



UNIVERSITY OF
Southampton



Preliminary Results in Tag Disambiguation using DBpedia

Andrés García-Silva[†], Martin Szomszor[‡], Harith Alani[‡], Oscar Corcho[†]

[†] {hgarcia, ocorcho}@fi.upm.es

Facultad de Informática

Universidad Politécnica de Madrid

Campus de Montegancedo s/n

28660 Boadilla del Monte, Madrid, Spain

[‡] {mns2, h.alani}@ecs.soton.ac.uk

School of Electronics and

Computer Science,

University of Southampton,

SO16 1BJ, UK.

1. Introduction
 - 1.1 Folksonomies & problems
 - 1.2 Association of semantics to tags
2. State of the Art
3. Context & Disambiguation
 - 2.1 Context in dictionary based approaches
 - 2.2 Context in folksonomies
4. Tagora sense repository
5. Disambiguation approach
6. Preliminary results
7. Conclusions and future work

1. Introduction

1.1 Folksonomies & problems

1.2 Association of semantics to tags

2. State of the Art

3. Context & Disambiguation

2.1 Context in dictionary based approaches

2.2 Context in folksonomies

4. Tagora sense repository

5. Disambiguation approach

6. Preliminary results

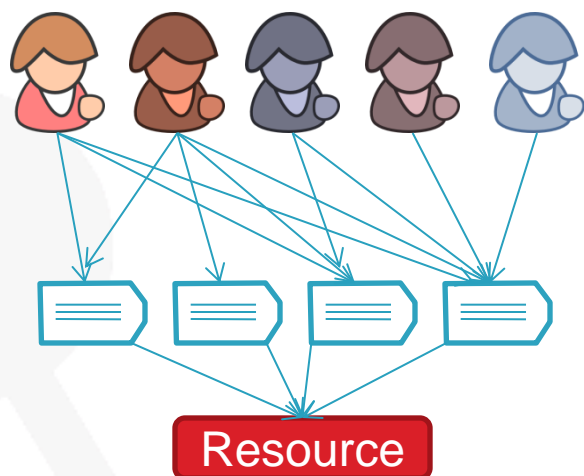
7. Conclusions and future work

- Folksonomy*
 - The result of free tagging information and objects for one's own retrieval (Anything with an URL)
 - Tagging is done in a social environment
 - People use their own vocabulary

