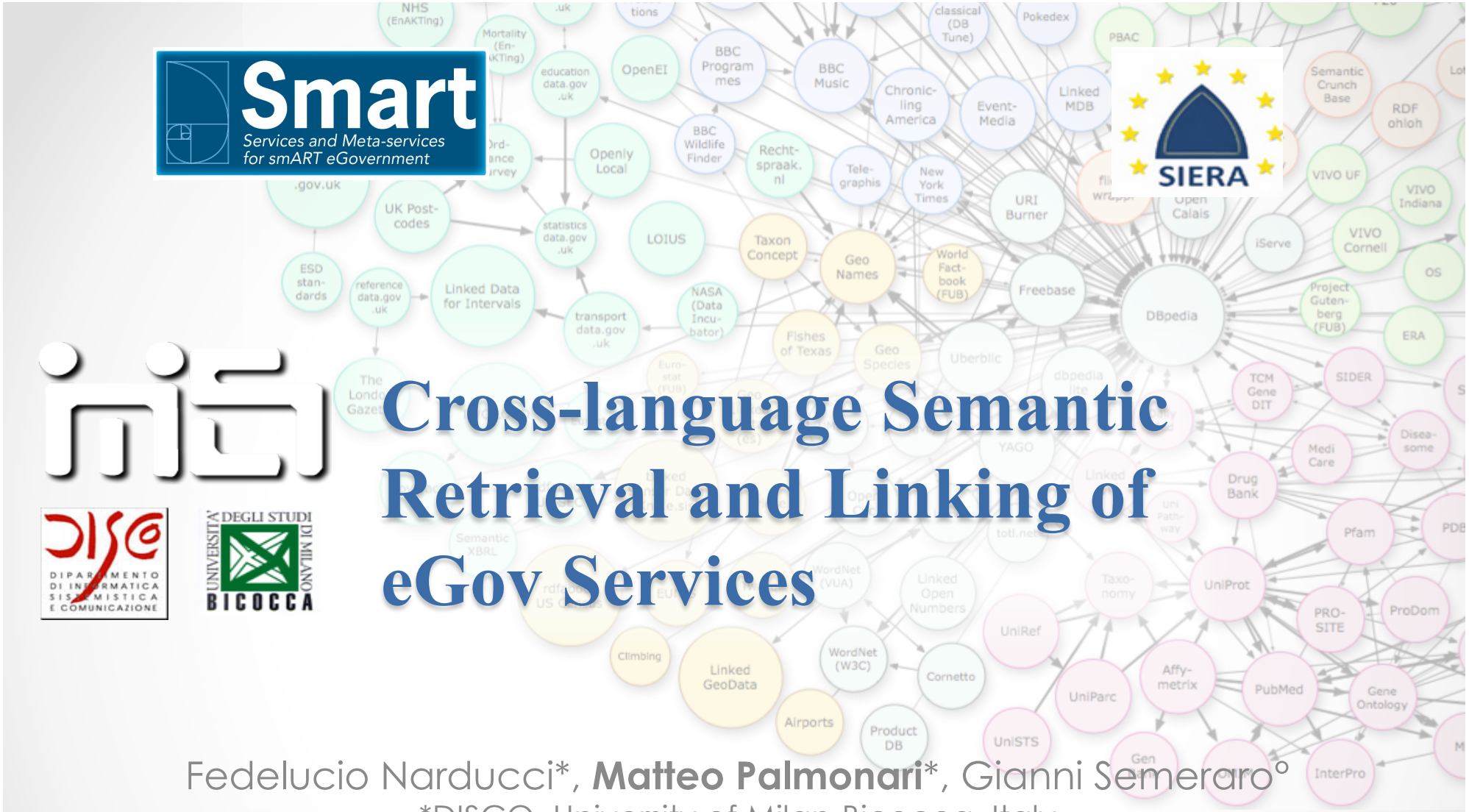


Cross-language Semantic Retrieval and Linking of eGov Services

Fedelucio Narducci*, **Matteo Palmonari***, Gianni Semeraro°
*DISCO, University of Milan-Bicocca, Italy
*Department of Computer Science, University of Bari Aldo Moro, Italy

ISWC 2013
Sydney, Australia

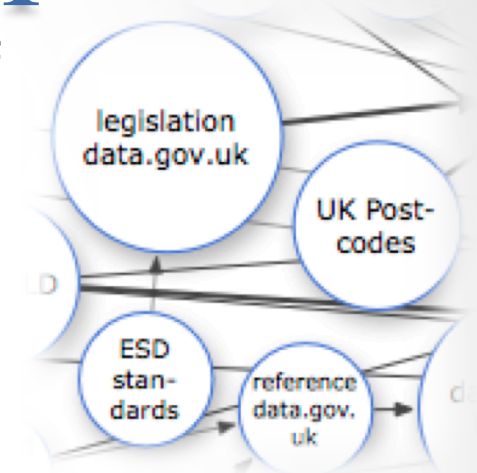
Cross-language Linking of Open Government Data

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

*http://logd.tw.rpi.edu/iogds data analytics

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

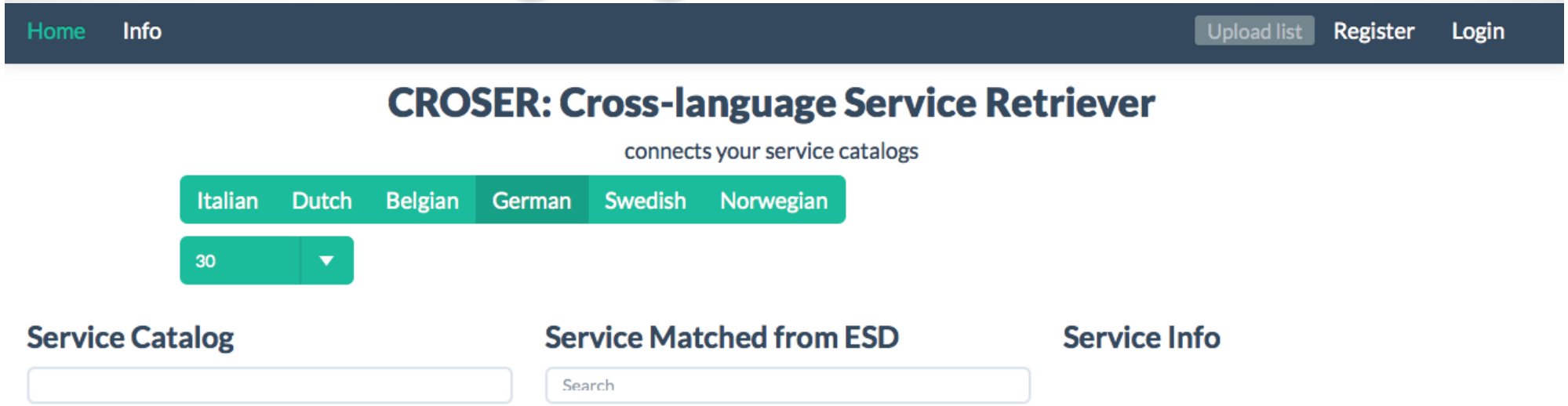
- not a mere "translation" problem
- cultural bias

