



# Cross-language Semantic Retrieval and Linking of eGov Services

Fedelucio Narducci\*, **Matteo Palmonari\***, Gianni Semeraro<sup>o</sup>

\*DISCO, University of Milan-Bicocca, Italy

\*Department of Computer Science, University of Bari Aldo Moro, Italy



# ISWC 2013

Sydney, Australia



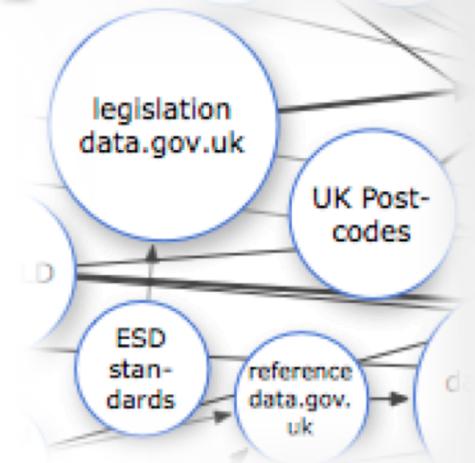
# Cross-language Linking of Open Government Data

- A large amount of Open Government Data in many languages\*:
  - 1,000,000+ datasets published online (February 2013)
  - 40 different countries
  - **24 different languages**



# Cross-language Linking of eGov Service Descriptions

- Government service catalogs are part of the LOD cloud
  - Electronic Service Delivery (ESD)-toolkit
  - European Local Government Service List (**LGSL**)
    - 2000+ interlinked public services in 6 languages
    - Each country maintains its list of public services



# Cross-language Linking of eGov Services

## *Why it is Useful*

- Advantages for PAs:
  - Compare local service offerings with best practices in other countries
  - Support interoperability among PAs of different countries and other service providers
  - Enrich service descriptions with additional information via links to LGSL (e.g., link to life event ontologies)
- Advantages for citizens
  - Find eGov services when in a foreign country
  - Towards cross-language service access

**Costly and Error Prone Activity**

Catalogs of several hundreds of services

# Cross-language Linking of eGov Services

## *Why it is Challenging*



≈ sameAs links

Semantic heterogeneity

- not a mere “translation” problem
- cultural bias

Ultra-short descriptions

- Challenging **cross-language matching** problem
- Most of the approaches:
  - use **structural information** [Spohr et al. 2011, Fu et al. 2011, Wang et al. 2009] or **long textual descriptions** [Knoth et al. 2011]
  - or report problems when automatic **translation** return descriptions **with heterogeneous vocabulary** [Hertling & Paulheim 2012]

# CroSeR

## Cross-language Service Retriever

[Home](#) [Info](#)

[Upload list](#) [Register](#) [Login](#)

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

30



Service Catalog

Service Matched from ESD

Search

Service Info

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR

## Cross-language Service Retriever

[Home](#) [Info](#)

[Upload list](#) [Register](#) [Login](#)

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

30



Service Catalog

Service Matched from ESD

Search

Service Info

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR

## Cross-language Service Retriever

[Home](#) [Info](#)

[Upload list](#) [Register](#) [Login](#)

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

• [Load a catalog](#)

30



Service Catalog

Service Matched from ESD

Search

Service Info

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR

## Cross-language Service Retriever

[Home](#) [Info](#)

[Upload list](#) [Register](#) [Login](#)

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

• [Load a catalog](#)

30 ▼

#### Service Catalog

• [Select a source service](#)

51	Straßen- und Brückenunterhaltung, Straßenverkehrstechnik
63	Hausumringe Bremerhaven
65	Fäll- und Schnittgenehmigung nach der Bremischen Baumschutzverordnung
79	Spritzen - gebrauchte - Entsorgung
133	Briefwahl

#### Service Matched from ESD

#### Service Info

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR

## Cross-language Service Retriever

Home Info

Upload list Register Login

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

• Load a catalog

30 ▼

#### Service Catalog

• Select a source service

Br
51 Straßen- und Brückenunterhaltung, Straßenverkehrstechnik
63 Hausumringe Bremerhaven
65 Fäll- und Schnittgenehmigung nach der Bremischen Baumschutzverordnung
79 Spritzen - gebrauchte - Entsorgung
133 Briefwahl

