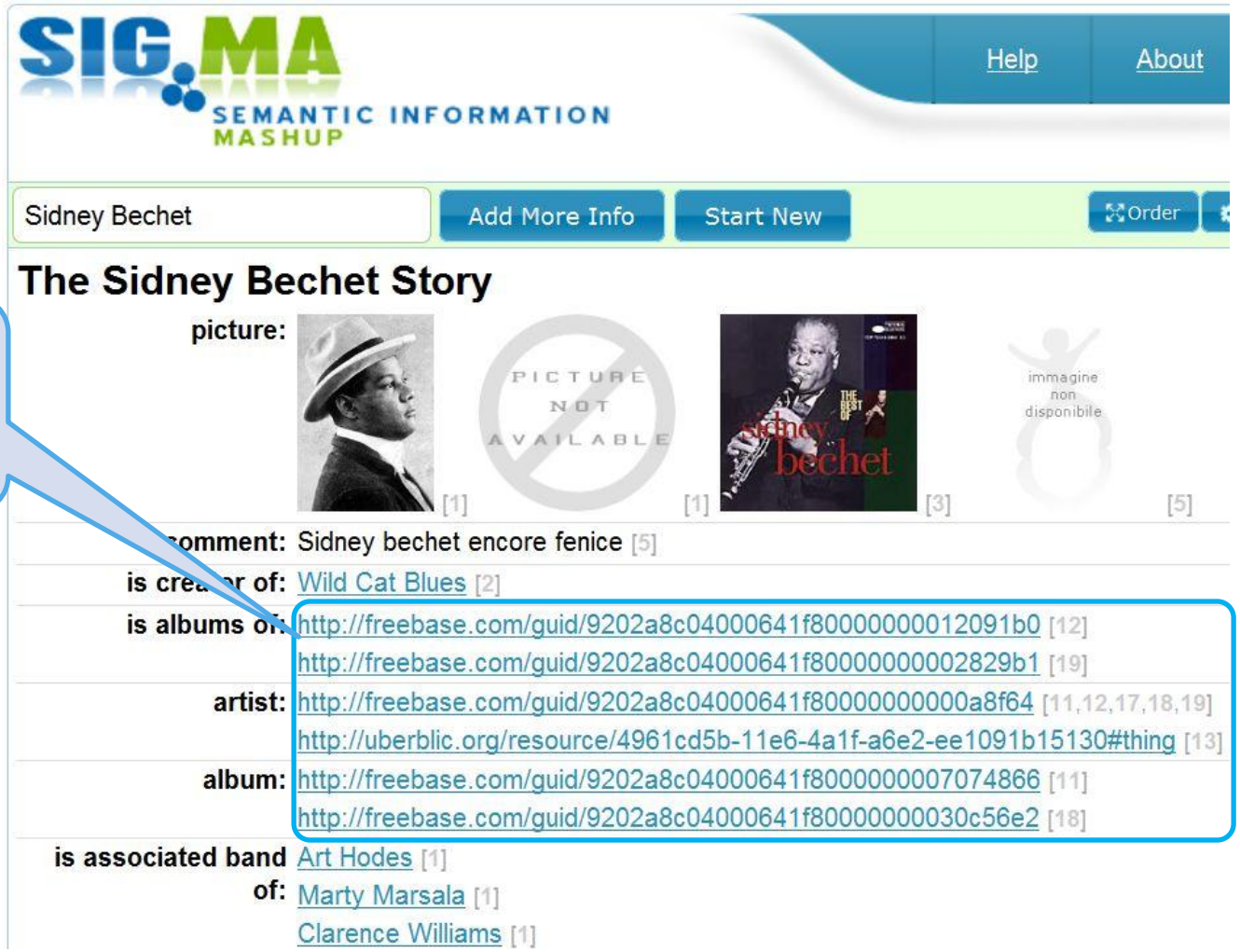
- Where labels are necessary
  - Displaying labels instead of URIs to end-users
  - Searching over the Web of Data
  - Document annotation

KIT – Karlsruhe Institute of Technology
Institute for Applied Informatics and Formal Description Methods

Motivation

Scenario: linked data browsing

[SIGMA]



Is this meaningful to users?