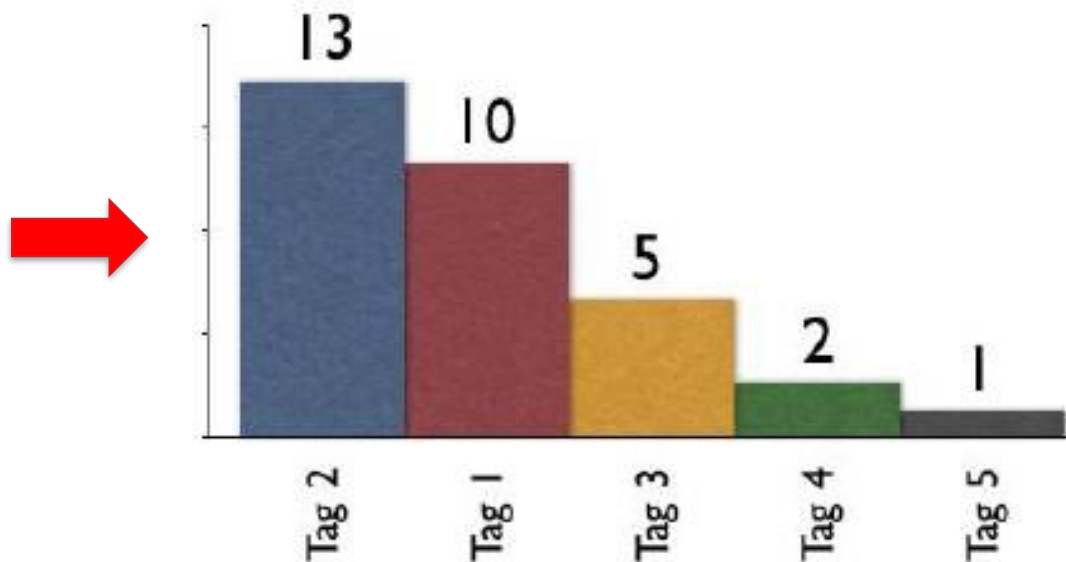


* <http://www.vanderwal.net/folksonomy.html>

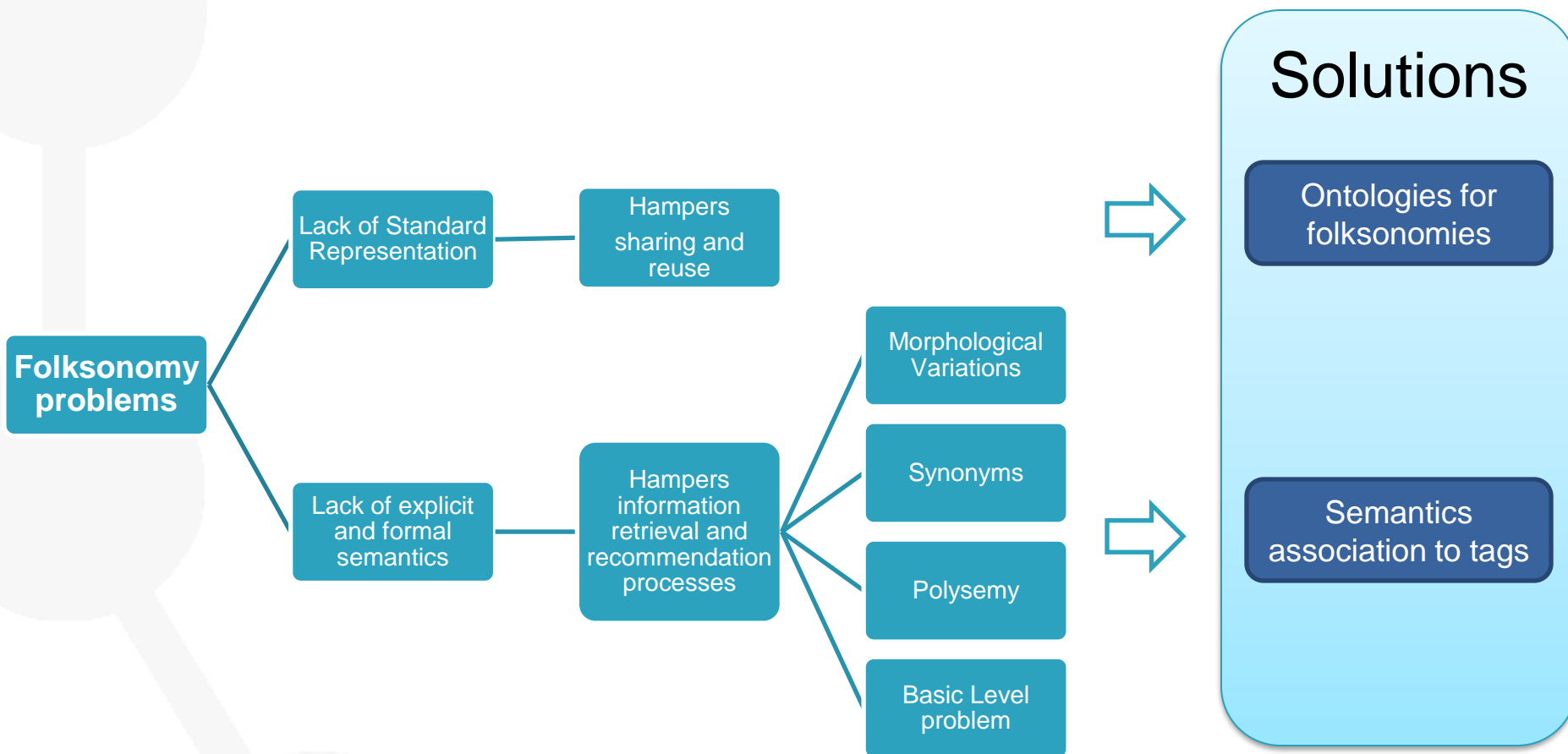
Broad folksonomy*

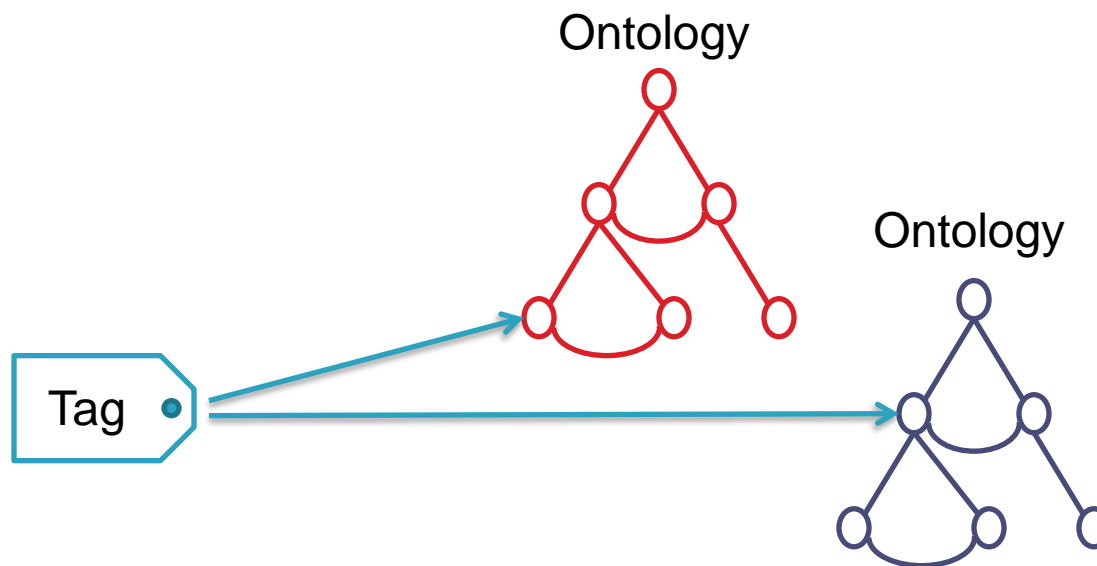


Vocabulary emergence



* www.vanderwal.net





Semantics association to tags



1. Introduction

1.1 Folksonomies & problems

1.2 Association of semantics to tags

2. State of the Art

3. Context & Disambiguation

2.1 Context in dictionary based approaches

2.2 Context in folksonomies

4. Tagora sense repository

5. Disambiguation approach

6. Preliminary results

7. Conclusions and future work

Technique

Semantic

Automatic

Our contribution:

- Automatic approach to associate semantics to tags relying on **DBpedia** as semantic resource.
- Disambiguation algorithm inspired by well known information retrieval techniques
- Several context definition as a way to ameliorate tagging data scarceness in the disambiguation process.

- No explicit semantics

- Specia, L., Motta, E., Integrating Folksonomies with the Semantic Web. In *Proceedings of the 4th European Conference on the Semantic Web: Research and Applications*, Innsbruck, Austria (2007)
- Hamasaki, M., Matsuo, Y., Nisimura, T., Takeda, H., Ontology Extraction using Social Network. In *International Workshop on Semantic Web for Collaborative Knowledge Acquisition*, Hyderabad, India (2007)
- Angeletou, S., Sabou, M., Motta, E., Semantically Enriching Folksonomies with FLOR. In *1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008)*, Tenerife, Spain (2008).

