# Very Large Cross-lingual Resources at OAEI 2008

Laura Hollink
Véronique Malaisé
Vrije Universiteit Amsterdam

# Task description

- Mappings between three resources:
  - GTAA
    - Thesaurus of the Dutch Institute for Sound and Vision.
    - 4 Facets: Subject, Location, People, Name
  - WordNet
    - Lexical database
  - DBPedia
    - ‘structured version of Wikipedia’
- 3 sets  of mappings
- exactMatch, broadMatch, narrowMatch.

# Rationale

- Different languages
  - Archive with GTAA metadata in Dutch only
  - Broaden user group
  - Integrated access to archives of other countries?
- Large resources
  - Disambiguation becomes serious problem
- Heterogeneous resources
  - Different structure
  - Weak or inconsistent structure
  - Large parts have no counterparts, when to stop mapping?

# Cross-Lingual

- GTAA in Dutch:
  - Preferred labels
  - Alternative labels
  - Scope notes
- WordNet in English
  - Word-senses
  - Glosses
- DBPedia in English and (most of the time) Dutch
  - Titles
  - Abstracts

# Different Schema's

- GTAA in SKOS
  - Skos:concepts with pref- and altLabels
  - Narrower/Broader relations between the concepts
- WordNet
  - Synsets with word-senses
  - Hyponym relations between synsets.
- DBPedia
  - Things with titles and abstracts
  - links to dbpedia categories, rdf:type links to yago classes
  - Hierarchical structure of yago classes and categories.