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

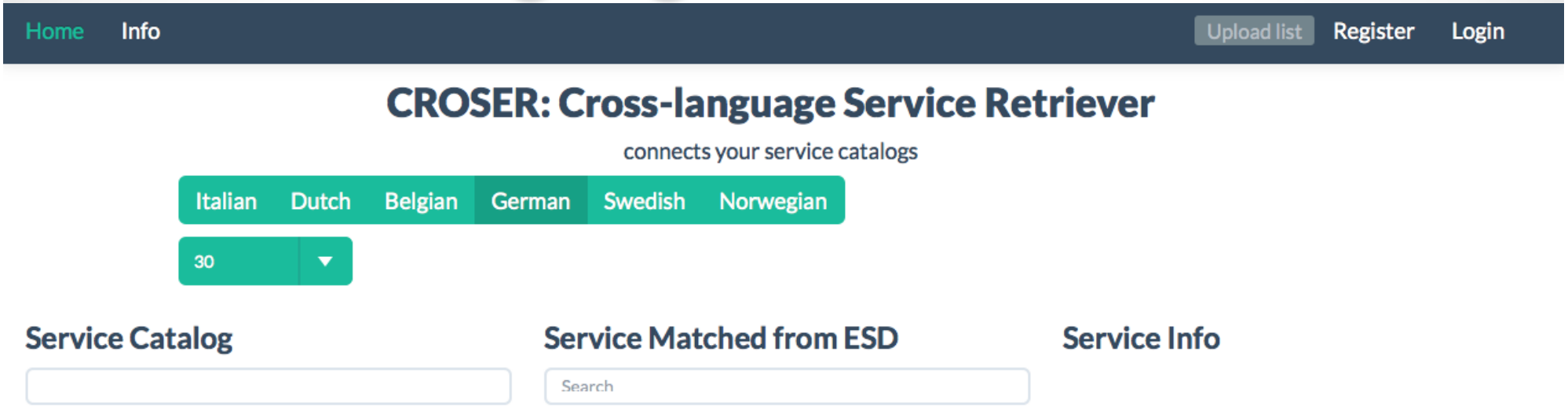


Web tool to **support the linkage** of a source eGov service catalog represented in **any language to** a target catalog represented in **English**

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

CroSeR

Cross-language Service Retriever

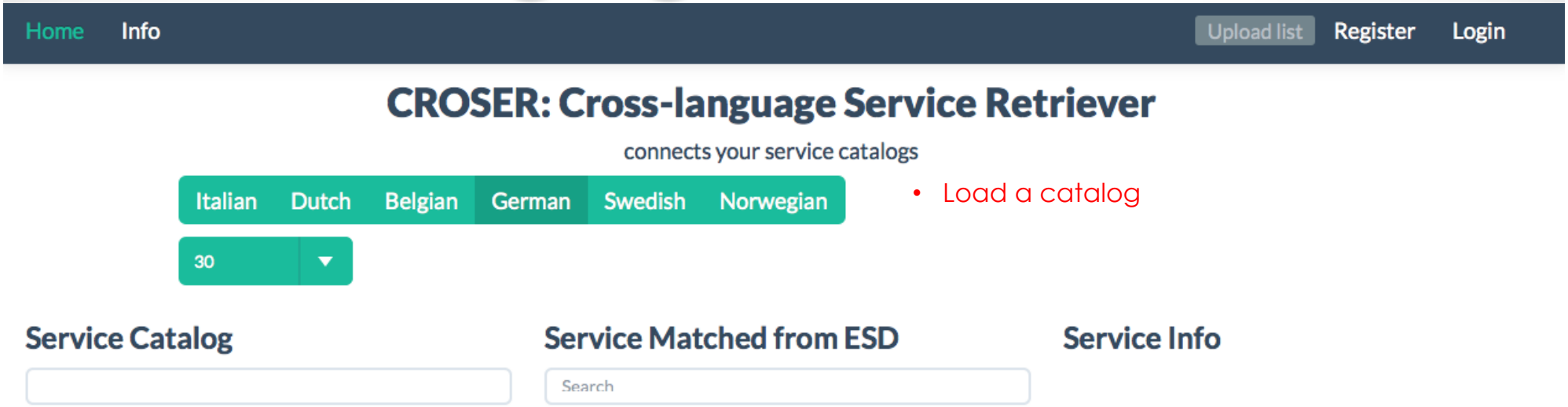


Web tool to **support the linkage** of a source eGov service catalog represented in **any language to** a target catalog represented in **English**