#### Service Matched from ESD

Search	
1 Voting	↑ ↔ ↓
2 Proxy voting	↑ ↔ ↓
3 Postal voting	↑ ↔ ↓
4 Members - elections - polling stations	↑ ↔ ↓
5 Election results	↑ ↔ ↓
6 Members - elections - results	↑ ↔ ↓
7 Electoral register	↑ ↔ ↓

#### Service Info

• Look at the retrieved services (link recommendations)

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR

## Cross-language Service Retriever

[Home](#) [Info](#)

[Upload list](#) [Register](#) [Login](#)

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

• [Load a catalog](#)

30 ▼

#### Service Catalog

• [Select a source service](#)

Br
51 Straßen- und Brückenunterhaltung, Straßenverkehrstechnik
63 Hausumringe Bremerhaven
65 Fäll- und Schnittgenehmigung nach der Bremischen Baumschutzverordnung
79 Spritzen - gebrauchte - Entsorgung
133 Briefwahl

#### Service Matched from ESD

Search	
1 Voting	↑ ↔ ↓
2 Proxy voting	↑ ↔ ↓
3 Postal voting	↑ ↔ ↓
4 Members - elections - polling stations	↑ ↔ ↓
5 Election results	↑ ↔ ↓
6 Members - elections - results	↑ ↔ ↓
7 Electoral register	↑ ↔ ↓

#### Service Info

• [Look at the retrieved services \(link recommendations\)](#)

• [Link](#)  
skos broader / exact / narrower match

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR

## Cross-language Service Retriever

Home Info

Upload list Register Login

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

30

- Load a catalog

- Look at target service description (if available)

#### Service Catalog

- Scan the list or search for a specific service

Br	
51	Straßen- und Brückenunterhaltung, Straßenverkehrstechnik
63	Hausumringe Bremerhaven
65	Fäll- und Schnittgenehmigung nach der Bremischen Baumschutzverordnung
79	Spritzen - gebrauchte - Entsorgung
133	Briefwahl

#### Service Matched from ESD

Search	
1	Voting
2	Proxy voting
3	Postal voting
4	Members - elections - polling stations
5	Election results
6	Members - elections - results
7	Electoral register

#### Service Info

<http://id.esd-toolkit.eu/service/1013>  
Provision of a facility whereby people who cannot attend the polling station on an election day can have postal ballot papers sent to them.

- Look at the link recommendations
- Link  
skos broader / exact / narrower match

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR

## Cross-language Service Retriever

[Home](#) [Info](#)

[Upload list](#) [Register](#) [Login](#)

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

30

#### Service Catalog

Br
51 Straßen- und Brückenunterhaltung, Straßenverkehrstechnik
63 Hausumringe Bremerhaven
65 Fäll- und Schnittgenehmigung nach der Bremischen Baumschutzverordnung
79 Spritzen - gebrauchte - Entsorgung
133 Briefwahl

#### Service Matched from ESD

Search	
1 Voting	↑ ↔ ↓
2 Proxy voting	↑ ↔ ↓
3 Postal voting	↑ ↔ ↓
4 Members - elections - polling stations	↑ ↔ ↓
5 Election results	↑ ↔ ↓
6 Members - elections - results	↑ ↔ ↓
7 Electoral register	↑ ↔ ↓