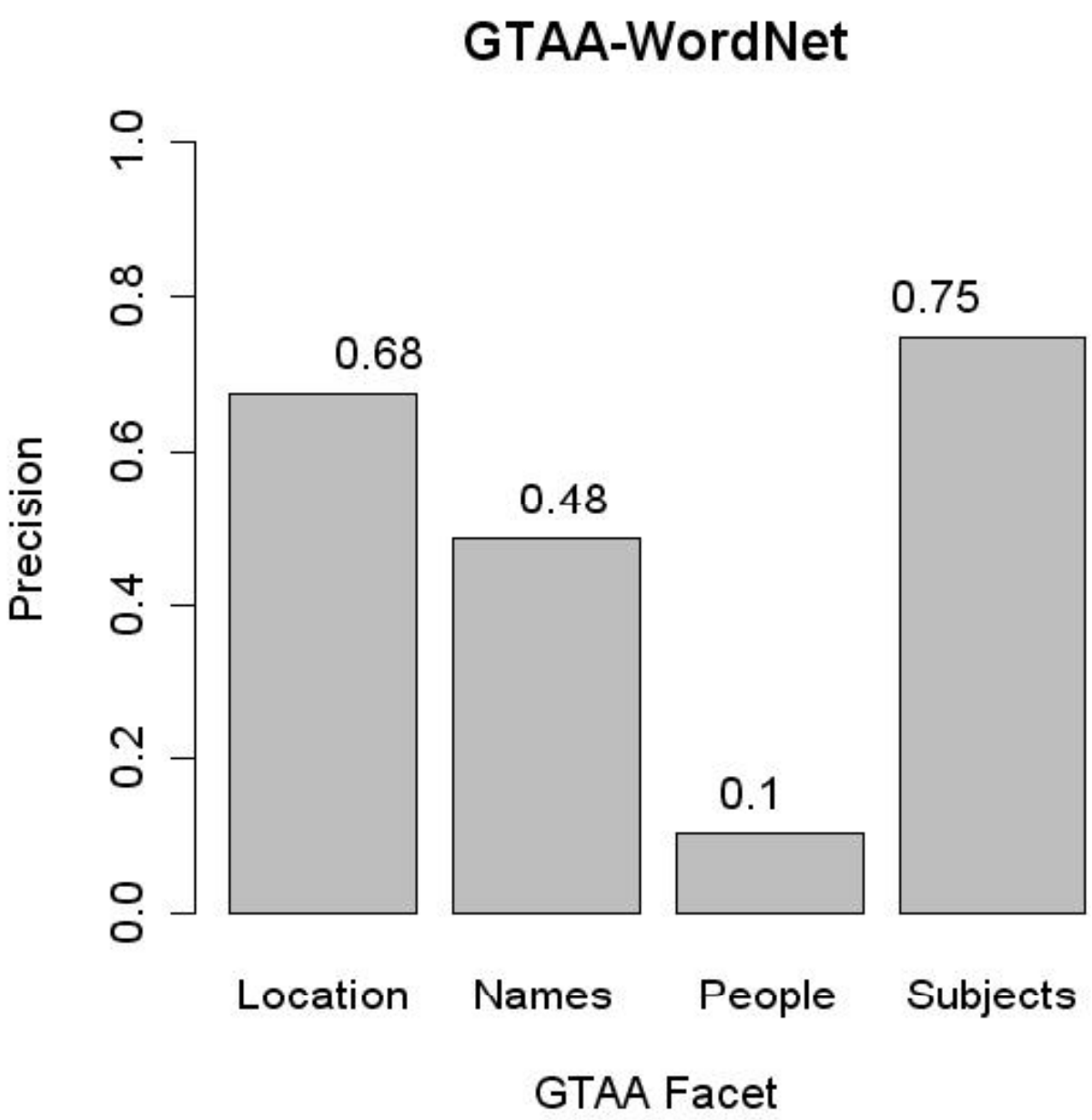
# Results

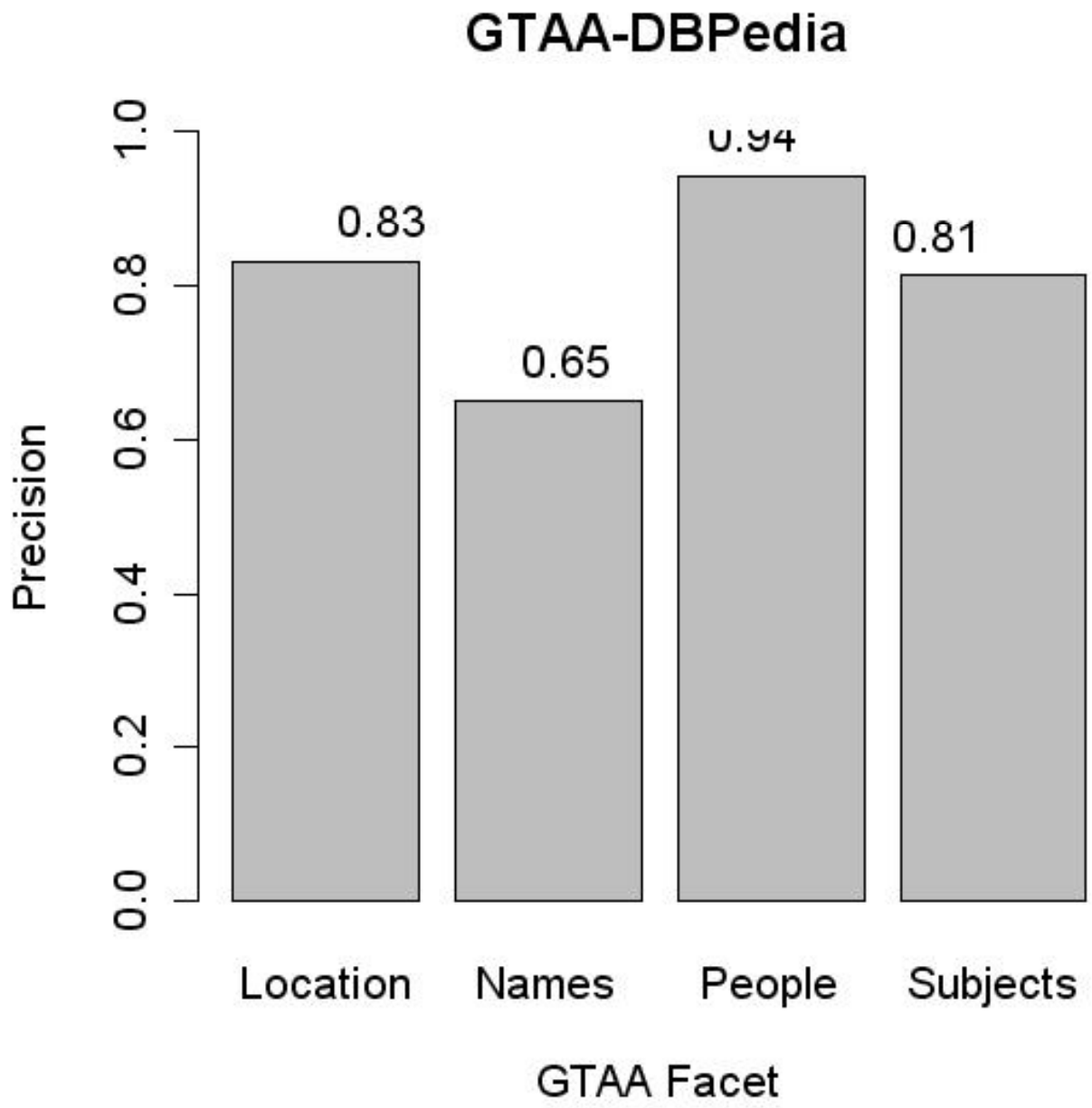
- One Participant: DSSIM

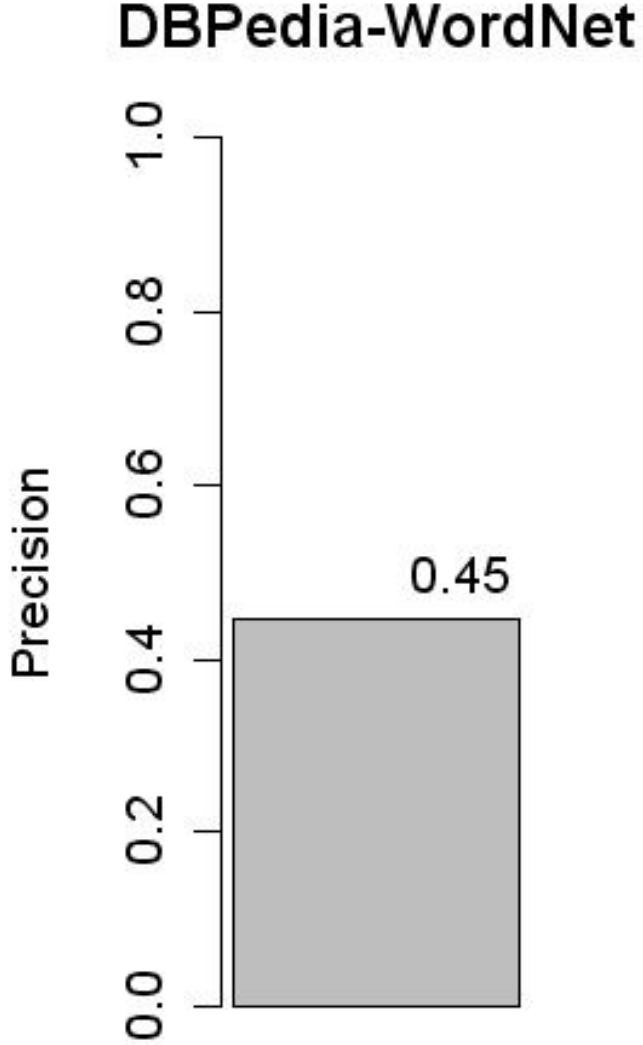
| Vocabulary | #Concepts | #Mappings to | | |
|---|---|---|---|---|
| | In Vocabulary | WordNet | DBPedia | GTAA |
| WordNet | 82,000 | NA | 28,974 | 2,405 |
| DBPedia | 2,700,000 | 28,974 | NA | 13,156 |
| GTAA | 160,000 | 2,405 | 13,156 | NA |
| Subject | 3,800 | 655 | 1,363 | NA |
| People | 97,000 | 82 | 2,238 | NA |
| Name | 27,000 | 681 | 3,989 | NA |
| Location | 14,000 | 987 | 5,566 | NA |