RELATED WORK

Related work (1/3)

- Linked data browsers – how they deal with the problem of missing labels
  - Display URI
  - Display last part of URI
  - Let user select labeling properties
- Linked data summarization & verbalization

Related work (2/3)

- Semantic search engines such as Falcons, Sindice, MicroSearch, Watson, SWSE, Swoogle provide keyword-based searches
  - Rely on existence of nodes that are labeled or on meaningful URIs

Related work (3/3)

- [Azlinayati et al.] analyzed identifiers and labels in 219 ontologies
  - Teminological data
  - Web of Data mainly consists of instance data

CHALLENGES

Challenges

- Multitude of labeling properties
- Missing labels
- Label selection & ambiguity
- Multilinguality

LABELING PROPERTIES

Labeling properties

http://www.w3.org/2000/01/rdf-schema#label
http://xmlns.com/foaf/0.1/nick
http://purl.org/dc/elements/1.1/title
http://purl.org/rss/1.0/title
http://xmlns.com/foaf/0.1/name
http://purl.org/dc/terms/title
http://www.geonames.org/ontology#name
http://xmlns.com/foaf/0.1/nickname
http://swrc.ontoware.org/ontology#name
http://sw.cyc.com/CycAnnotations_v1#label
http://rdf.opiumfield.com/lastfm/spec#title
http://www.proteinontology.info/po.owl#ResidueName
http://www.proteinontology.info/po.owl#Atom
http://www.proteinontology.info/po.owl#Element
http://www.proteinontology.info/po.owl#AtomName
http://www.proteinontology.info/po.owl#ChainName
http://purl.uniprot.org/core/fullName
http://purl.uniprot.org/core/title

BTC2010 data

36 labeling properties identified in 3,167,799,445 ntriples.

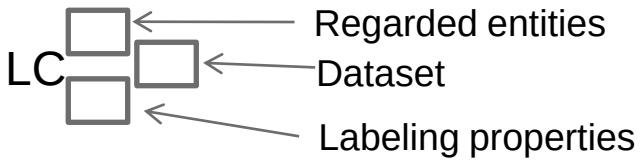
METRICS

Metrics

1. Completeness
2. Efficient accessibility
3. Unambiguity
4. Multilinguality

Completeness

- All non-information resources should have labels
- Labeling completeness metric LC
- Ratio of regarded entities with at least one label
- Notation:



LC(D) should be 1

1) $LC_{rdfs}^{NIR}(D)$

2) $LC_{rdfs+}^{NIR}(D)$

3) $LC_{lp}^{NIR}(D)$

Efficient accessibility (1/2)

- URIs without labels can be dereferenced
- Example

```
ex:Bonn ex:location ex:Germany .
ex:Bonn rdfs:label "Bonn" .
```

  - Need to dereference `ex:location` and `ex:Germany` before displaying first triple
- LE: ratio of all mentioned URIs with at least one label

Efficient accessibility (2/2)

- Metric parameter for a set of entities with known labels (FOAF, GoodRelations, ...)
- Example

```
ex:Basil foaf:img ex:basil.jpg
ex:Basil rdfs:label "Basil"
```

$$LE_{rdfs}^{foaf^-}(D)=1 \qquad LE_{rdfs}^{-}(D)=0.5$$

- LE should be 1

KIT – Karlsruhe Institute of Technology
Institute for Applied Informatics and Formal Description Methods

Unambiguity

- An entity can have multiple labels (e.g. synonyms) that are not differentiated (e.g. by language)
- LU_f is the ratio of all entities that have exactly one preferred label according to a selection procedure *f*
- Example

```
ex:loc rdfs:label „place“ .
ex:loc rdfs:label „location“ .
```

- LU_f should be 1

Multilinguality

- Language tags used on literals to state their natural language
- Display according to user's language preferences
- Example