#### Service Info

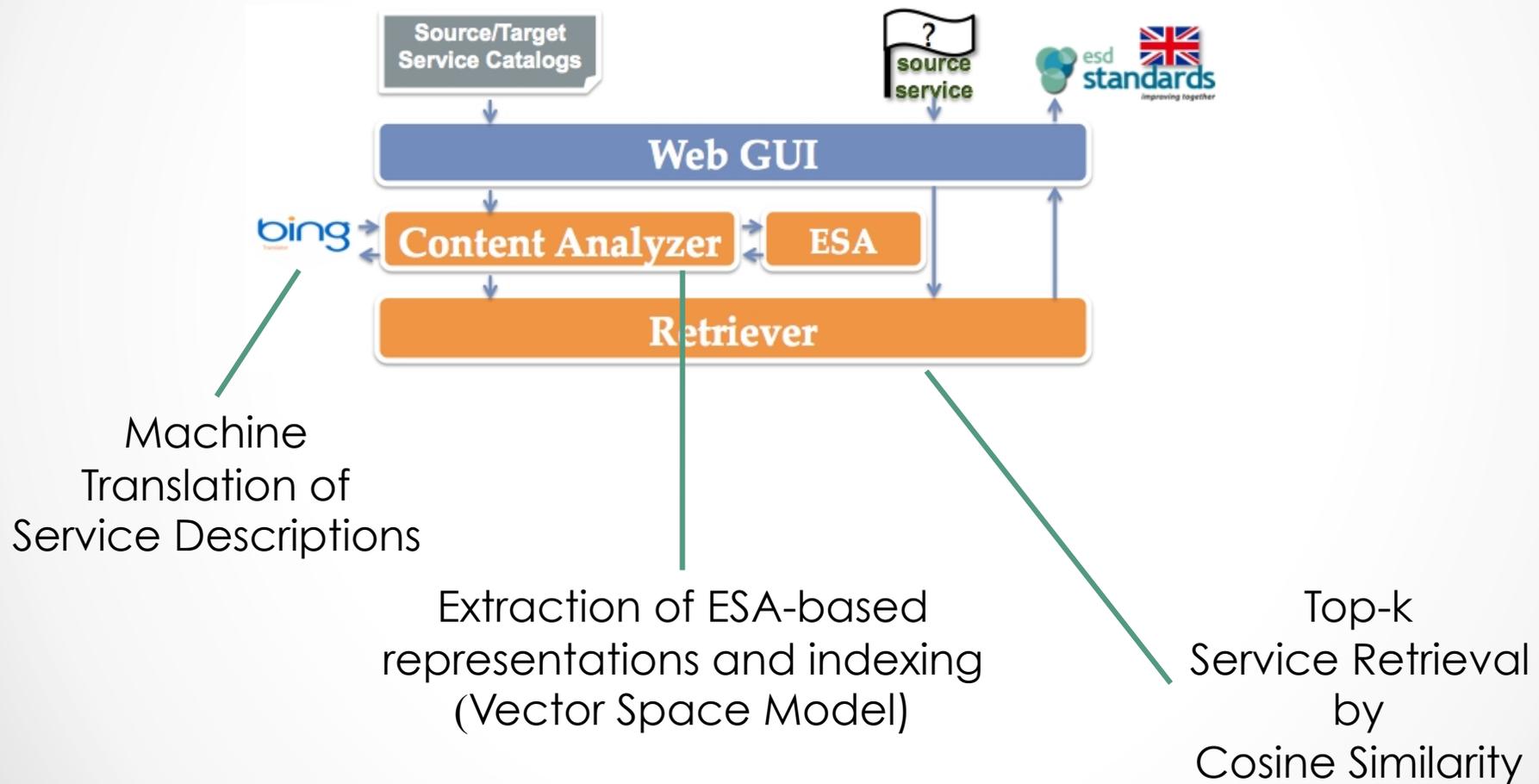
<http://id.esd-toolkit.eu/service/1013>

Provision of a facility whereby people who cannot attend the polling station on an election day can have postal ballot papers sent to them.

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# CroSeR: Matching Approach



# Explicit Semantic Analysis in CroSeR

Wikipedia-based representation of natural language expressions with the ESA matrix

		Wikipedia articles				
		ESA	Job Interview	Employment Agency	...	Unemployment benefits
Terms occurring in Wikipedia articles	unemployment		0,65	0,84	...	0,92
	...		TF-IDF	TF-IDF	TF-IDF	TF-IDF
	Term k		TF-IDF	TF-IDF	Tf-IDF	TF-IDF

- A **set of terms** is represented by the centroid of the vectors associated with the individual terms
  - E.g.: “Unemployment Support” → Job Interview (0,42), Employment Agency (0.55), ..., Unemployment Benefits(0.62)
- Feature generation + light-weight disambiguation