# Evaluation Process

- Precision
  - GTAA-WordNet & GTAA-DBPedia
    - Inspection of 100 mappings per GTAA facet.
  - WordNet-DBPedia
    - Inspection of 100 mappings
  - Correct or Incorrect or Narrow/Broader/Related
- Recall
  - GTAA-WordNet & GTAA-DBPedia
    - Comparison to a reference alignment of 100 GTAA People and 100 GTAA Subjects

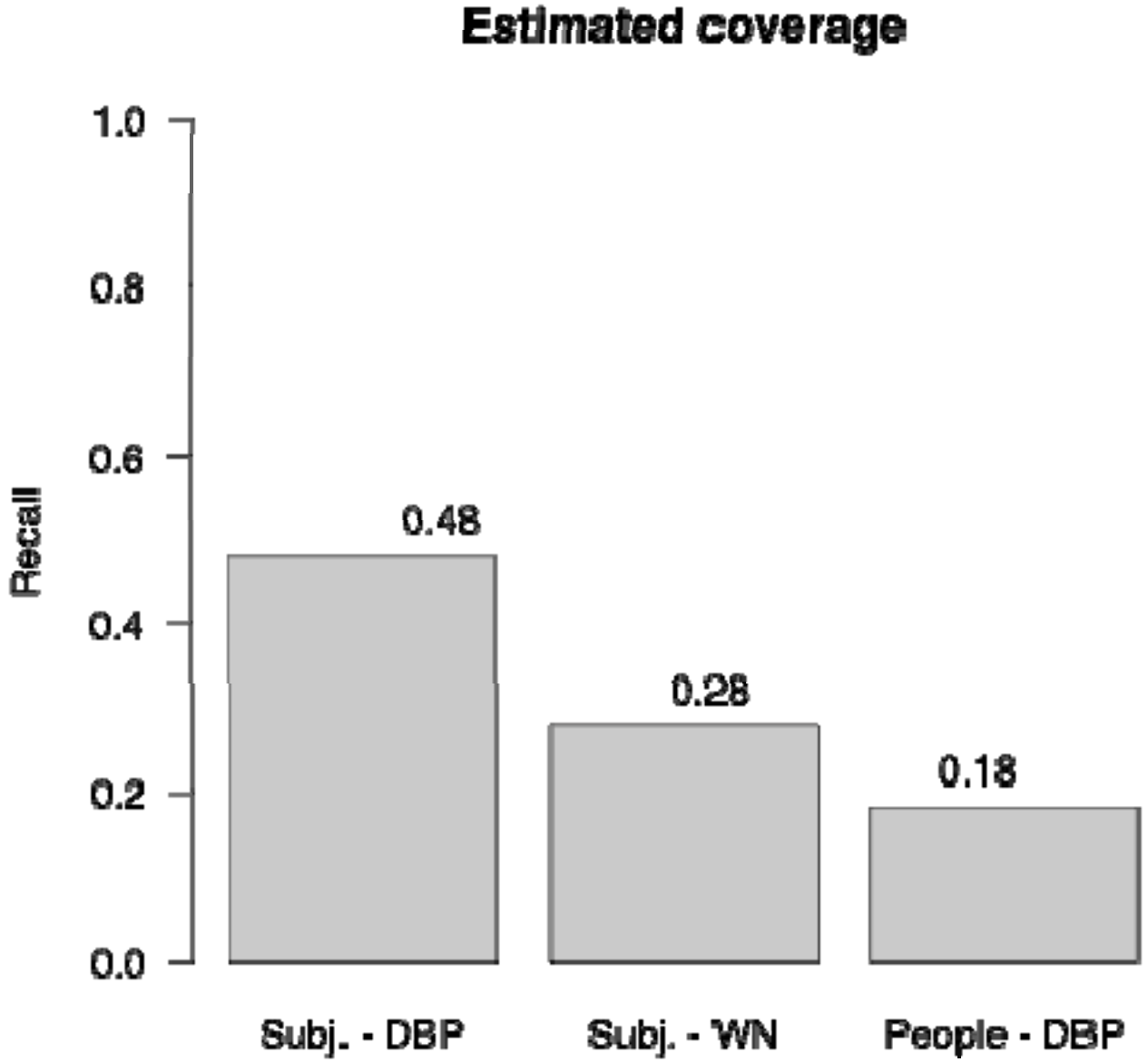

GTAA-WordNet
Precision
0.0
0.2
0.4
0.6
0.8
1.0
0.68
0.48
0.1
0.75
Location
Names
People
Subjects
GTAA Facet

