

Session 1: NLP and the Multilingual Semantic Web: Challenges and Opportunities

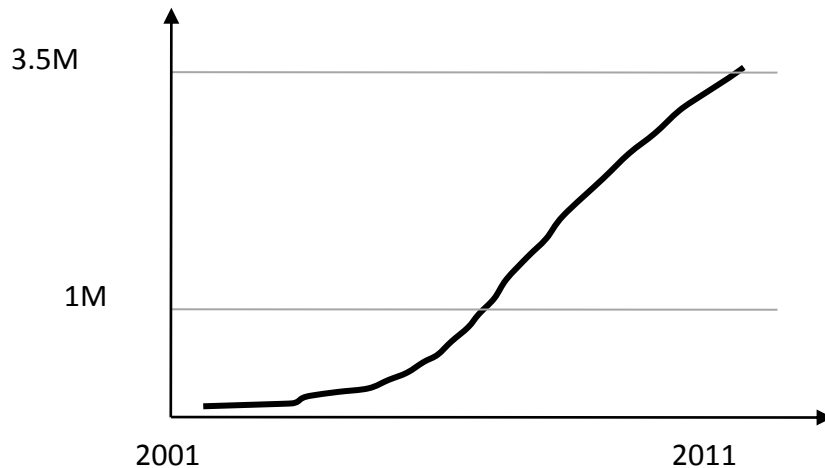
Tobias Wunner

**Digital Research Enterprise Institute (DERI)
National University of Ireland, Galway (NUIG)**



Monnet is supported
by the European Union
under Grant No. 248458

- Wikipedia
 - 250 languages



articles	languages
3.6M	English (1Bln words)
~ 1M	German, French
< 800k	Spanish, Polish, Italian
< 700k	Russian, Dutch, Chinese
< 200k	Slovenian
< 20k	Afrikaans, Irish

From: http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

- Similar resources
 - Hudong (Chinese)
 - 3.9m articles (Wikipedia English 3.6m)

The image shows two side-by-side screenshots of encyclopedia entries for 'Traditional Chinese Medicine'. The left screenshot is from Hudong (互动百科), a Chinese encyclopedia. It features a blue header with the Hudong logo and navigation tabs for 'Terms', 'Articles', and 'Pictures'. The entry title is '中医' (Traditional Chinese Medicine) with a star rating. Below the title, it lists categories: 'Open Category: 中医 中医学 中国传统文'. The entry text mentions it was created by 'vampire女王' and lists collaborators: '湖', '贾荣曼', and '冯济相'. There is a small photo of a person in a white lab coat working with herbs, captioned '中药房在抓药'.

The right screenshot is from Wikipedia, showing the 'Wikipedia entry' for 'Traditional Chinese Medicine'. It has a blue header with the Hudong logo and navigation tabs for 'Term', 'Article', and 'Pictures'. The entry title is 'Traditional Chinese Medicine' with a star rating. Below the title, it lists categories: 'Open Category:'. The entry text mentions it was created by 'the vampire queen' and lists collaborators: 'Wikipedia editors', 'medicine bone disease network', 'lagoon', 'Jia Rong Man', and 'Feng Ji-phase'. There is a small photo of a person in a white lab coat working with herbs, captioned 'Chinese medicine, is "The Origin and Formation in China, with the overall concept of the medical diagnosis and treatment and so on."'. A large red 'Google Translate' watermark is overlaid on the bottom right of the Wikipedia entry.