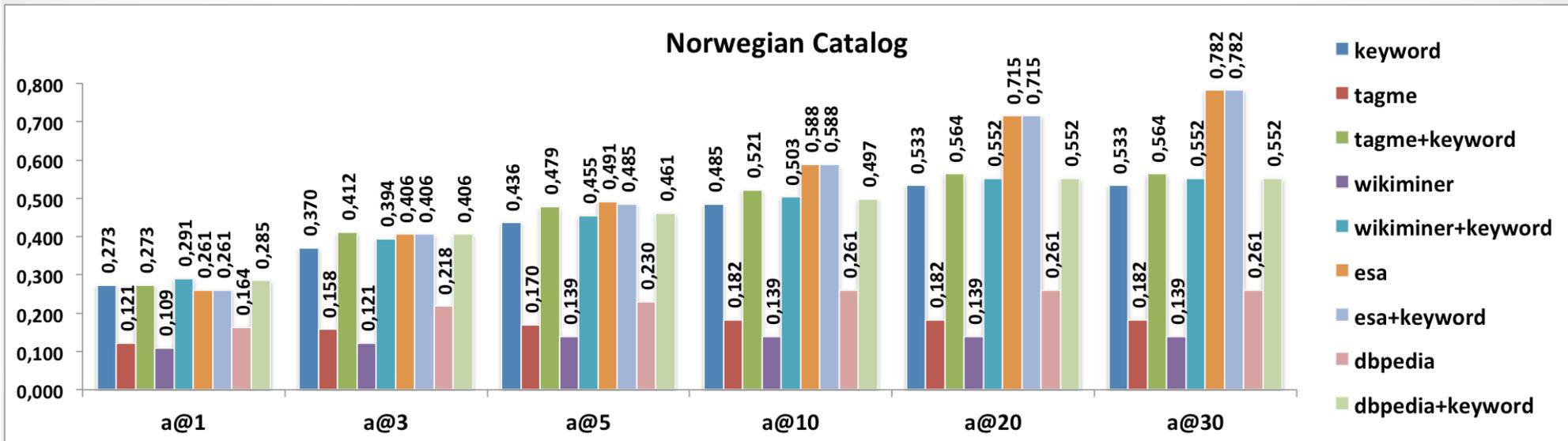
# Experimental Evaluation: Design

- Dataset
  - Any language LGSL vs. English LGLS
    - Dutch (#225)
    - German (#190)
    - Flemish (#341)
    - Norwegian (#165)
    - Swedish (#66)
  - TOT = #997 vs. #1425
- Methodology
  - **Gold standard:** ≈sameAs links defined by human experts in the LGSL
  - **Accuracy** @1 ... @30
  - **MRR** (Mean Reciprocal Rank )
- Comparative evaluation against baseline and other techniques based on similar principles\*
  - **Esa**
  - **Esa + keyword**  
vs
  - Keyword (baseline)
  - Tagme
  - Tagme + keyword
  - Wikiminer
  - Wikiminer + keyword
  - Dbpedia Spotlight
  - Dbpedia Sp + keyword