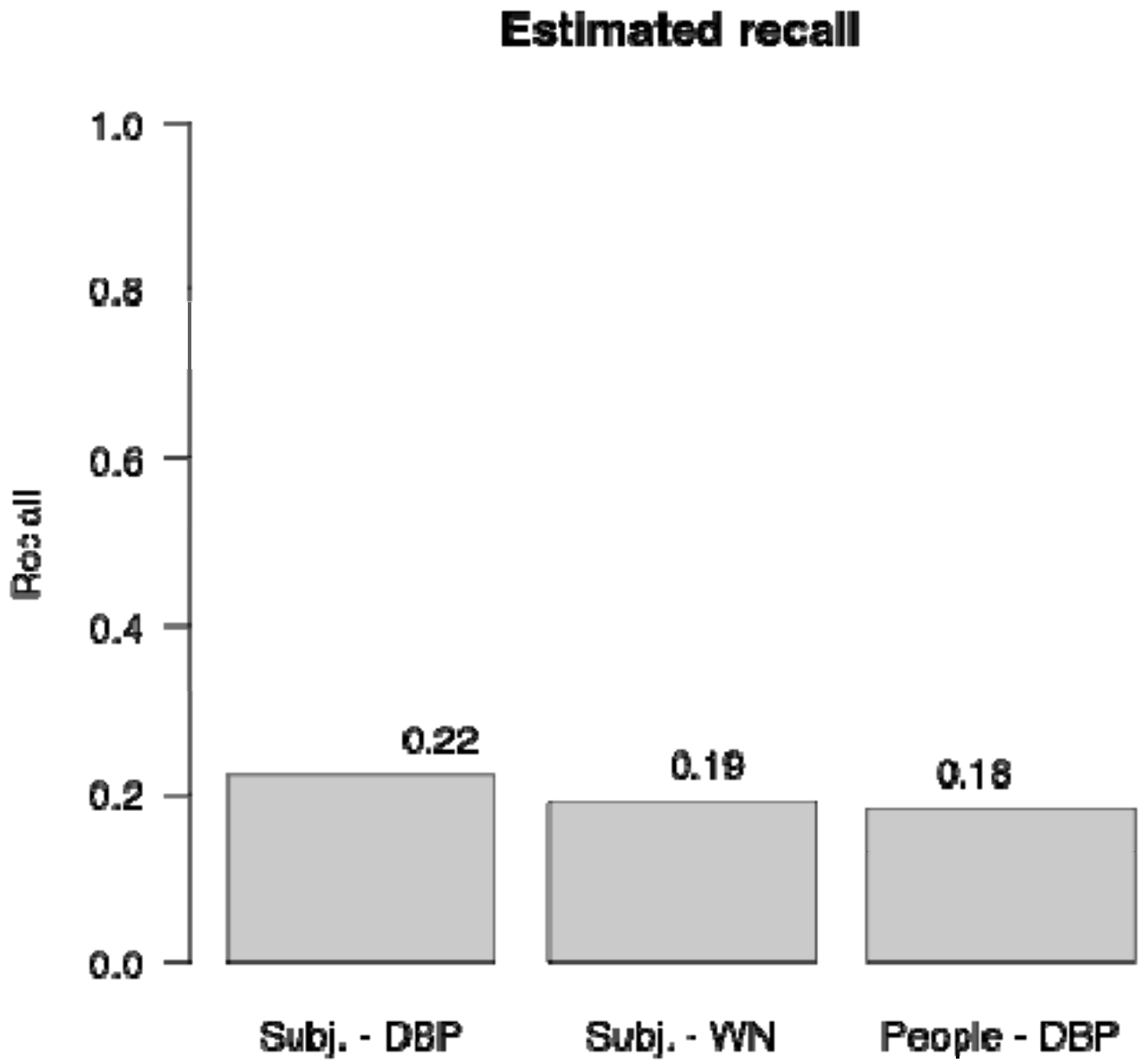

GTAA-DBPedia
Precision
0.0
0.2
0.4
0.6
0.8
1.0
0.83
0.65
0.94
0.81
Location
Names
People
Subjects
GTAA Facet


DBPedia-WordNet
Precision
0.0
0.2
0.4
0.6
0.8
1.0
0.45

# Pre-existing WN-DBPedia Mappings

- Type Links
  - Air_New_Zealand  wordnet-type  synset-airline-noun-2
  - No overlap between DSSIM mapping and wordnet-type links.
  - DSSIM mappings of things with a wordnet-link performed less than DSSIM mappings of things without a wordnet-link.
- Yago
  - most DBPedia "things" are instances in the YAGO ontology.
  - Dbpedia categories are classes in the YAGO ontology, subclasses of wordnet synsets.
  - “Crazy”_Joe_Davola rdf:type  FictionalCharacter
- Overlap between DSSIM results and Yago?
- We are looking (mainly) for exactMatches, not type links.


Estimated coverage
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.48
0.28
0.18
Subj. - DBP
Subj. - WN
People - DBP


Estimated recall
Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.22
0.19
0.18
Subj. - DBP
Subj. - WN
People - DBP

# Conclusions

- Also other types of links than exactMatch are necessary:
  - Subject:pausbezoeken -> List_of_pastoral_visits_of_Pope_John_Paul_II_outside_Italy.
  - Location:Venezuela -> synset-Venezuelan-noun-1
  - Subject:Verdedigingswerken -> fortification

# Conclusions

Context would help a lot.

- GTAA facet information <skos:inScheme>
  - Person:GoghVincentvan -> synset-vacationing-noun-1
  - Location:Harlem -> synset-hammer-noun-8
  - Location:Melbourne -> synset-Melbourne-noun-1
- Titles of DBPedia ‘referring pages’ used as alternative labels.
    - Person:SummerGordon -> Sting_(musician)
- But: no longer a generally applicable tool.

# Reflection on the evaluation process

- Disambiguation of DBPedia/WordNet concepts very hard, also for evaluator.
  - Subject:leguanen -> Iguana or Iguanidae?
  - Multiple mappings are reasonable.
- When is a mapping ‘Related’?
- DBPedia disambiguation pages.
- GTAA contains may Dutch-specific concepts
  - Diogenes = TV program
- Take into account the confidence measures

# Input from you

- What do OAEI participants think of this task?
- How can we improve it?

# Why Not?

- Too much preprocessing required?
- The three resources have different schema’s?
- Tool was not built for resources that large?
- Tool was not built for multi-lingual matching?

Thank you!