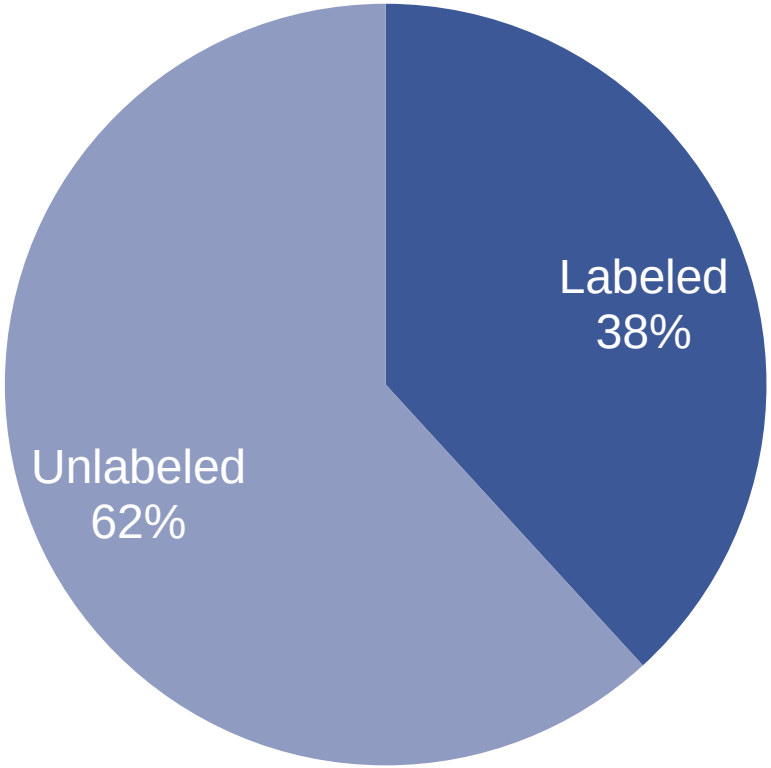
  `"Bonn"@en or "  "@ko or  "Bonna"@la`
- LLN: number of label languages
- LLC^{lang}: completeness for language *lang*