\*Experiments with CL-ESA [Sorg et al. 2'12] have also been carried out

# Experiments: Accuracy

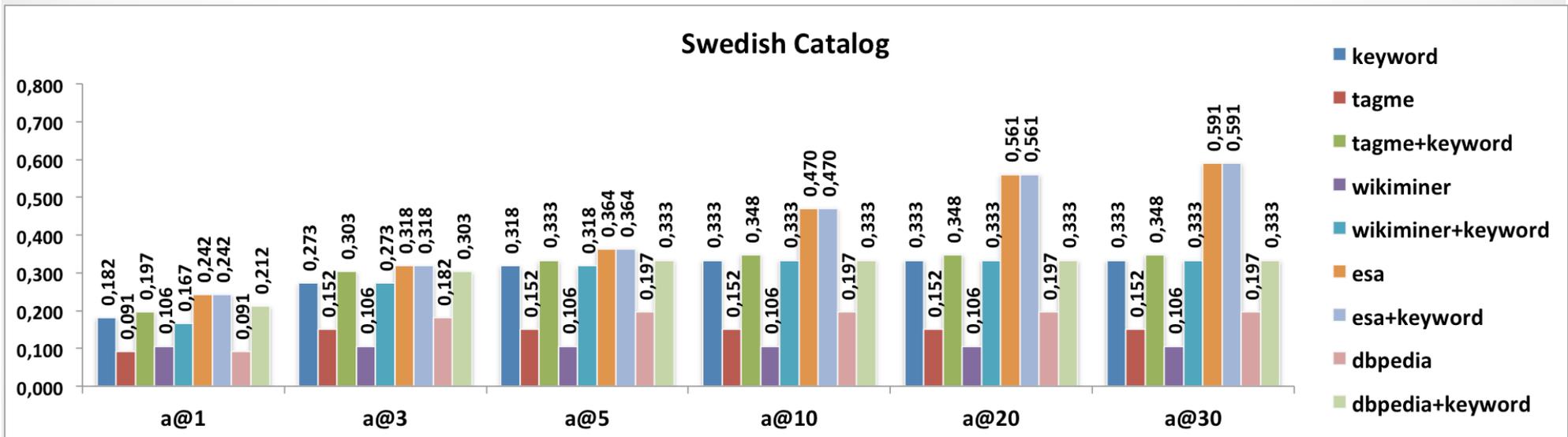
## Dataset in which Best Average Accuracy



- Best approach in terms of Accuracy@ $n$  is ESA
- ESA becomes more effective when more services are returned
- Merging the Wikipedia-based representation with keywords does not improve accuracy in ESA

# Experiments: Accuracy

## Dataset in which Worst Average Accuracy



- Low accuracy for every approach
- Best approach in terms of Accuracy@ $n$  is ESA
- ESA relative performance is more evident
  - ESA more effective for any  $n$

# Experiments: Mean Reciprocal Rank

Representation	Dutch	Belgian	German	Norwegian	Swedish
keyword	<b>0.333</b>	0.320	0.242	0.273	0.182
tagme	0.120	0.094	0.147	0.121	0.091
tagme+keyword	0.316	0.334	0.258	0.273	0.197
wikifi	0.080	0.114	0.116	0.109	0.106
wikifi+keyword	0.324	0.326	0.258	<b>0.291</b>	0.167
esa	0.311	0.326	<b>0.289</b>	0.261	<b>0.242</b>
esa+keyword	0.311	<b>0.328</b>	<b>0.289</b>	0.261	<b>0.242</b>
dbpedia	0.182	0.202	0.163	0.164	0.091
dbpedia+keyword	0.329	0.334	0.274	0.285	0.212

- ESA average rank: between 3rd and 5th position
- Suboptimal MRR in two datasets: higher coverage is achieved under the condition that the list of retrieved services is extended

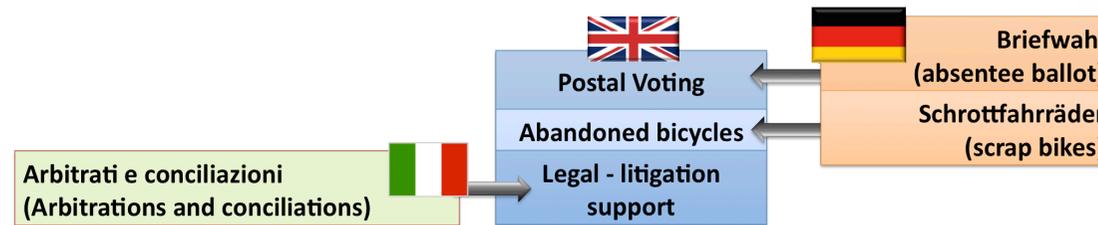
# Experiments: Mean Reciprocal Rank

Representation	Dutch	Belgian	German	Norwegian	Swedish
keyword	<b>0.333</b>	0.320	0.242	0.273	0.182
tagme	0.120	0.094	0.147	0.121	0.091
ta	Difficult matching task: user intervention is needed				
wikifi	0.080	0.114	0.116	0.109	0.106
wikifi+keyword	0.224	0.226	0.258	0.201	0.167
	Looking at a reasonable number of recommendations significantly reduce the linking effort (e.g., 30/1425)				
dbpedia	0.182	0.202	0.163	0.164	0.091
dbpedia+keyword	0.329	0.334	0.274	0.285	0.212