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

CroSeR

Cross-language Service Retriever



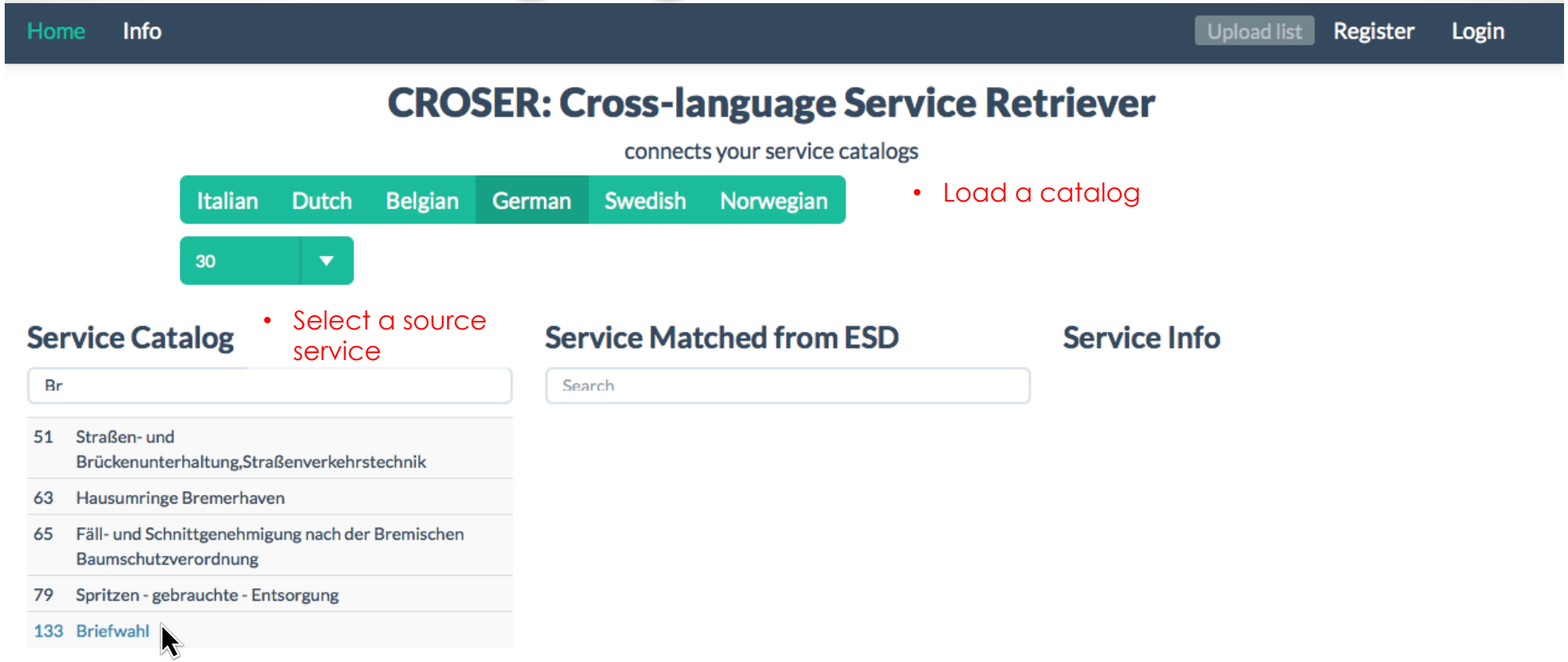
- Load a catalog

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

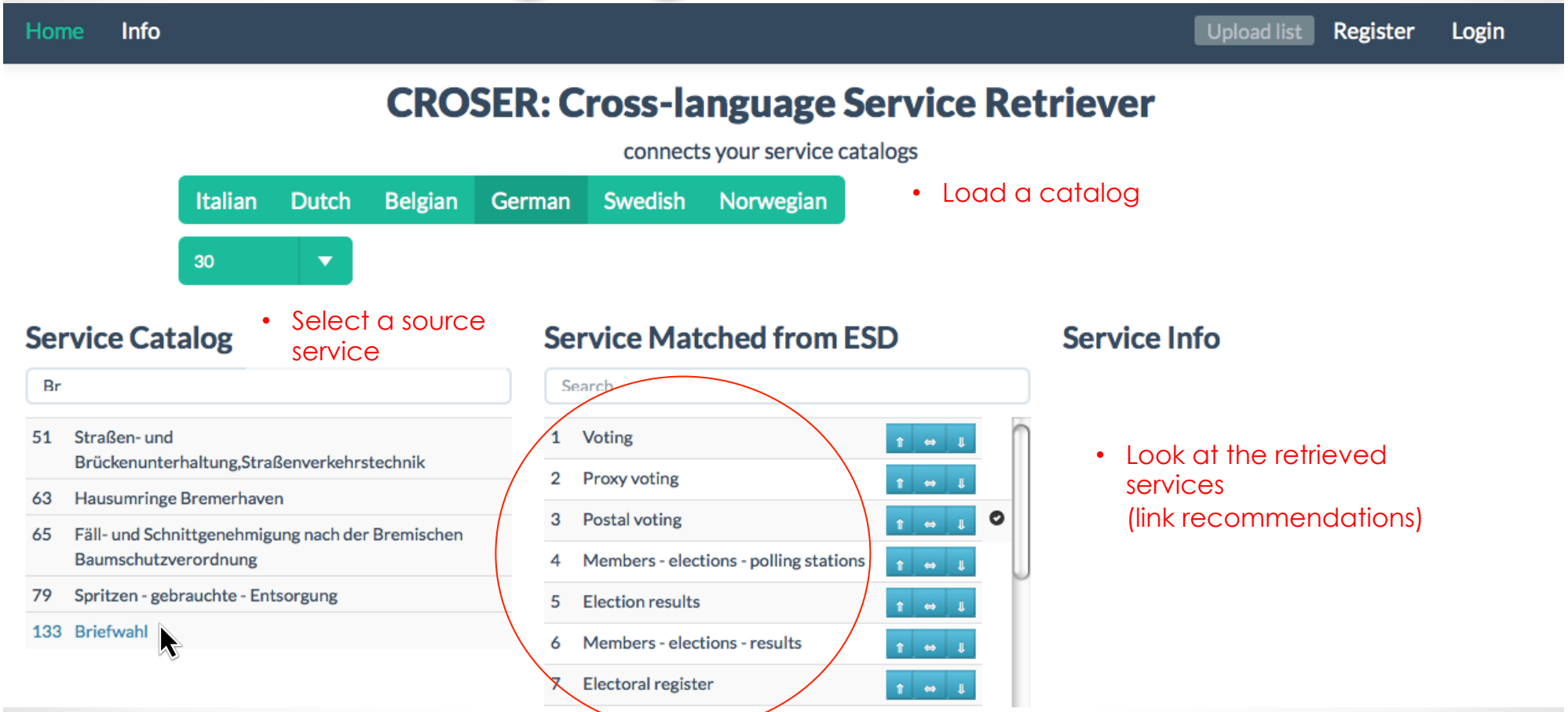


Web tool to **support the linkage** of a source eGov service catalog represented in **any language to** a target catalog represented in **English**

