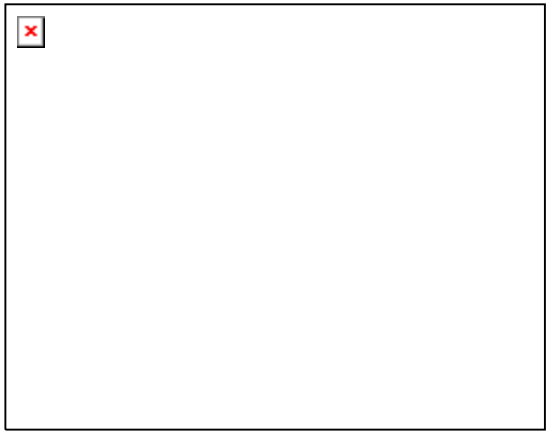
Semantically Enriching Folksonomies with



Sofia Angeletou, Marta Sabou
and Enrico Motta

NeOn

KNOWLEDGE MEDIA INSTITUTE KMi

The Open University

Semantic Web2.0

"The combination of Semantic Web formal structures and Web2.0 user generated content can lead the Web to its full potential".

Web2.0

- easy upload
- free tagging
  - requiring minimal annotation effort
- open, dynamic and evolving vocabulary
- .. leading to a content intensive web

...however..

tagging systems' characteristics

- content retrieval mechanisms are limited:
  - keyword based search
  - tag cloud navigation
- search may suffer of poor precision and recall due to:
  - basic level variation problem
    - whale VS orca
  - syntactic inconsistencies
    - singular VS plural
    - concatenated/misspelled tags

..an example

- looking for flickr images of
  "animals which live in the water"

5/24 ≈ 21% relevant

.. some missed photos

dolphin

From f0rbe5

whale

From gerb

dolphin

From dobar

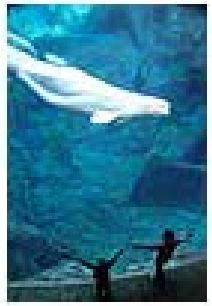
From ichie

dolphin

From *rosacelo*

whale

From ScottS101



From Markus

whale

From ScottS101

sea elephant

From mtchm

seal

From _rebekka

whale

From f0rbe5

modifying the query..

- "animal habitat water"
- "animal sea"
- "animal water"

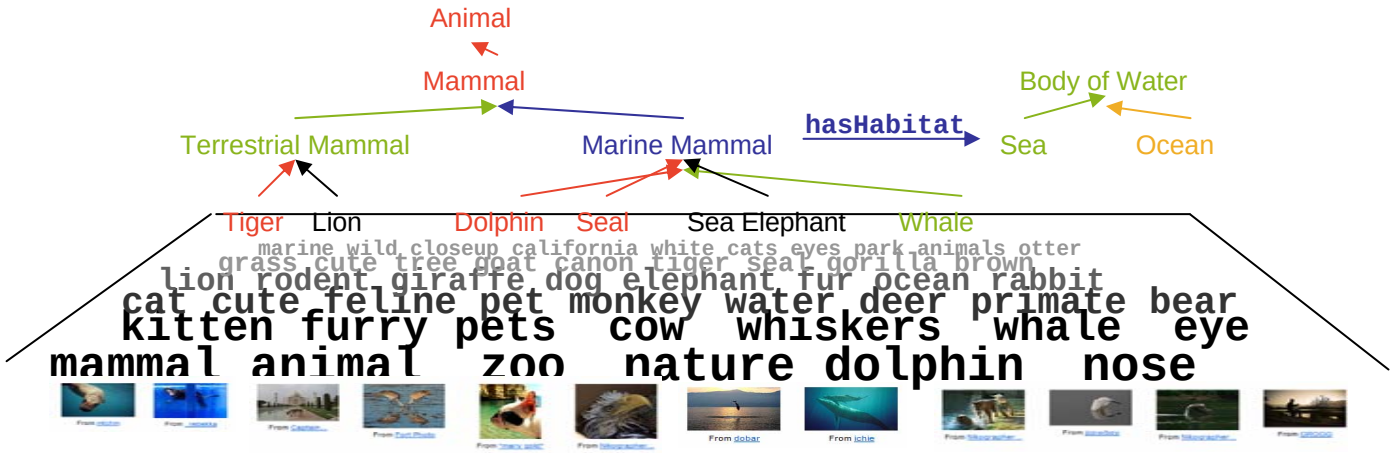
- similar results

...also:

- not easy for the user to form the most effective query

our goal

- Improve content retrieval in folksonomies
  - enhance precision and recall in search
  - enable complex queries
  - support intelligent navigation
- by applying a semantic layer on top of folksonomy tagspaces