- Term variations

widely accepted term

"traditional chinese medicine"

About 7,000,000 results (0.13 seconds)

rarely used term variation

"chinese traditional medicine"

About 330,000 results (0.30 seconds)

- Linguistic variations
 - Singular plural variation (English)

Participle + noun singular

"received payment"

About 3,190,000 results (0.12 seconds)

noun plural + noun singular

"payments receivable"

About 73,400 results (0.20 seconds)

- Case variation (Irish, translation=traditional Chinese medicine)

genitive case

"leighis Sínis traidisiúnta"

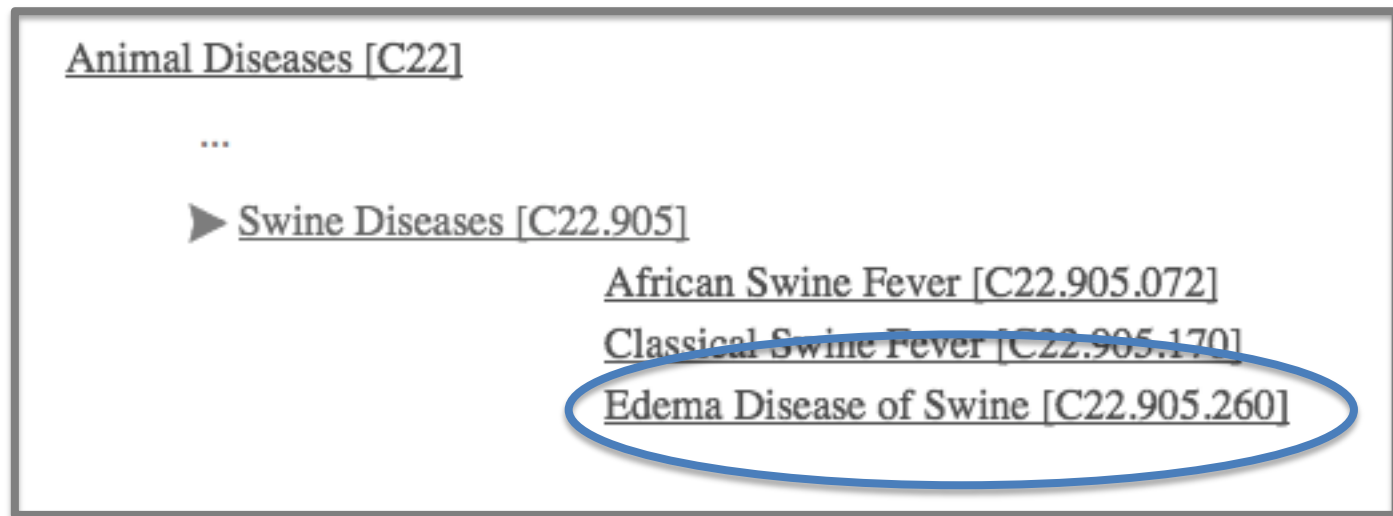
About 906 results (0.17 seconds)

nominative case

"leigheas Sínis traidisiúnta"

About 3,770 results (0.16 seconds)

- There is no correct / incorrect on the Web
- But standards (formal agreements)
 - MeSH (Medical Subject Headings) classification
 - Example “edema disease of swine” (vs. “edema swine disease”)



From <http://www.nlm.nih.gov/mesh/MBrowser.html>

- ...is multilingual

Multilingual literals (STW - German economy Thesaurus)

<descriptor/13555-4>

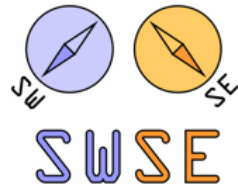
```
skos:inScheme <../stw> ;
skos:prefLabel "Autobahn"@de, "Motorway"@en ;
skos:altLabel "Autobahnnetz"@de, "Autoroute"@de,
"Freeway"@en, "Highway"@en ;
```

Multilingual vocabularies (Rechtspraak.nl -Dutch)

:aard_van_de_schade

```
rdf:type best:Concept , skos:Concept ;
skos:altLabel "zaakschade"@nl , "aard van de schade"@nl, ... ;
skos:broader :invulling_beoordelingscriterium ;
```

- How to search?
 - Semantic Web Query Language (SPARQL)
 - Semantic Web Search Engines



- How to search literals?
- DBPedia SPARQL endpoint
- Query: "All resources with German String
Traditionelle chinesische Medizin"

Query

Default Graph URI

http://dbpedia.org

(Security restrictions of this server do not allow administrator can change them, according to t

Query text

```
select * where {  
  ?subject| rdfs:label "Traditionelle chinesische Medizin"@de  
}
```