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

CroSeR

Cross-language Service Retriever

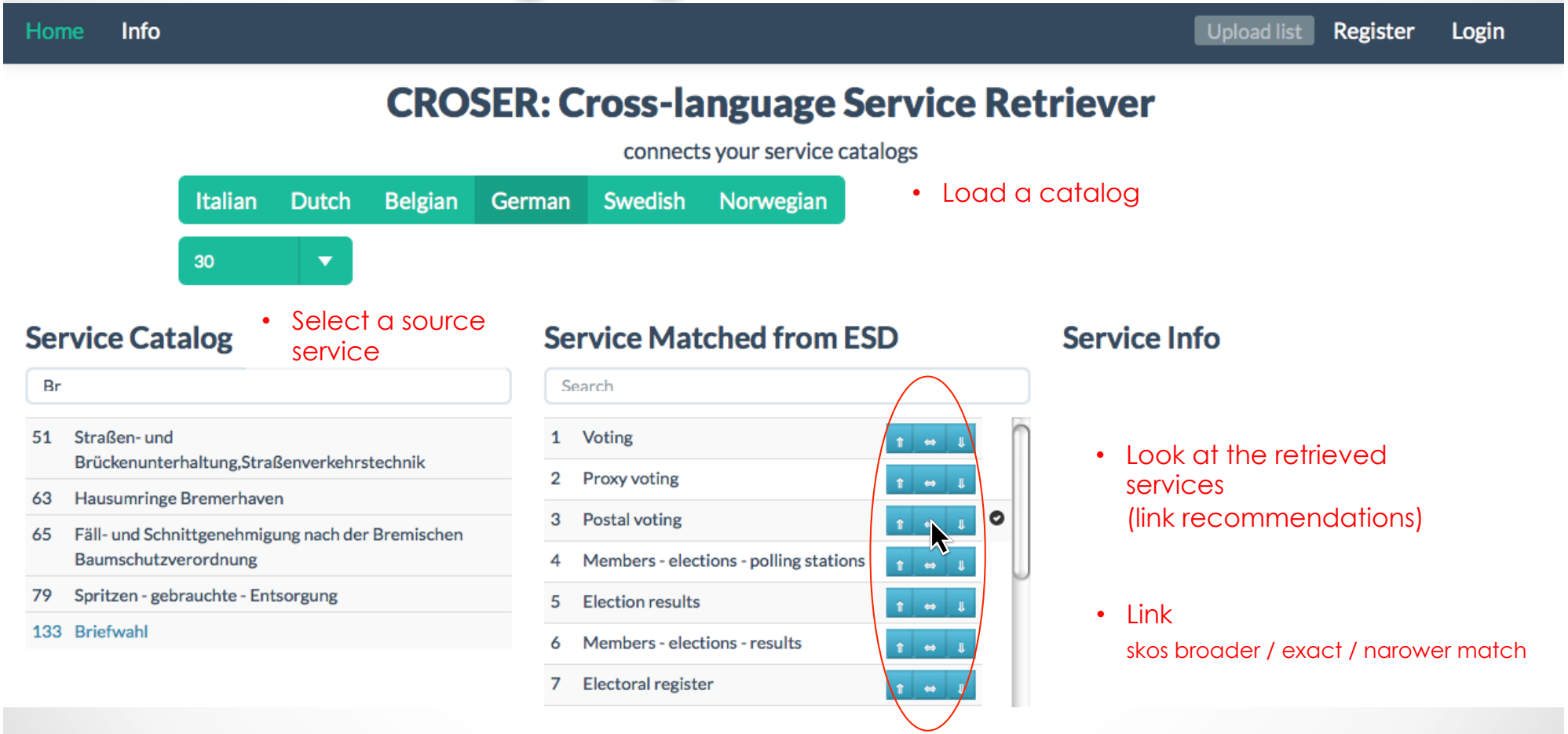


Web tool to **support the linkage** of a source eGov service catalog represented in **any language to** a target catalog represented in **English**