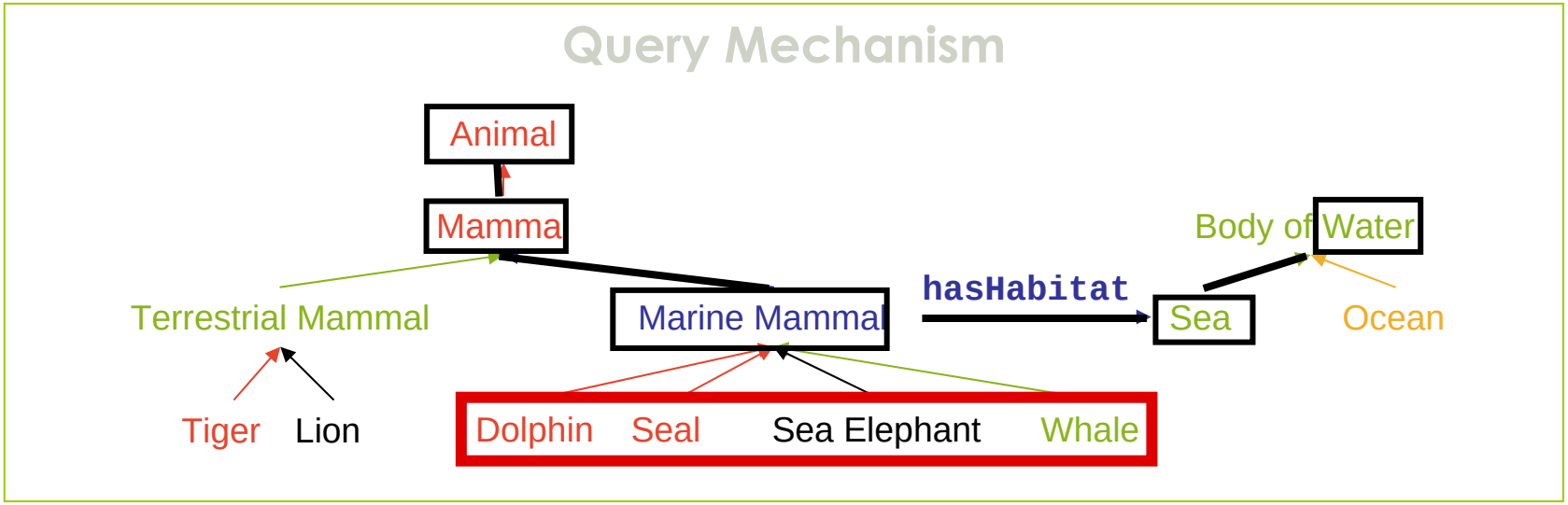


our goal

STEP1: Semantically Enriching Folksonomies

our goal

STEP2: Querying Folksonomies through the Semantic Layer



“Dolphin OR Seal OR Sea Elephant OR Whale”

21/24 ≈ 87% relevant

existing work on folksonomy enrichment

- tag clustering based on co-occurrence frequency, to identify groups of related tags
  - works well in certain contexts, but does not bring 'explicit semantics' into the system
    - co-occurrence has no formal meaning (still not able to address the problem of "animal living in water")
- existing semantic approaches limited in their semantic coverage
  - some use a thesaurus
  - others use a pre-defined ontology
- some cases require human intervention
- domain specific