1. Introduction
 - 1.1 Folksonomies & problems
 - 1.2 Association of semantics to tags
2. State of the Art
- 3. Context & Disambiguation**
 - 2.1 Context in dictionary based approaches
 - 2.2 Context in folksonomies
4. Tagora sense repository
5. Disambiguation approach
6. Preliminary results
7. Conclusions and future work

- First disambiguation approach relying on a dictionary (Lesk, 1998)
 - Definitions of the word to disambiguate & of each word in the context.
 - Context: The words appearing in the sentence
 - Definitions of the words in the context are compared against the definitions of the word to disambiguate.

Problems: When the definitions are short. (Sanderson, 2000)

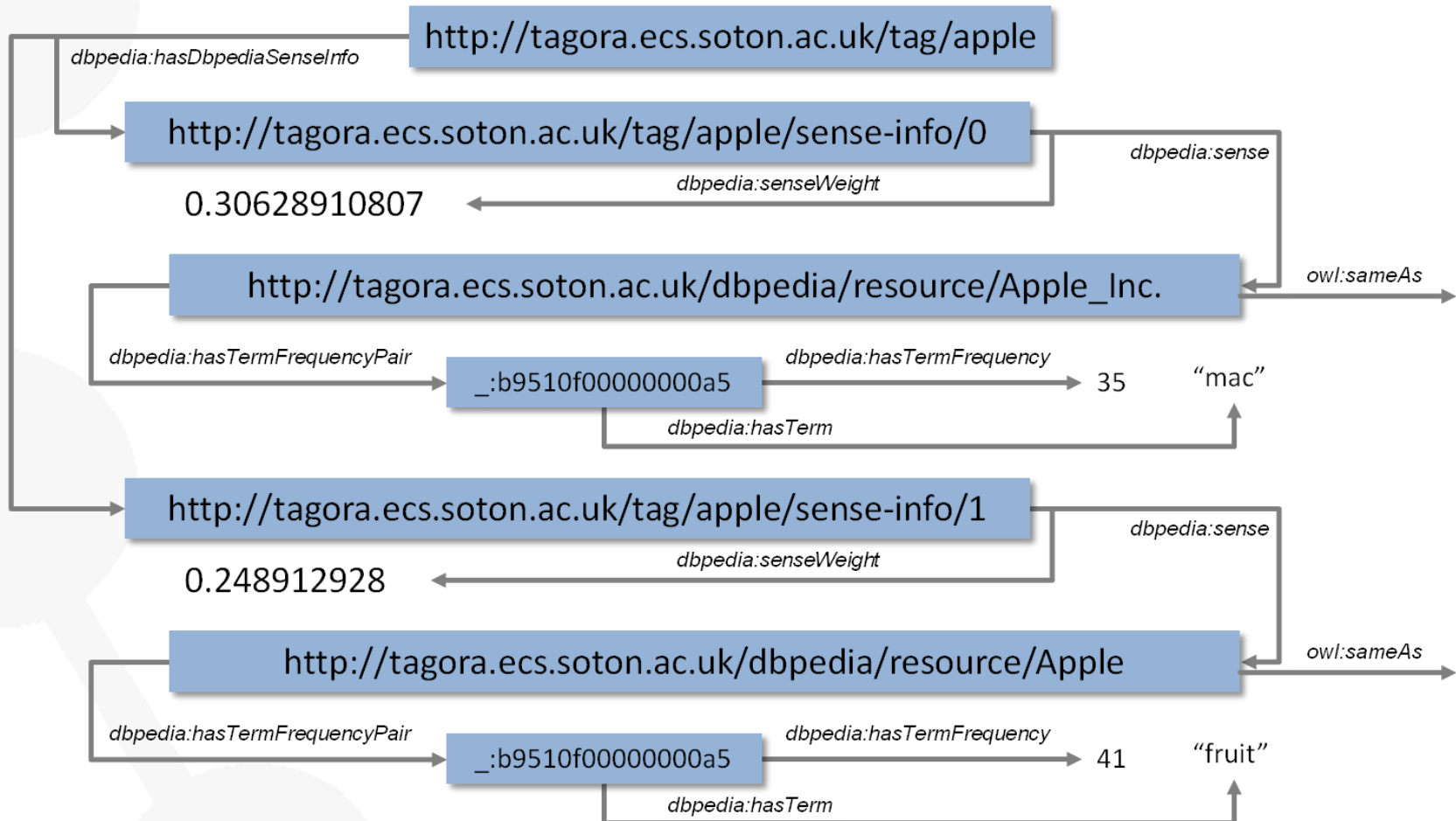
- Lesk, M., "They said true things, but called them by wrong names" – vocabulary problems in retrieval systems. in *Proc. 4th Annual Conference of the University of Waterloo Centre for the New OED* (1998)
- Sanderson, M., Retrieving with Good Sense. In *Information Retrieval* 2(1): 47-67 (2000)