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

CroSeR

Cross-language Service Retriever

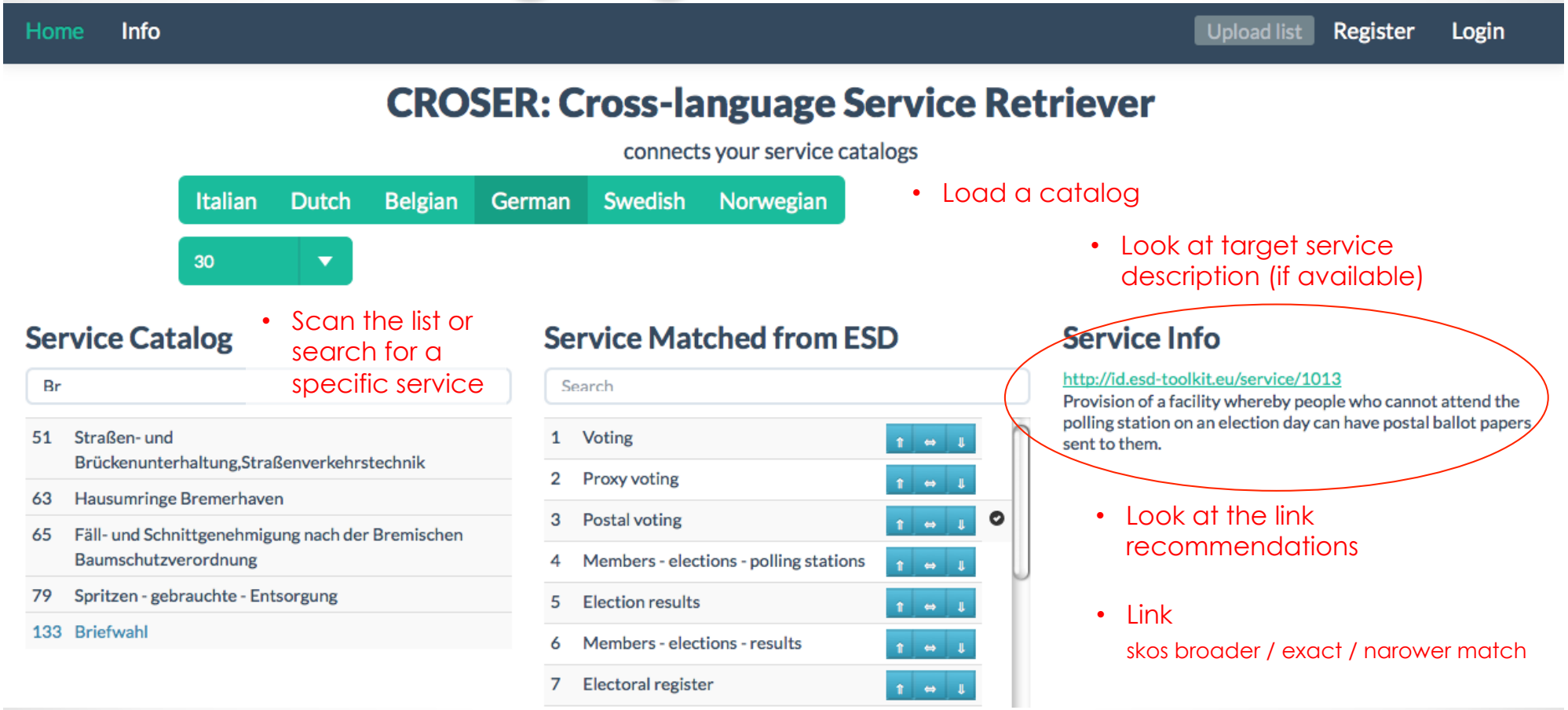


Web tool to **support the linkage** of a source eGov service catalog represented in **any language to** a target catalog represented in **English**

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

CroSeR

Cross-language Service Retriever



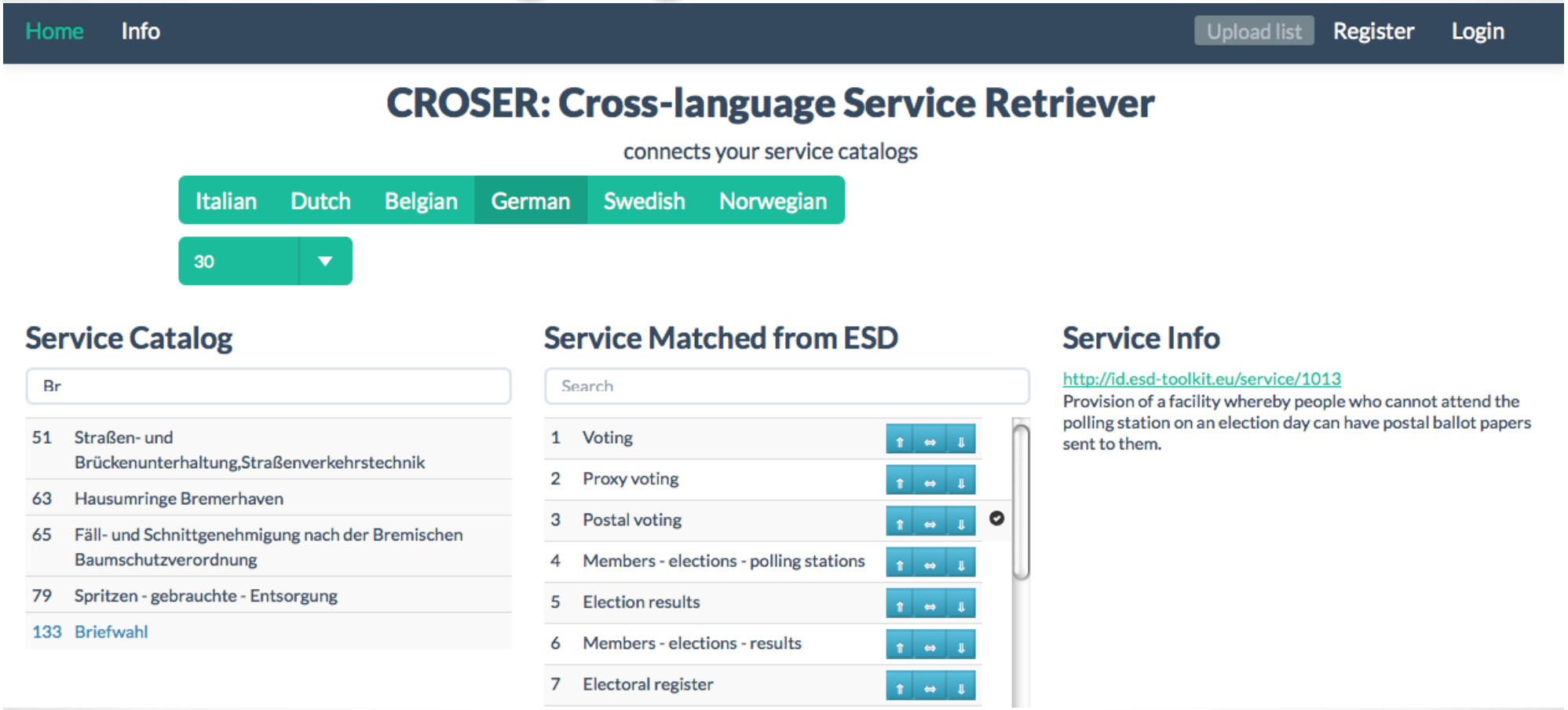
- Load a catalog
- Look at target service description (if available)
- Scan the list or search for a specific service
- Look at the link recommendations
- Link
  skos broader / exact / narower match