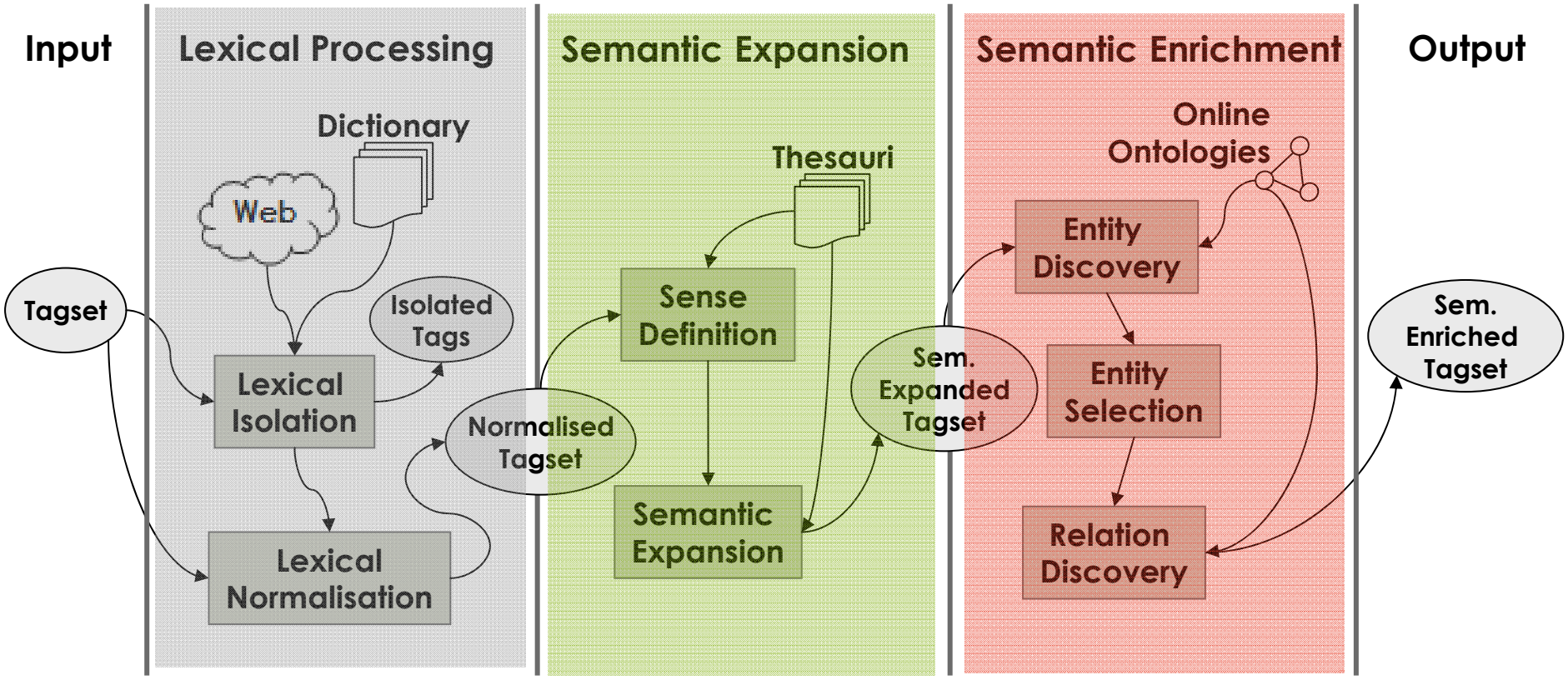
our approach

- automatic semantic enrichment of tagspaces
- exploiting the entire Semantic Web as well as other sources of background knowledge
- domain independent
- enrichment includes the semantic neighbourhood of a concept found in an ontology

Details for http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#Whale (view as graph)
Back

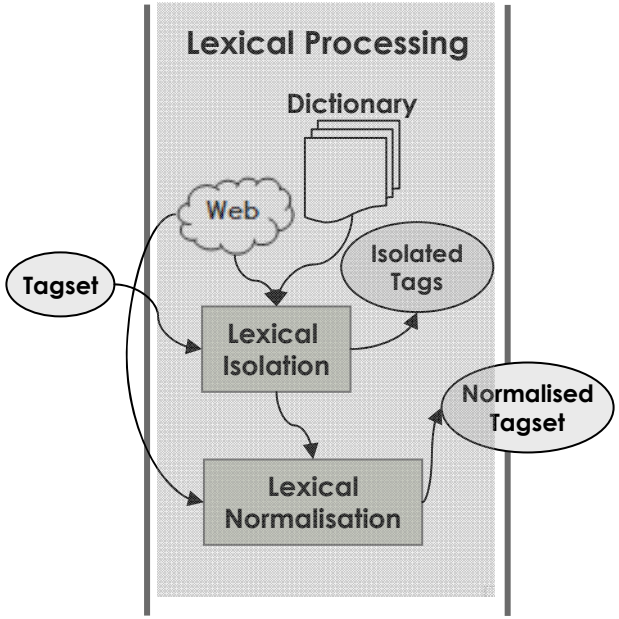
- In http://secse.atosorigin.es:10000/ontologies/cyc.owl
  - *Class*
  - **guid**: bd58dc2a-9c29-11b1-9dad-c379636f7270
  - **comment**: The collection of all whales, including most types of cetacean. #$Whale is a subset of #$Mammal, including the larges
  - **subClassOf**: http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#Cetacean
  - **type**: http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#ClarifyingCollectionType
  - **type**: http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#BiologicalTaxon
  - http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#BlueWhale: **subClassOf**
  - http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#FinbackWhale: **subClassOf**
  - http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#SeiWhale: **subClassOf**
  - http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#ToothedWhale: **subClassOf**
  - http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#RightWhale: **subClassOf**
  - http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#BowheadWhale: **subClassOf**
  - http://paoli.open.ac.uk/watson-cache/5/4b6/466d/b79a5/f1f2298c82/d30800bfaeeff211a#BaleenWhale: **subClassOf**

FLOR


Input
Lexical Processing
Dictionary
Web
Tagset
Isolated Tags
Lexical Isolation
Normalised Tagset
Lexical Normalisation
Semantic Expansion
Thesauri
Sense Definition
Semantic Expansion
Sem. Expanded Tagset
Semantic Enrichment
Online Ontologies
Entity Discovery
Entity Selection
Relation Discovery
Sem. Enriched Tagset
Output