Results

subject

http://dbpedia.org/resource/Traditional_Chinese_medicine

- How to search literals?
 - Sindice Query (supported)

“Search all literals with String”

Search interface type: [Simple](#) [Advanced](#) [Guru](#) [Query Language Documentation](#)

keyword(s)

Group By Dataset: ☐ Sorted by:

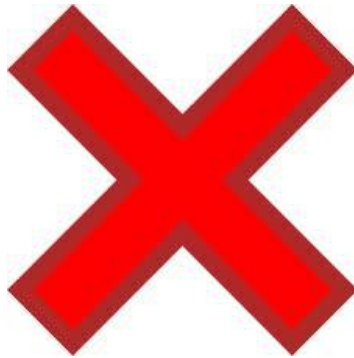
Sindice search: 中医 found 36 documents (in 0.01 seconds)

Results

```
<http://raynix.cn/...> dc:title "
...
```

- How to search words in resource names?
- SPARQL Query (only possible with regexes... very slow)

”All resources with word traditional”



- How to search words in resource names?
- Sindice Query:

“All resources with word chinese & medicine” (unordered)

Search interface type: [Simple](#) [Advanced](#) **[Guru](#)** [Query Language Documentation](#)

keyword(s)

ntriple(s)

<traditional AND chinese> * *

Results

http://dbpedia.org/resource/traditional_Chinese_medicine (RDF)

+ 2009-08-28 - 301 triples in 43.5 kB

http://dbpedia.org/resource/traditional_Chinese_medicine ([Search](#)) Inspect: ([Cache](#)) ([Live](#))

http://dbpedia.org/resource/traditional_Chinese (RDF)

+ 2009-08-28 - 127 triples in 17.5 kB

http://dbpedia.org/resource/traditional_Chinese ([Search](#)) Inspect: ([Cache](#)) ([Live](#))

- Multilingual/Ontology-based Information Extraction (BioCaster, OpenCalais)
- Multilingual Access to Linked Data (DBPedia)
- Ontology Localization (LabelTranslator)
- Ontology-based Natural Language Generation (CLANN)

- Aggregates and processes health news
- Annotates news based on a multilingual ontology
- Uses proprietary format and SKOS-XL to maintain terminology

<http://born.nii.ac.jp>

concept = measles

Synonyms	en: Rubeola (DISEASE) en: Hard measles (DISEASE) en: Morbilli (DISEASE) en: 14-day measles (DISEASE) en: Rubella (DISEASE) en: Measles (DISEASE) fr: Rougeole (DISEASE) fr: Rougeole (DISEASE) jp: 麻疹 (DISEASE) jp: 風疹 (DISEASE) ...
	ko: 마진 (DISEASE) zh: 麻疹 (DISEASE) es: Sarampión (DISEASE) th: โรคหัด (DISEASE) vi: Bệnh sởi (DISEASE)

- Global Health Monitor [en]**

In combination with Malta [Lat: 35.946153, Long: 14.384604]

Found on Google News (2011-04-25)

»Search for biomedical references on NCBI, HighWire, GoPubMed, Google Scholar

<http://born.nii.ac.jp>

- Challenges
 - Multilingual adaptation
 - Adaptation of information extracion rules to other domains
 - Use of proprietary format is undesirable

- Semantic markup of unstructured text
- Multilingual (English, French, Spanish)

- English

- 39 entities
 - 75 relations

Entities

Anniversary, City, Company, Container, FaxNumber, Holiday, IndustryTerm, MusicGroup, NaturalFeature, OperationalProduct, ProgrammingLanguage, SportsEvent, SportsGame, SportsLeague

Events and Facts

Acquisition, Alliance, AnalystEarnings, BonusSharesIssuance, BusinessRelationship, CompanyCustomer, CompanyEarnings

- Domain tuned (Finance, Biomedical)
- Only 15 base entities for non-English, no relations
- Demo

<http://viewer.opencalais.com>

John McCrae, Elena Montiel Ponsado and Tobias Wunner

are presenting natural language processing models

for the semantic web at ESWC 2011 in Greece.