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

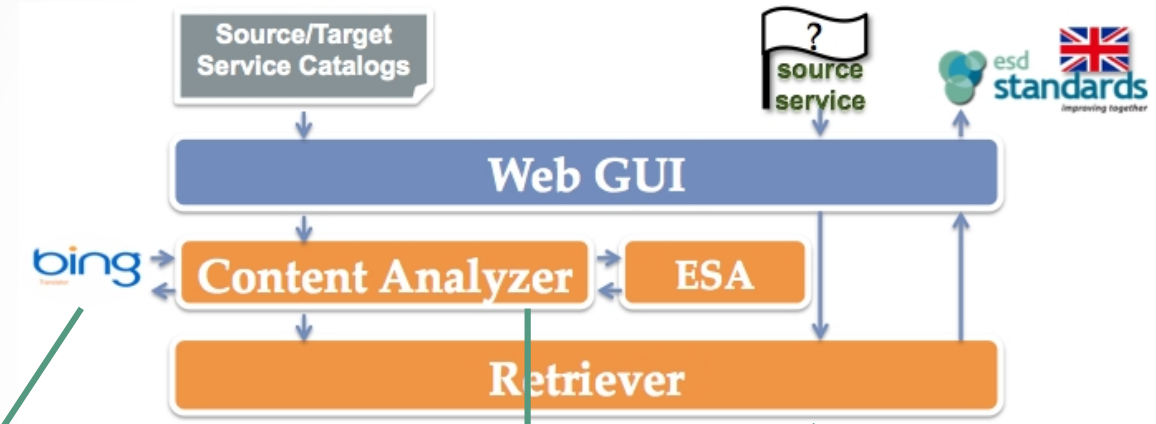
Cross-language Service Retriever



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 Approcah



Machine Translation of Service Descriptions

Extraction of ESA-based representations and indexing (Vector Space Model)

Top-k Service Retrieval by Cosine Similarity

Explicit Semantic Analysis in CroSeR

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

Wikipedia articles

Terms occurring in Wikipedia articles

ESA	Job Interview	Employment Agency	…	Unemployment benefits
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)
    - Norvegian (#165)
    - Swedish (#66)

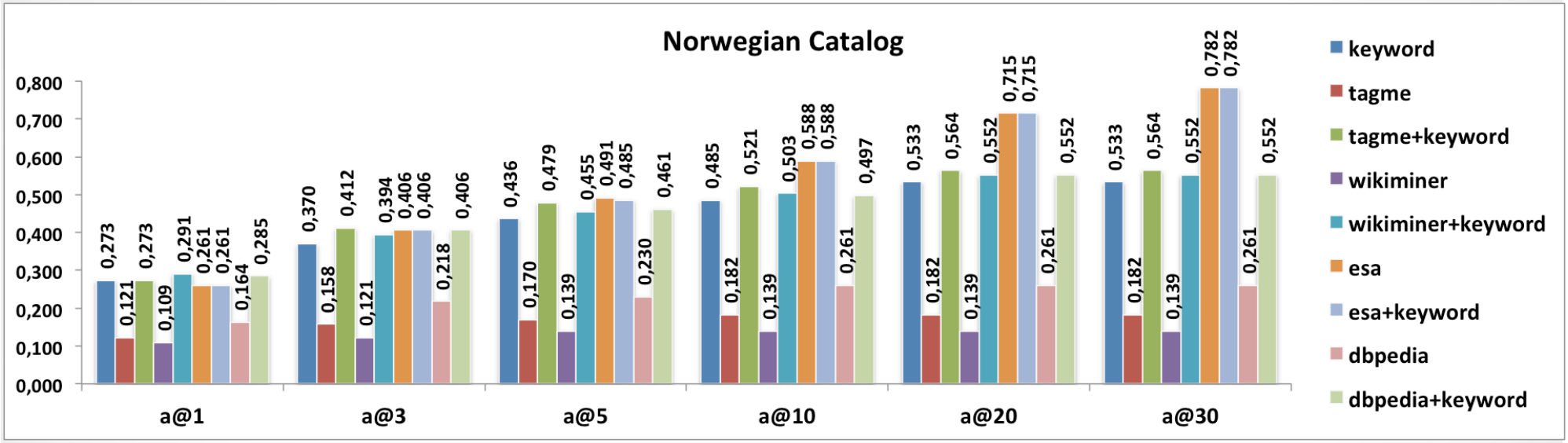
    TOT = #997 vs. #1425
- Methodology
  - **Gold standard:** ≈sameAs links defined by human experts in the LGSL
  - **Accuracy** @1 … @30
  - **MRR** (Mean Reciprocal Rank)
- Comparative evaluation against baseline and other techniques based on similar principles*
  - **Esa**
  - **Esa + keyword**

    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

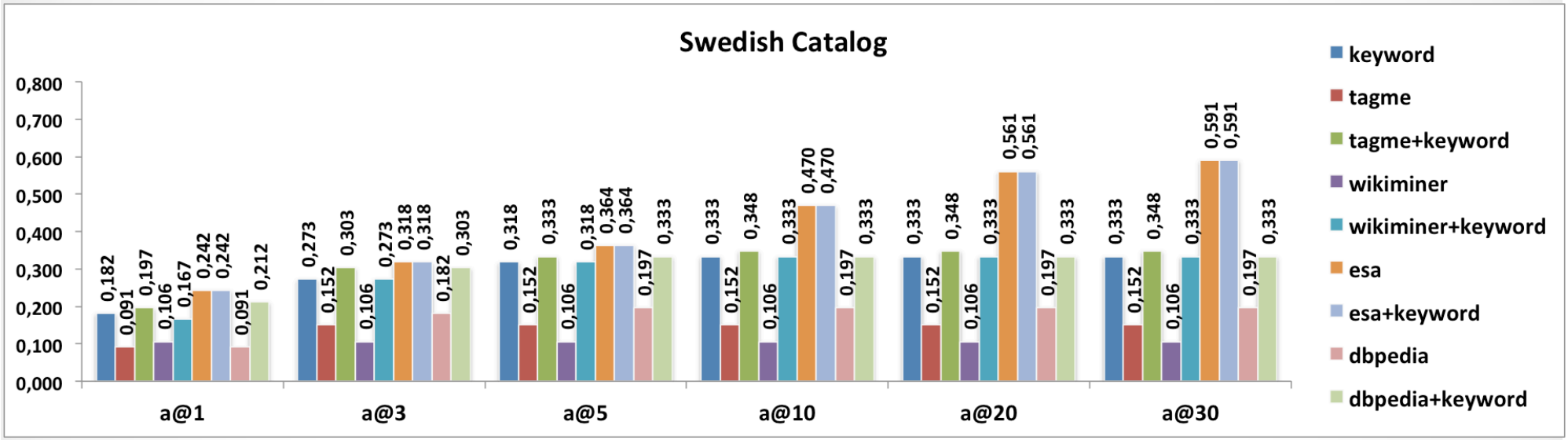
Dataset in which Best Average Accuracy



- 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	**0.333**	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	**0.291**	0.167
esa	0.311	0.326	**0.289**	0.261	**0.242**
esa+keyword	0.311	**0.328**	**0.289**	0.261	**0.242**
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	**0.333**	0.320	0.242	0.273	0.182
tagme	0.120	0.094	0.147	0.121	0.091
ta					
wikifi	0.080	0.114	0.116	0.109	0.106
wikifi+keyword	0.324	0.326	0.258	**0.291**	0.167
dbpedia	0.182	0.202	0.163	0.164	0.091
dbpedia+keyword	0.329	0.334	0.274	0.285	0.212