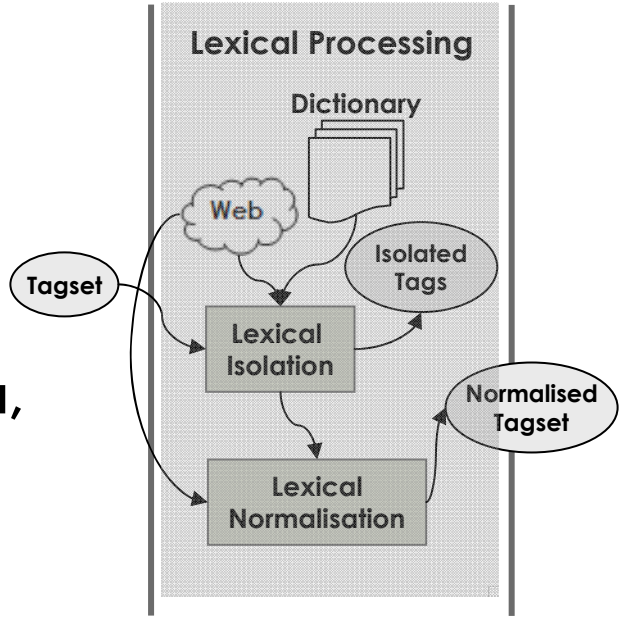
1.1.Lexical Isolation

- isolate tags that can't be processed by the next steps of FLOR
  - special characters
    ":P","(raw -> jpg)"
  - non English
    "sillon","arbol"
  - numbers
    "356days","tag1"

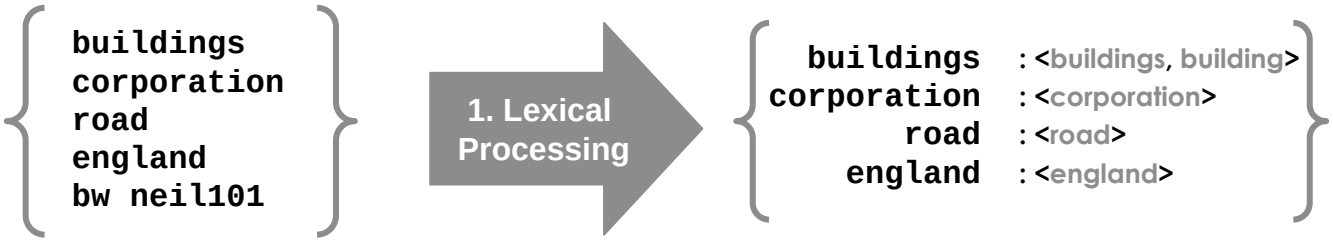


1.2.Lexical Normalisation

- **enhance anchoring**
  - **Folksonomies:** `santabarbara`
  - **Semantic Web:**
    Santa-Barbara **or** ***Santa+Barbara***
  - **WordNet:** *Santa Barbara*
  - **Produce the following:**
    - **{santaBarbara santa.barbara, santa_barbara, santa(space)barbara, santa-barbara, santa+barbara, ..}**

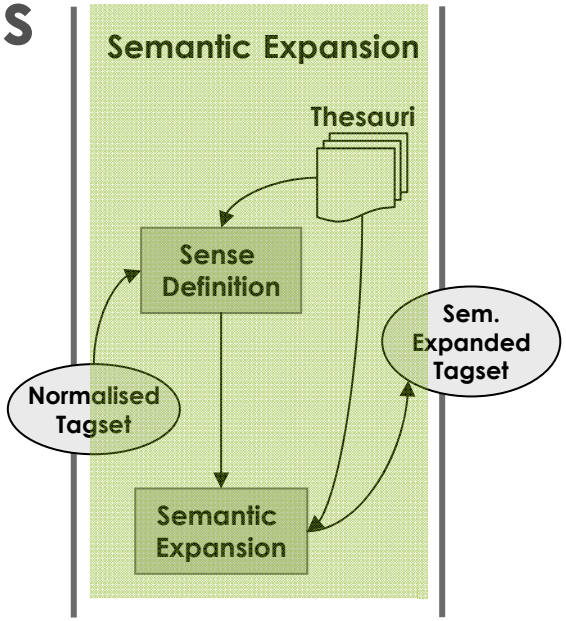


FLOR methodology



2. Sense Definition & Semantic Expansion

- Goals:
  1. Define appropriate sense for each tag (based on the context)
  2. Expand the tag with Synonyms and Hypernyms



2.1.Sense Definition

Wu & Palmer Conceptual Similarity[1]