Entities:

☒ Country

☒ Greece

☒ Industry Term

☒ natural language processing

☒ semantic web

☒ Person

☒ Elena Montiel Ponsado

☒ John McCrae

☒ Tobias Wunner

☒ Technology

☒ natural language processing

Events & Facts:

☒ Generic Relations

☒ Tobias Wunner, natural language processing, present

☒ natural language processing, model

- Challenges
 - Multilingual adaptation of lexicon and extraction rules
 - Domain adaptation of lexicon and extraction rules

- Multilingual ontology editor
- Linguistic annotations (Num., POS, Gender)
- ... a for better translation

Identifier

URI:

LexicalEntry-6	noun	French	×		
LexicalEntry-1	noun	Spanish	×		
LexicalEntry-5	noun	French	×		
LexicalEntry-7	noun				

Number + Gender

▼ Lexicalizations

Entries

Label	G. Number	Gender	Dialect	Language	
Rivière	Singular	Feminini		French	×

part of speech

▼ Lexical Entry Senses

Entries

Identifier	Language			
Sense-1	French	×		

Definitions

Definition	Language
Cours d'eau moyennement abondant qui se jette dans un fleuve, dans la mer ...	French

***Ontology Navigator**

- Fenómeno_hidrográfico
 - Aguas
 - Aguas_Continetales
 - Aguas_Subterráneas
 - Aguas_Superficiales
 - Aguas_Corrientes
 - Aguas_Corrientes_Artificiales
 - Aguas_Corrientes_Naturales
 - Arroyo
 - Río
 - Torrente
 - Corriente
 - Aguas_Quietas
 - Masa_de_Agua_Artificial
 - Aguas_Dulces
 - Aguas_Saladas
 - Glaciar

Entity Properties

Identifier

URI: <http://www.owl-ontologies.com/Ontology1>

Lexical Entries

Entries

Identifier	Part Of Speech	Language
LexicalEntry-3	noun	Spanish
LexicalEntry-2	noun	Spanish
LexicalEntry-4	noun	English
LexicalEntry-6	noun	French
LexicalEntry-1	noun	Spanish
LexicalEntry-5	noun	French
LexicalEntry-7	noun	Catalan

Lexicalizations

Lexical Entry Senses

Usage Context

Sources

“river”@en

“rivière”@fr

“fleuve”@fr

Ambiguous!

- Challenges
 - Ontology context-aware translations
 - Use linguistic features in the lexicon for better translation

- Controlled Language ANnotations (CLANN)
- To write domain specific grammars (meeting minutes)
- Intermediate representation

Domain ontology (e.g. meeting minutes)



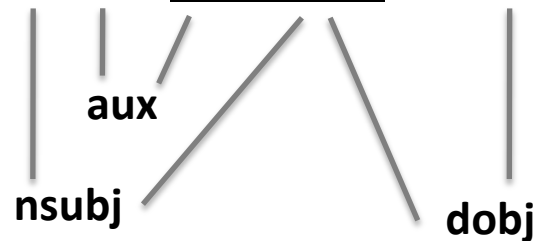
MLink Grammer



LinkedGrammar

- Example

“John will present lemon model.”



```
:Sentence1 :hasRootNode [  
  rdf:type :TextNode ;  
  :hasText "present" ;  
  :hasSubType :Verb ;  
  :hasObject [ rdf:type :TextNode ;  
                :hasText "model" ;  
                :hasObjectModifier [ rdf:type :TextNode ;  
                                      :hasText "lemon" .  
                ] ] ] ]
```

- Challenges
 - From text to triples?
 - Multilingual adaptation
 - Domain adaptation (meeting minutes)

- Web and Semantic Web is
 - “Lingual” (variations within one language)
 - Multilingual (between languages and cultures)
- Applications need domain and multilingual adaptation
 - Lexicon updates / extensions
 - Extraction rules updates / extensions
- What do we need?
 - Efficient adaptation and sharing of linguistic resources between ontology-based NLP applications