- Contexts in folksonomies



- Linked data enabled service endpoint
 - Metadata about tags and their possible senses.
 - Wikipedia pages -> Disambiguation or Redirection links
 - Terms and frequencies
 - DBpedia resource related to each Wikipedia page.
 - Query using:
 - REST -> <http://tagora.ecs.soton.ac.uk/tag/apple>
 - SPARQL end-point.
 - Result: RDF document
- DBpedia coverage:
 - **2.6** million things, 213,000 people, 328,000 places, 57,000 music albums, 36,000 films, 20,000 companies.
 - Wordnet as of 2006 contains about 150,000 words
 - Named entity recognition
 - Classes, Instances, and semantic relations

Linked data representation of tag senses



1. Introduction
 - 1.1 Folksonomies & problems
 - 1.2 Association of semantics to tags
2. State of the Art
3. Context & Disambiguation
 - 2.1 Context in dictionary based approaches
 - 2.2 Context in folksonomies
4. Tagora sense repository
- 5. Disambiguation approach**
- 6. Preliminary results**
7. Conclusions and future work

- The algorithm selects among a set of candidate DBpedia resources, the one that describe better the meaning of an ambiguous tag according to its context.
- The candidate DBpedia resources and the tag context are represented as vectors using a common vocabulary.
 - The common vocabulary is the union of the most frequent terms in each wikipedia page related to each Dbpedia resource.
- The tag context vector is compared against each DBpedia resource vector using the cosine of the angle as similarity measure.

$$Sim(V_{context}, V_{sense}) = \cos \theta = \frac{V_{context} \cdot V_{sense}}{|V_{context}| |V_{sense}|}$$

- The most similar DBpedia entry is selected as the one representing the meaning of the analyzed tag

- Tagging activity:
 - User u has tagged the resource $r = \text{http://www.nature.com}$ with the tags $nature$, $news$, $science$.
- **Context**(u , $nature$, r) = { $nature$, $news$, $science$ }
- **Senses**($Nature$) = { $dbpedia:Nature$, $dbpedia:Nature_journal$ }
- $\text{Terms}(dbpedia:Nature) = \{(life,62), (nature,46), (earth,32)\}$
- $\text{Terms}(dbpedia:Nature_journal) = \{(nature,77), (science,29), (scientific,25)\}$
- **Voc**($nature$) = { $life$, $nature$, $earth$, $science$, $scientific$ }
- $V_{context} = (0,1,0,1,0)$
- $V_{nature} = (62,46,32,0,0)$
- $V_{nature(journal)} = (0,77,0,29,25)$
- $\text{Sim}(V_{context}, V_{nature}) = 0,389$
- $\text{Sim}(V_{context}, V_{nature(journal)}) = \mathbf{0,872}$