WordNet

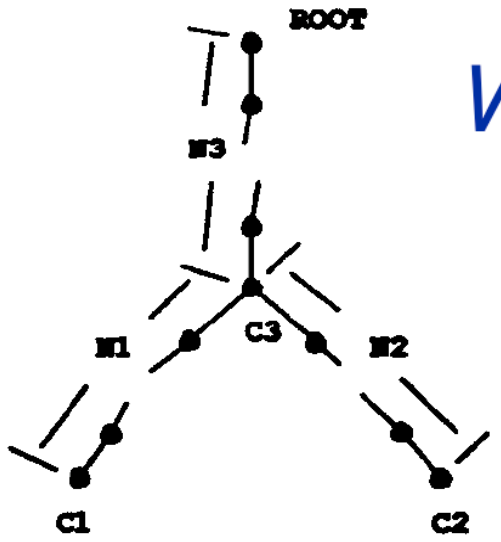


Figure 4. The concept similarity measure

The conceptual similarity between C1 and C2 is:

$$ConSim(C1, C2) = \frac{2*N3}{N1+N2+2*N3}$$

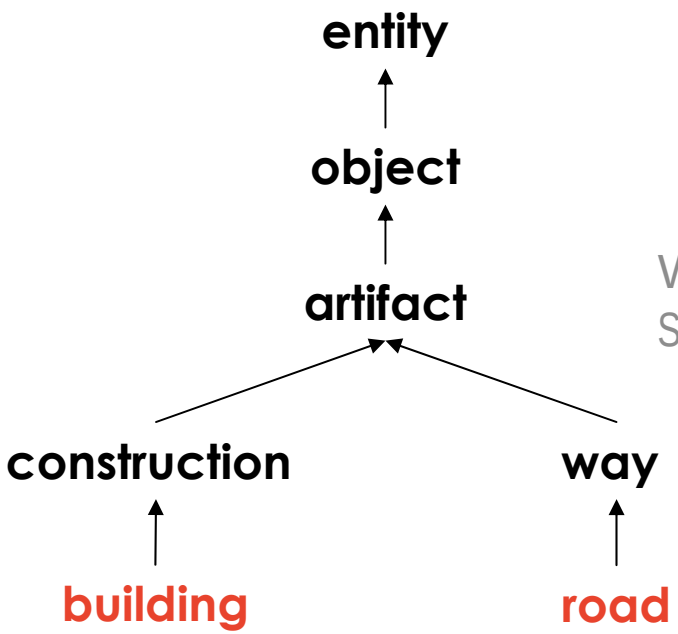
1. Z. Wu and M. Palmer. Verb semantics and lexical selection. In 32nd Annual Meeting of the Association for Computational Linguistics, 1994.

2.1.Sense Definition

building
corporation
road
england

Using the Wu and Palmer similarity formula on WordNet calculate the pairwise similarity for all combinations of tags.

WordNet



Wu and Palmer
Similarity: 0.666

2.1.Sense Definition

building
corporation
road
england

WordNet

Wu and Palmer
Similarity: 0.363

the occupants of a building; "*the entire building complained about the noise*"

2.1.Sense Definition

Selected Senses

building	a structure that has a roof and walls and stands more or less permanently in one place; "*there was a three-story building on the corner*"
corporation	a business firm whose articles of incorporation have been approved in some state
road	an open way (generally public) for travel or transportation
england	a division of the United Kingdom

2.2.Semantic Expansion

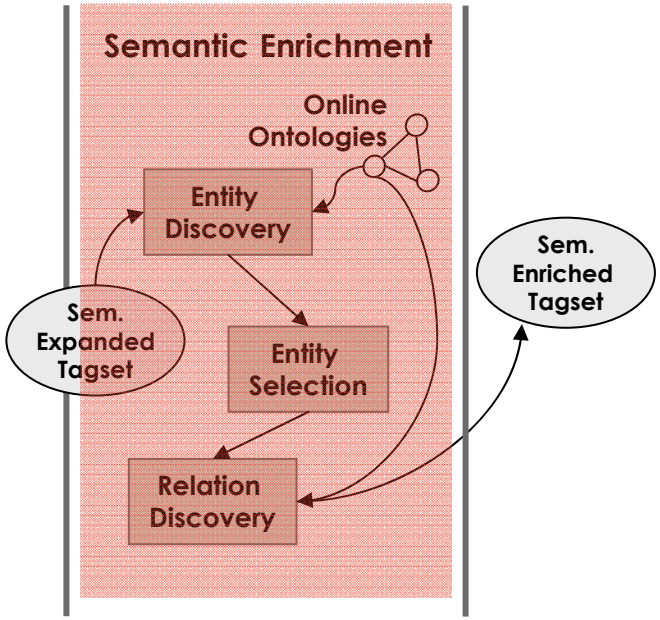
The synonyms and hypernyms from the selected senses are used to expand the tags