Measurements on BTC 2010 corpus

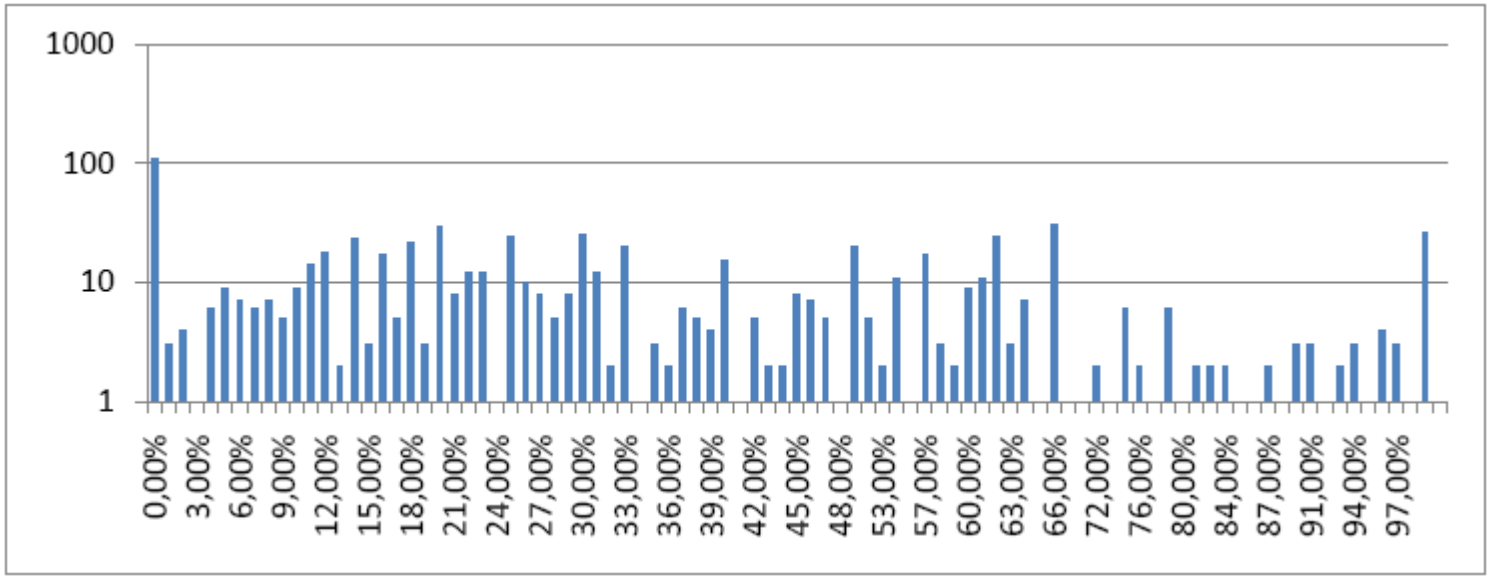
RESULTS

Results: Completeness

Non-information resources



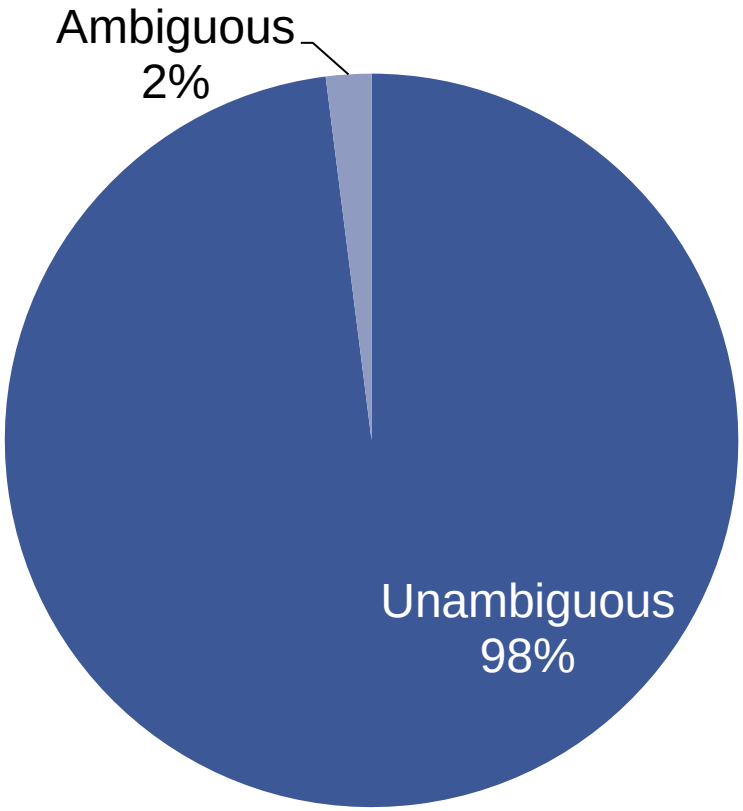
Results: Efficient accessability



741 data sets, about 5 data sets per PLD
top: 10 most occurring vocabulary namespaces