Some user x has tagged a picture r with the tags *ice*, *iceskating*, *nottingham*, and *skating*.



| | |
|---|--------------|
| <i>ice</i> | |
| dbpedia/resource/Ice | 0,911 |
| dbpedia/resource/Ice_(comics) | 0,735 |
| <i>skating</i> | |
| dbpedia/resource/Artistic_roller_skating | 0,671 |
| dbpedia/resource/Figure_skating | 0,569 |
| dbpedia/resource/Freestyle_slalom_skating | 0,000 |
| dbpedia/resource/Ice_skating | 0,893 |
| dbpedia/resource/Road_skating | 0,451 |
| dbpedia/resource/Roller_skating | 0,394 |
| dbpedia/resource/Skateboarding | 0,197 |
| dbpedia/resource/Snowboarding | 0,000 |
| dbpedia/resource/Speed_skating | 0,549 |
| dbpedia/resource/Tour_skating | 0,831 |
| <i>nottingham</i> | |
| dbpedia/resource/East_Nottingham_Township,_Pennsylvania | 0,000 |
| dbpedia/resource/Elizabeth_I_of_England | 0,000 |
| dbpedia/resource/Nottingham | 0,750 |
| dbpedia/resource/Nottingham,_New_Hampshire | 0,386 |
| dbpedia/resource/Nottingham_Cooperative | 0,524 |
| dbpedia/resource/Nottingham_Township,_Harrison_County,_Ohio | 0,000 |
| dbpedia/resource/Nottingham_Township,_Pennsylvania | 0,000 |
| dbpedia/resource/Nottinghamshire | 0,428 |
| dbpedia/resource/Sheriff_of_Nottingham | 0,640 |
| dbpedia/resource/West_Nottingham_Township,_Pennsylvania | 0,000 |

Issues affecting the disambiguation process

| DBpedia Entry | Freq (<i>nottingham</i>) | Terms in $Voc(nottingham)$ |
|------------------------------|-------------------------------|-------------------------------|
| ../Nottingham | 181 | 16 |
| ../Sheriff_of_Nottingham | 9 | 3 |
| ../Nottingham_Cooperative | 12 | 12 |
| ../Nottinghamshire | 27 | 15 |
| ../Nottingham,_New_Hampshire | 14 | 17 |

- Dbpedia resources with lower number of terms in $Voc(tag)$
 - *When the context has few terms, it's likely that shorter sense vectors are more similar to the context vector than longer sense vectors.*
- When none of the tags in the context appears in $Voc(tag)$ the sense selection is carried out in terms of $freq(tag)$

1. Introduction
 - 1.1 Folksonomies & problems
 - 1.2 Association of semantics to tags
2. State of the Art
3. Context & Disambiguation
 - 2.1 Context in dictionary based approaches
 - 2.2 Context in folksonomies
4. Tagora sense repository
5. Disambiguation approach
6. Preliminary results
7. Conclusions and future work

- Conclusions:
 - Inspired by IR techniques we have presented a tag disambiguation algorithm relying on DBpedia & Wikipedia information.
 - Vector representation of the tag contexts and Dbpedia resources using a common vocabulary based on term frequency.
 - Preliminary results and some identified problems:
 - Few terms in the context
 - Terms in the context do not appear in the common vocabulary
 - Few terms in the Dbpedia resource
 - We have presented different definitions of contexts for tagging activities as a way to ameliorate tagging data scarceness.

- Future Work
 - Improve current results
 - Test the approach in a large tag set with the different contexts.
 - Sophisticated similarity measures
 - Study tagging activities in specific domains.
 - Extract more context information from the tagged resource (text documents)
 - Evaluation of the approach
 - Evaluation using *Precision* and *Recall*
 - Test bed and Standard evaluation metrics
 - Applications
 - Use DBpedia semantic information to evolve domain ontologies
 - Use DBpedia semantic information to improve searching and recommendation processes.