	Synonyms	Hypernyms	
buildings: <	<edifice>,	< structure, construction, artefact, ...>	>
corporation: <	<corp>,	< firm, business, concern,..>	>
road: <	<route>,	<way, artefact, object,..>	>
england : <	< >,	<European_Country, European_Nation, land,..>	>

FLOR methodology

3.Semantic Enrichment

- The final phase, links the tags with Ontological Entities (Semantic Web Entities, SWEs)
  - Class
  - Property
  - Individual



3.1.Entity Discovery

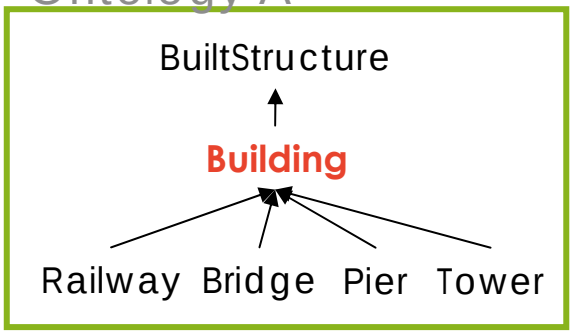
- Query the Semantic Web with 

- Identify all entities that contain
  - the tag OR
  - its lexical representations OR
  - its synonyms
- as
  - localname OR
  - label

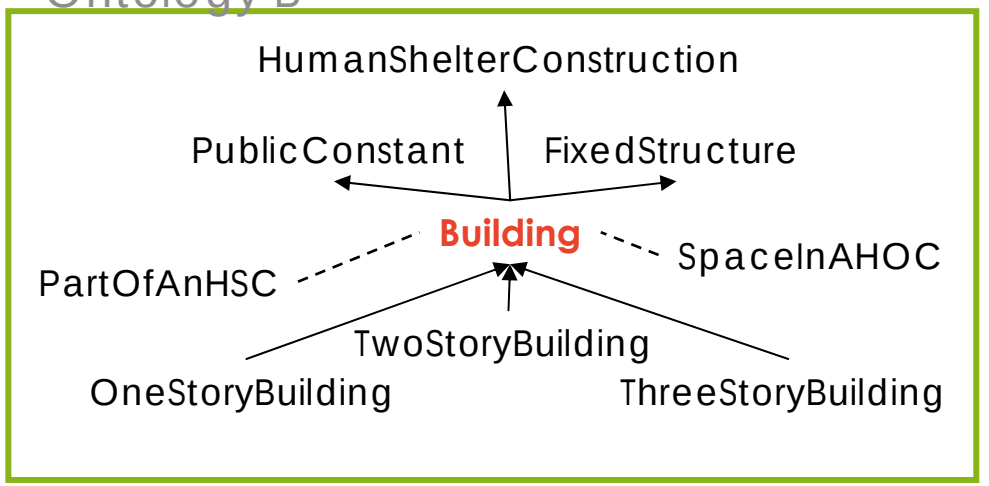
3.1.Entity Discovery

Watson results:

Ontology A

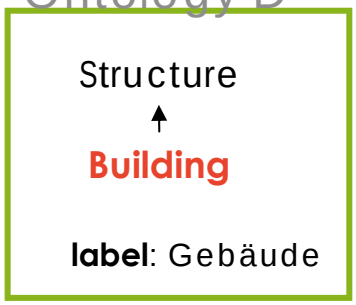


Ontology B



Ontology C

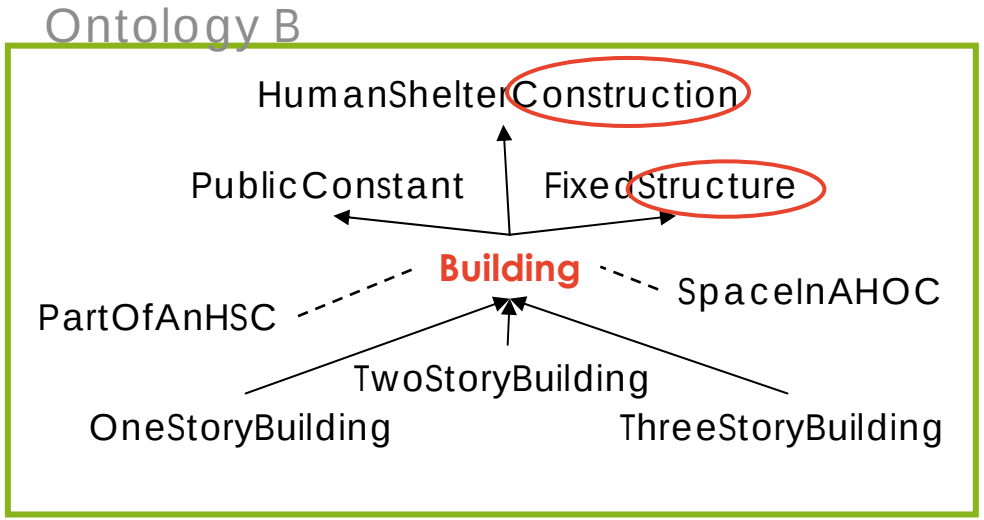
Ontology D



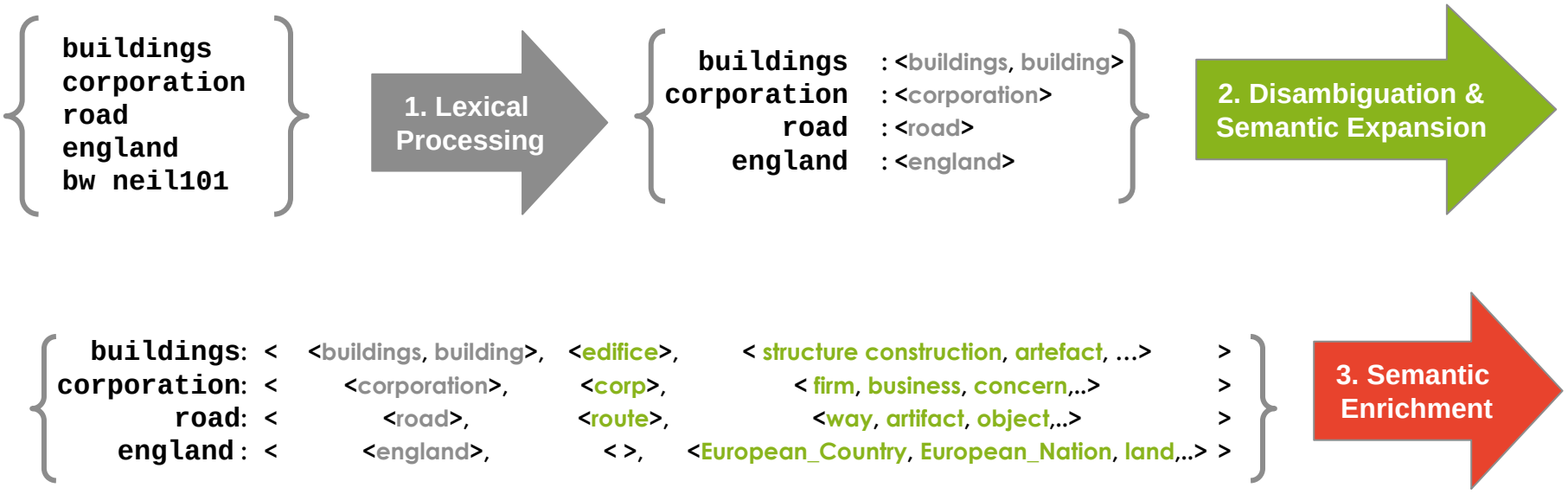
3.2.Entity Selection

the discovered Semantic Web Entities are compared against Semantically Expanded tags

buildings: < **<edifice>,** **<structure, construction, artefact, …>** **>**



FLOR methodology



preliminary experiments

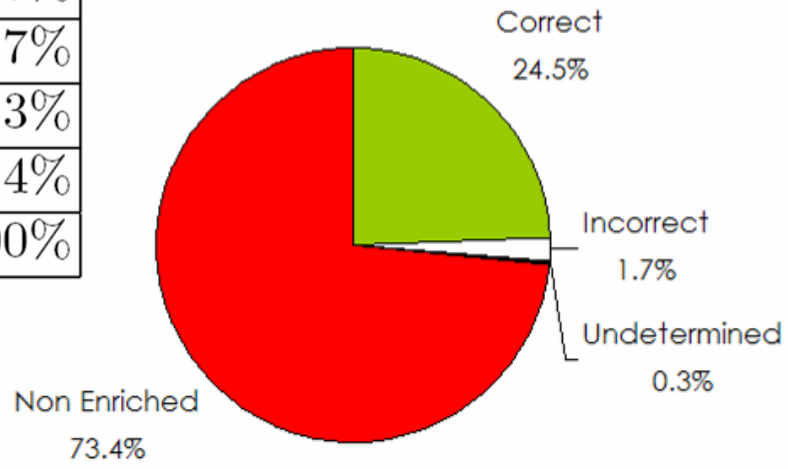
- randomly selected 250 flickr photos tagged with 2819 distinct tags
- the Lexical Isolation phase removed 59% of the tags, resulting to 1146 distinct tags and 226 photos
- the isolated tags included:
  - 45 two character tags (e.g., `pb`, `ak`)
  - 333 containing numbers (e.g., `356days`, `tag1`)
  - 86 containing special characters (e.g., `:P`, `(raw-> jpg)`)
  - 818 non English tags (e.g., `sillon`, `arbol`)