LE^{top}_{BTC}

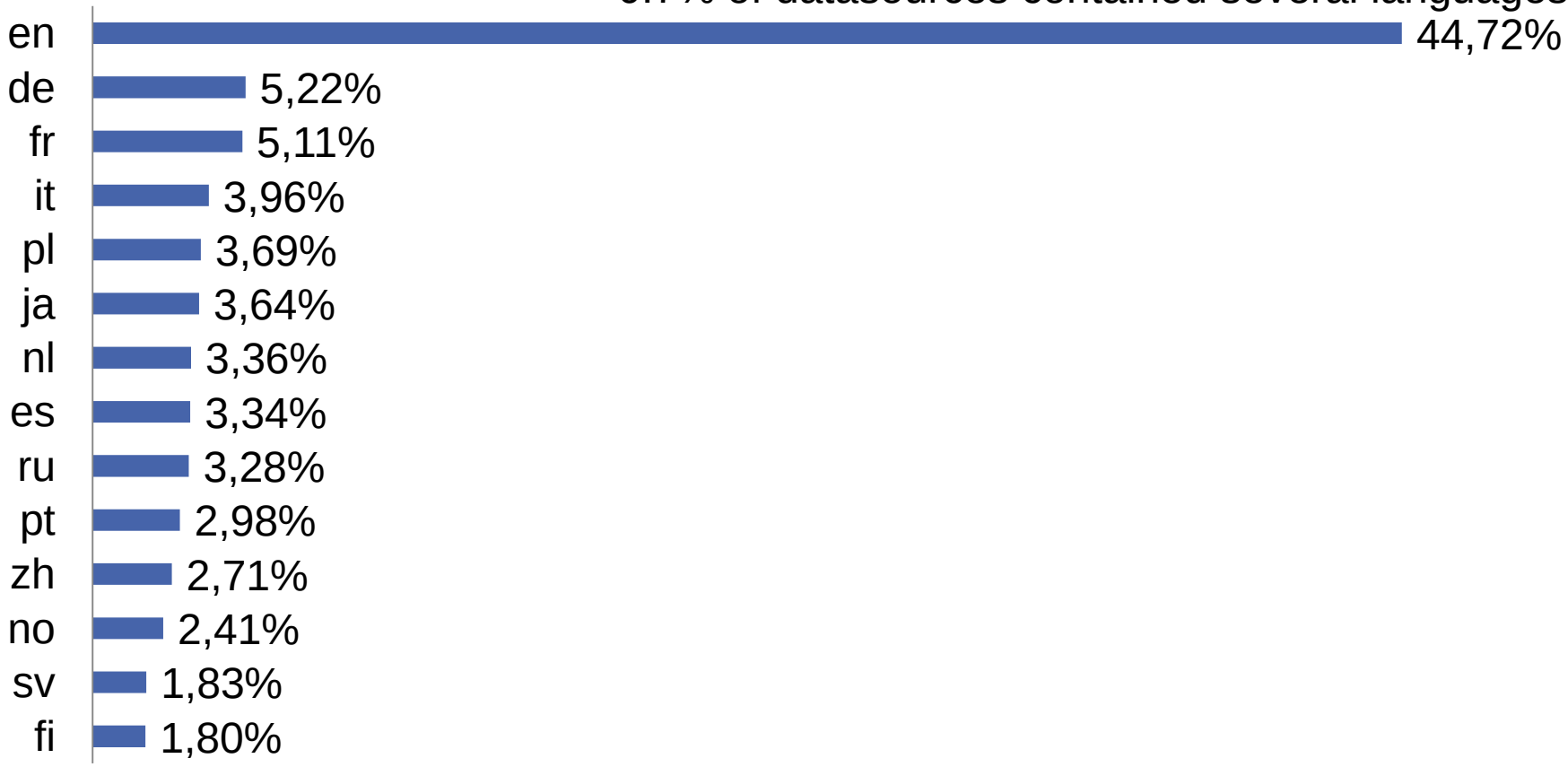
Results: Unambiguity



Unambiguity rate of 0.98

Result: Multilinguality

4.78% of NIR labels have language tag
2.2% of datasources contained one language
0.7% of datasources contained several languages