- ESA average rank: between 3rd and 5th position
- Suboptimal MRR in two datasets: higher coverage is achieved under the condition that the list of retrieved services is extended

# Experiments: Discussion

- CroSeR finds matchings that cannot be discovered by machine translation + keyword comparison



- CroSeR's recommendations can support the users to refine the links

“Absentee Ballot”

The screenshot shows two columns of search results. The left column contains results for the German term "Briefwahl" (133 results), and the right column contains results for the English term "Postal voting" (3 results). A red arrow points from the "133 Briefwahl" result to the "3 Postal voting" result. A green arrow points from the "133 Briefwahl" result to the "2 Proxy voting" result. A red circle with an equals sign is placed near the "133 Briefwahl" result. A blue oval highlights the "3 Postal voting" result. A blue oval also highlights the "2 Proxy voting" result. A blue oval highlights the "133 Briefwahl" result. A blue oval highlights the "3 Postal voting" result. A blue oval highlights the "2 Proxy voting" result.

51	Straßen- und Brückenunterhaltung, Straßenverkehrstechnik	1	Voting
63	Hausumringe Bremerhaven	2	Proxy voting
65	Fäll- und Schnittgenehmigung nach dem Baumschutzverordnung	3	Postal voting
79	Spritzen - gebrauchte - Entsorgung	4	Members - elections - polling stations
133	Briefwahl	5	Election results
		6	Members - elections - results

# Conclusions & Future Work

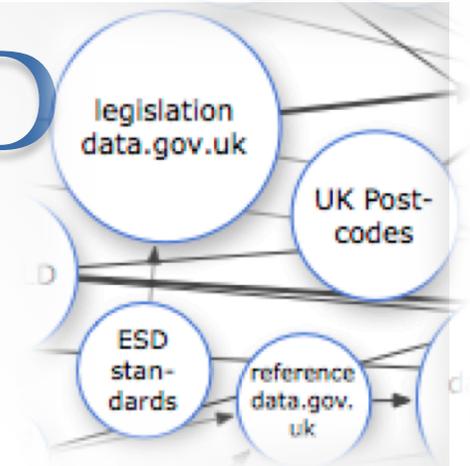
- Summary
  - CroSeR uses cross-language matching to recommend links among eGov service descriptions available in different languages
  - Good performance in semi-automatic linking settings
  - Unsupervised method based on Explicit Semantic Analysis
  - Language independent
- Future work
  - Collect additional information to improve the overall results
  - User study to further evaluate the quality of the recommendations (ongoing)
  - Application of similar approach to other cross-language matching problems

# Thanks, questions?

Demo at <http://siti-rack.siti.disco.unimib.it:8080/croser/>



# eGov Services LOD ESD & LGSL



- Electronic Service Delivery (ESD)-toolkit
  - Define the semantics of public sector services
  - The SmartCities project
    - Innovation network in the domain of the development and uptake of e-services in the whole North Sea region.
    - England, Netherlands, Belgium, Germany, Scotland, Sweden, and Norway
- European Local Government Service List (LGSL)
  - Each country responsible to build and maintain its list of public services
  - All of those services are interlinked to the services delivered by other countries
  - Linked to the LOD cloud