tag based results

- **Tag enrichment =** CORRECT
  - **if tag was linked to appropriate SWE**
- **Tag enrichment =** INCORRECT
  - **if tag was linked to un-appropriate SWE**
- **Tag enrichment =** UNDETERMINED
  - **If we were not able to determine the correctness of the enrichment**
- **Tag** NON ENRICHED
  - **if tag was not linked to any entity**

tag based results

Enrichment Result	# of Tags	Percentage
CORRECT	281	24.5%
INCORRECT	20	1.7%
UNDETERMINED	4	0.3%
NON ENRICHED	841	73.4%
Total	1146	100%



- 93 % enrichment precision
- 73.4% non enriched tags
  - selected a random 10% (85 tags) and were able to manually enriched 29, thus:
  - ~70% due to Knowledge Sparseness in Watson or Semantic Web
  - ~30% of the non-enriched tags due to FLOR algorithm issues

FLOR algorithm issues

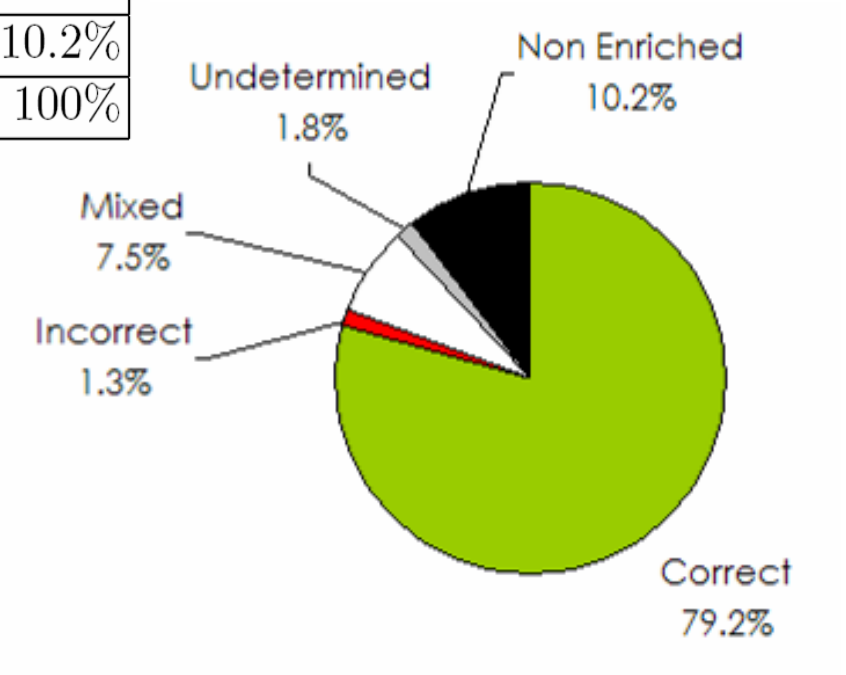
- 24% of non enriched tags defined incorrectly in Phase 2 (i.e., assigned to the wrong sense)
  - e.g., <square> assigned to <geometrical-shape> rather than <geographical-area>
- 55% of non enriched tags were differently defined in WordNet and in ontologies
  - e.g.,: `love`
    - WordNet: Love→ Emotion → Feeling → Psychological feature (*a strong positive emotion of regard and affection*)
    - Semantic Web: *Love* subClassOf *Affection*

photo based results

- **Photo enrichment =** CORRECT
  - **if all enriched tags CORRECT**
- **Photo enrichment =** INCORRECT
  - **if all enriched tags INCORRECT**
- **Photo enrichment =** MIXED
  - **if some tags INCORRECT and some tags CORRECT**
- **Photo enrichment =** UNDETERMINED
  - **if all enriched tags UNDETERMINED (i.e. could not decide on correctness)**
- **Photo** NON ENRICHED
  - **if none of the tags was enriched**

photo based results

Enrichment Result	# of Photos	Percentage
CORRECT	179	79.2%
INCORRECT	3	1.3%
MIXED	17	7.5%
UNDETERMINED	4	1.8%
NON ENRICHED	23	10.2%
Total	226	100%



future work

- Semantic Relatedness measure instead of similarity measure
- Process the Lexically Isolated tags using other background knowledge resources, e.g. Wikipedia.
- Relation discovery between tags with scarlet
- Step2: Intelligent Query Interface
- large scale evaluation

conclusions

- automatic semantic enrichment of tagspaces is possible
  - 93% precision in the 24.5% enriched tags
  - 79% enriched resources
- three phase architecture works well
  - identified the steps of each phase that require improvement

Thank you ☺

S.Angeletou@open.ac.uk

http://flor.kmi.open.ac.uk/