GUIDELINES

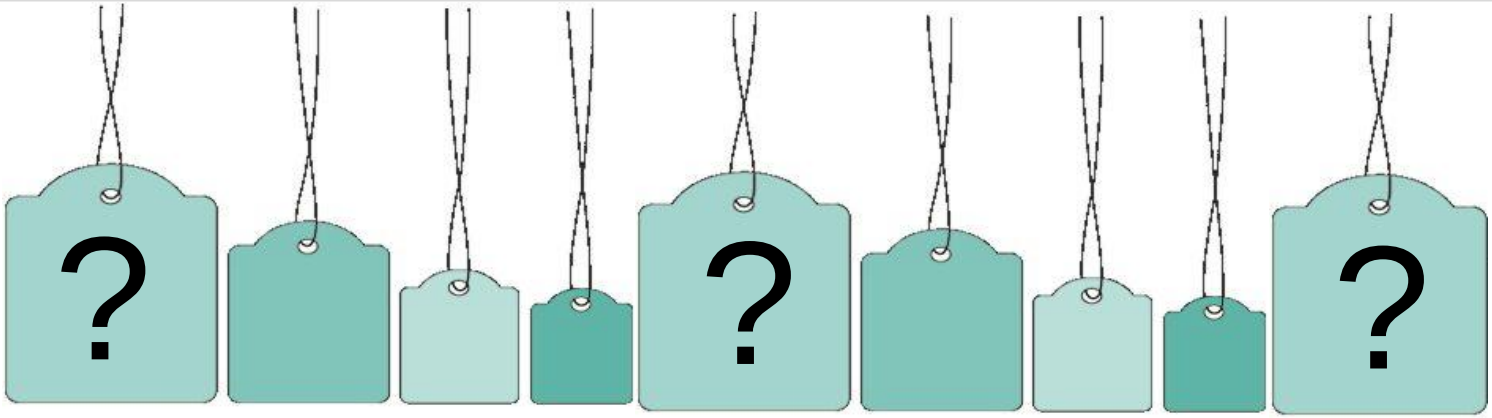
Guidelines

- Provide labels for all URIs mentioned in a given RDF graph
- Provide a complete set of labels in all supported languages
- Subproperty your labeling properties with `rdfs:label`
- Do not provide more than one preferred label for each URI

Conclusions

- Defined four parameterizeable metrics
- Suggested guidelines for labeling
- Many problems due to LOD principles
- Solution: serving data via SPARQL?
  - Application can exactly specify its need

- Labeling is essential for the Web of Data to become widely used

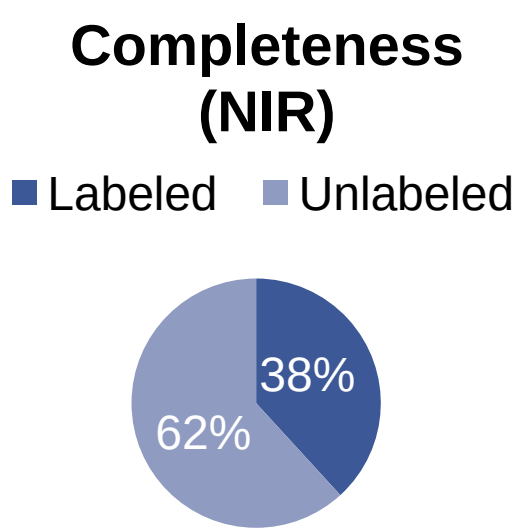


Thank you for your attention

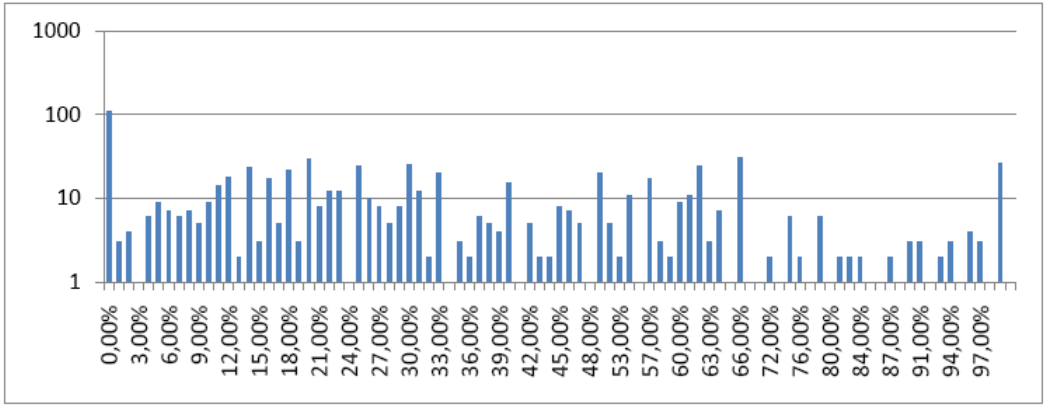
QUESTIONS?

Conclusions

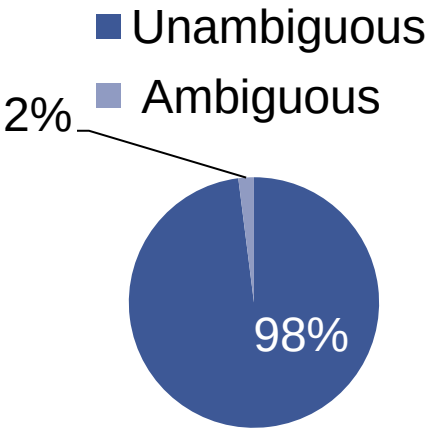
Completeness (NIR)



Efficient accessability



Unambiguity



Multilinguality