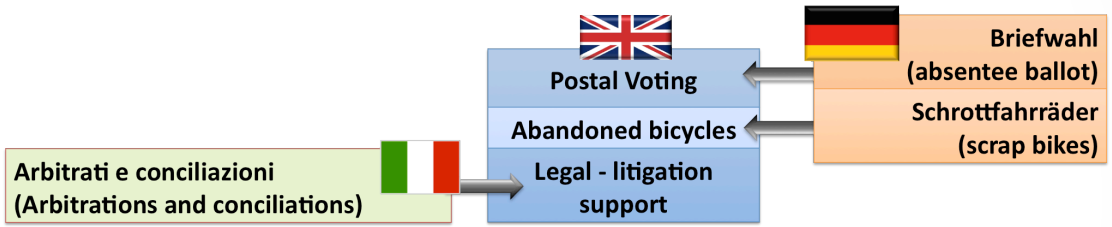
Difficult matching task: user intervention is needed

Looking at a reasonable number of recommendations significantly reduce the linking effort (e.g., 30/1425)

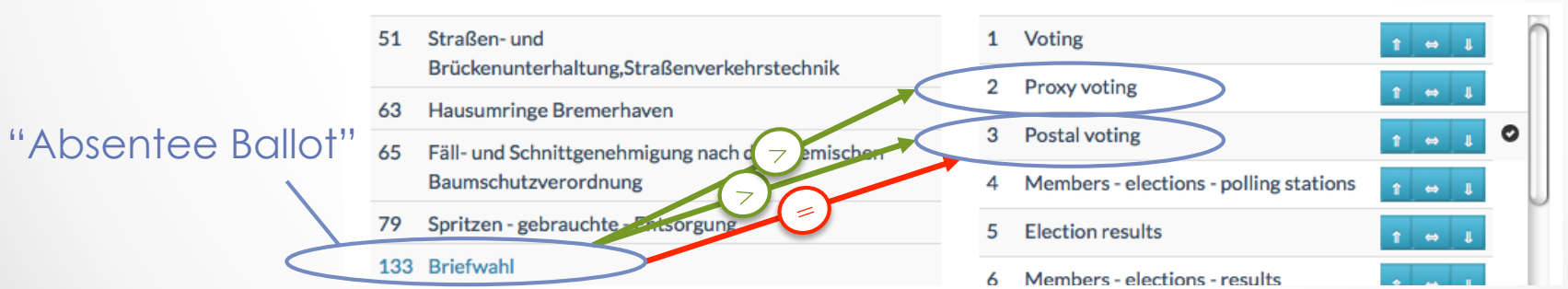
- ESA average rank: between 3rd and 5th position
- Suboptimal MRR in two datasets: higher coverage is achieved under the condition that the list of retrieved services is extended

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



Conclusions & Future Work

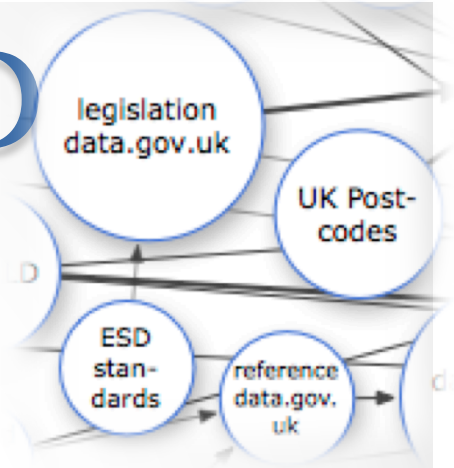
- Summary
  - CroSeR uses cross-language matching to recommend links among eGov service descriptions available in different languages
  - Good performance in semi-automatic linking settings
  - Unsupervised method based on Explicit Semantic Analysis
  - Language independent

- Future work
  - Collect additional information to improve the overall results
  - User study to further evaluate the quality of the recommandations (ongoing)
  - Application of similar approach to other cross-language matching problems

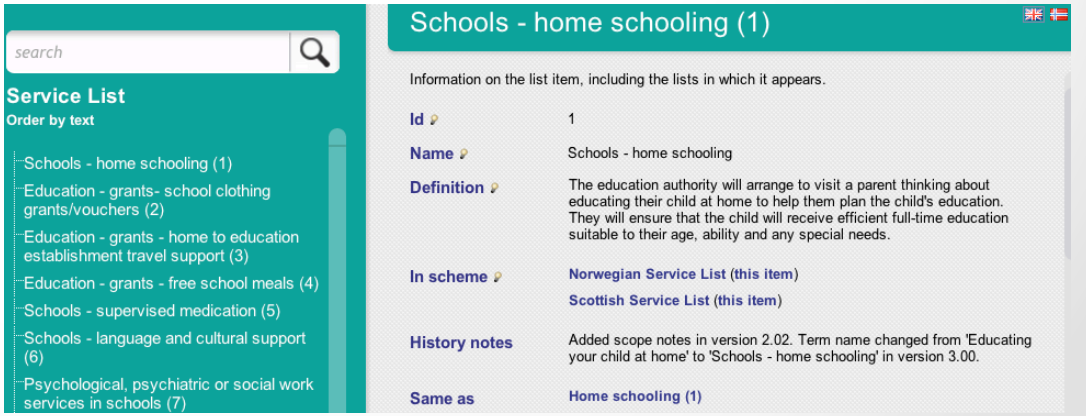
Thanks, questions?