The screenshot shows the user interface of the European Local Government Service List (LGSL). On the left, there is a search bar with the word 'search' and a magnifying glass icon. Below the search bar is a 'Service List' section with the text 'Order by text'. A list of services is displayed, including 'Schools - home schooling (1)', 'Education - grants- school clothing grants/vouchers (2)', 'Education - grants - home to education establishment travel support (3)', 'Education - grants - free school meals (4)', 'Schools - supervised medication (5)', 'Schools - language and cultural support (6)', and 'Psychological, psychiatric or social work services in schools (7)'. On the right, there is a detailed view for the selected service 'Schools - home schooling (1)'. This view includes a search bar, a title 'Schools - home schooling (1)', and a description: 'Information on the list item, including the lists in which it appears.' Below the description, there are fields for 'Id', 'Name', 'Definition', 'In scheme', 'History notes', and 'Same as'. The 'Definition' field contains the text: 'The education authority will arrange to visit a parent thinking about educating their child at home to help them plan the child's education. They will ensure that the child will receive efficient full-time education suitable to their age, ability and any special needs.' The 'In scheme' field lists 'Norwegian Service List (this item)' and 'Scottish Service List (this item)'. The 'History notes' field contains the text: 'Added scope notes in version 2.02. Term name changed from 'Educating your child at home' to 'Schools - home schooling' in version 3.00.' The 'Same as' field lists 'Home schooling (1)'.

# CroSeR

## Cross-language Service Retriever

Home Info

Upload list Register Login

### CROSER: Cross-language Service Retriever

connects your service catalogs

Italian Dutch Belgian German Swedish Norwegian

Load a catalog

30

Look at target service description (if available)

#### Service Catalog

Scan the list or search for a specific service

Br	
51	Straßen- und Brückenunterhaltung, Straßenverkehrstechnik
63	Hausumringe Bremerhaven
65	Fäll- und Schnittgenehmigung nach der Bremischen Baumschutzverordnung
79	Spritzen - gebrauchte - Entsorgung
133	Briefwahl

#### Service Matched from ESD

Search	
1	Voting
2	Proxy voting
3	Postal voting
4	Members - elections - polling stations
5	Election results
6	Members - elections - results
7	Electoral register

#### Service Info

<http://id.esd-toolkit.eu/service/1013>

Provision of a facility whereby people who cannot attend the polling station on an election day can have postal ballot papers sent to them.

Look at the link recommendations

Establish the link

skos broader / exact / narrower match

Web tool to **support the linkage** of a source eGov service catalog represented in **any language** to a target catalog represented in **English**

Based on **Machine Translation** and **Explicit Semantic Analysis (ESA)**

# ESA: Matrix

- The semantic relatedness between a term and a Wikipedia concept (article) is in terms of **TF-IDF score**

**Wikipedia articles**

	ESA	Concept 1	...	Concept n
Terms occurring in Wikipedia articles	Term 1	TF-IDF	TF-IDF	TF-IDF
	...	TF-IDF	TF-IDF	TF-IDF
	Term k	TF-IDF	TF-IDF	TF-IDF

- Several heuristics are applied in order to reduce the number of concepts and terms

# ESA: Wikipedia Concepts and Terms

Every Wikipedia **article** represents a **concept**

## Panthera

From Wikipedia, the free encyclopedia

***Panthera*** is a **genus** of the **family Felidae** (the cats), which contains four well-known living **species**: the **lion**, **tiger**, **jaguar**, and **leopard**. The genus comprises about half of the **big cats**. One meaning of the word ***panther*** is to designate cats of this family. Only these four cat species have the anatomical changes enabling them to **roar**. The primary reason for this was assumed to be the incomplete **ossification** of the **hyoid bone**. However, new studies show that the ability to roar is due to other **morphological** features, especially of the **larynx**. The **snow leopard**, *Uncia uncia*, which is sometimes included within *Panthera*, does not roar. Although it has an incomplete ossification of the hyoid bone, it lacks the special morphology of the larynx, which is typical for lions, tigers, jaguars and leopards.<sup>[1]</sup>

Species and subspecies

[edit]

<i>Panthera</i>
<span></span> <div>Tiger</div>
Scientific classification
Kingdom: <span>Animalia</span>
Phylum: <span>Chordata</span>

Panthera

Cat [0.92]

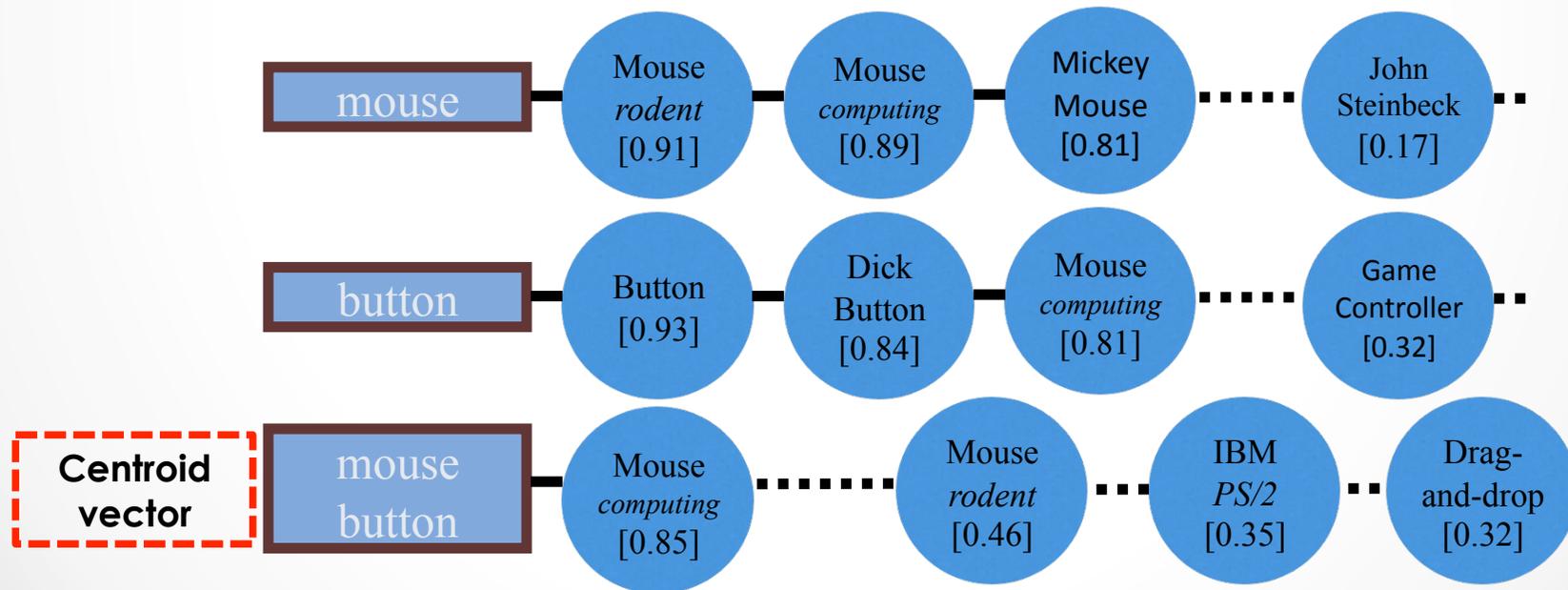
Leopard [0.84]

Roar [0.77]

Article **words (terms)** are associated with the **concept** (TF-IDF)

# ESA: Text Fragment Interpretation

The **semantic interpretation vector** of a text fragment is the **centroid** vector of the terms occurring in



# Experiment: Design

- Metrics

- **Accuracy@n**: is calculated considering only the first n retrieved services. If the correct service occurs in the top-n items, the service is marked as correctly retrieved (n = 1; 3; 5; 10; 20; 30)

- **MRR**

$$MRR = \frac{\sum_{i=1}^N \frac{1}{rank_i}}{N}$$

$rank_i$  is the rank of the correctly retrieved service  $i$  in the ranked list, and  $N$  is the number of the services correctly retrieved **with the configuration**.

- Motivation
- CroSeR
- Experimental Evaluation
- Conclusions

<b>Catalog</b>	<b>a@1</b>	<b>a@3</b>	<b>a@5</b>	<b>a@10</b>	<b>a@20</b>	<b>a@30</b>
Dutch				0.01	0.01	0.01
Belgian	0.05	0.01	0.01	0.01	0.01	0.01
German				0.01	0.01	0.01
Norwegian				0.01	0.01	0.01
Swedish				0.01	0.01	0.01