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

eGov Services LOD
ESD & LGSL

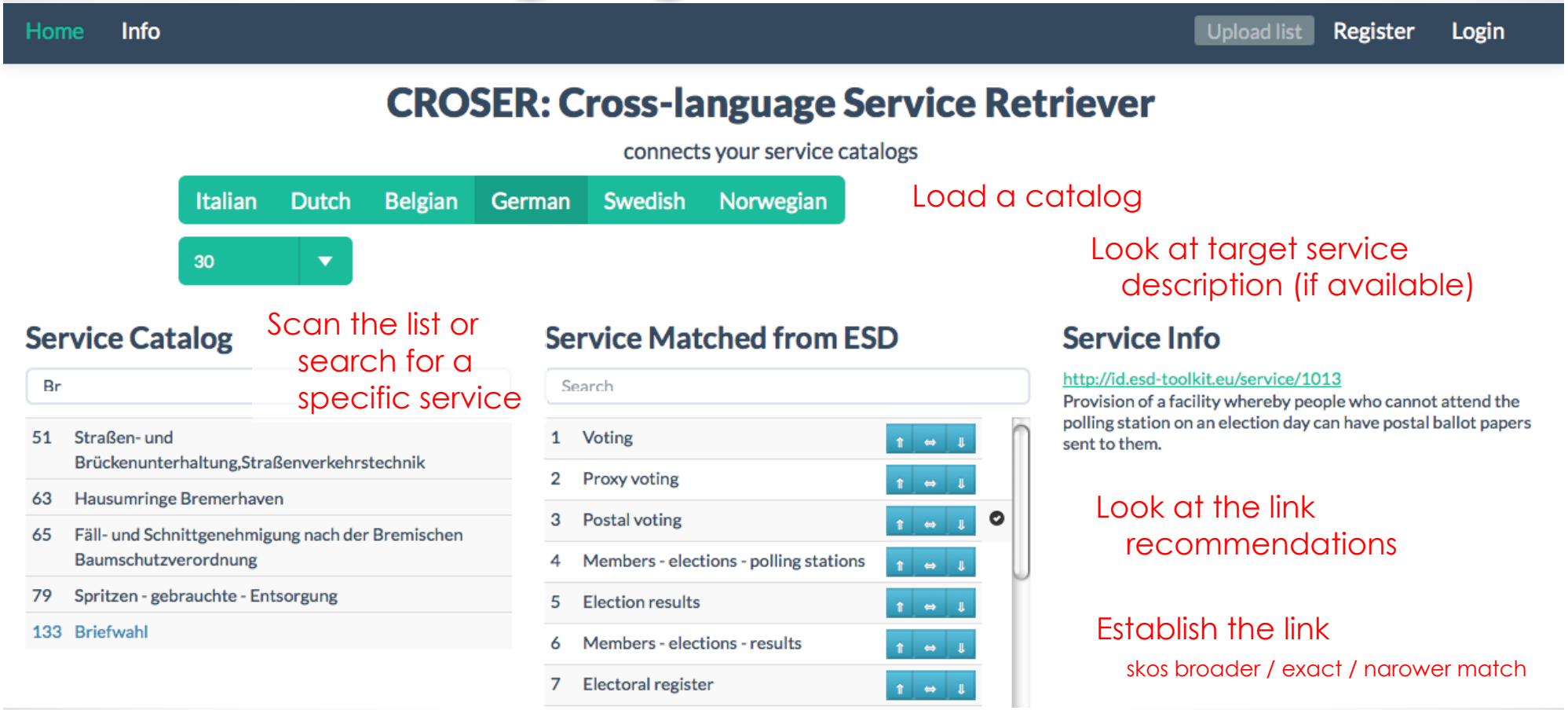


- Electronic Service Delivery (ESD)-toolkit
  - Define the semantics of public sector services
  - The SmartCities project
    - Innovation network in the domain of the development and uptake of e-services in the whole North Sea region.
    - England, Netherlands, Belgium, Germany, Scotland, Sweden, 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



CroSeR

Cross-language Service Retriever

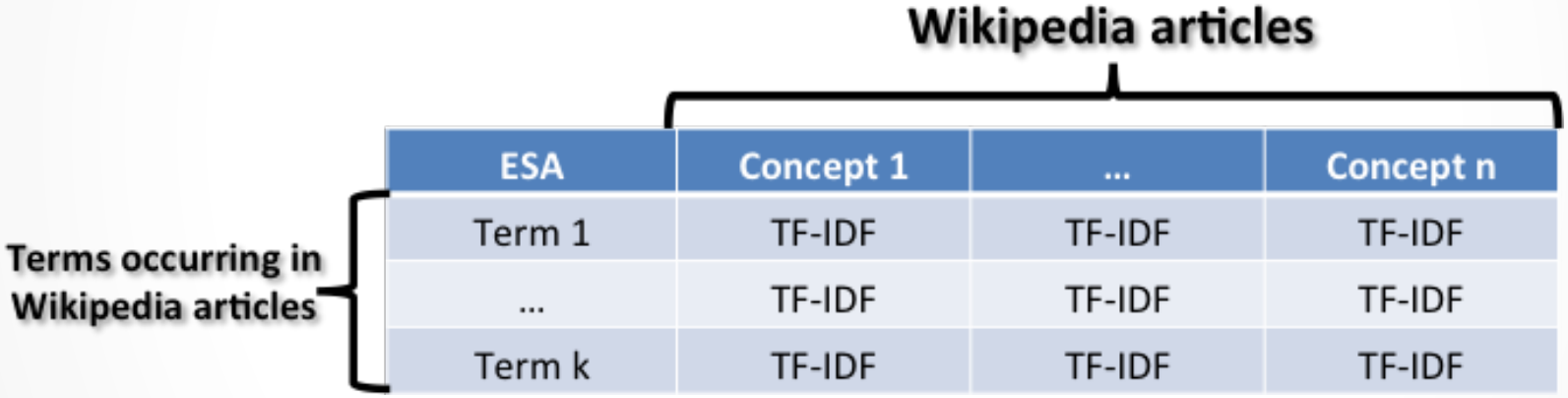


Web tool to **support the linkage** of a source eGov service catalog represented in **any language to** a target catalog represented in **English**

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**

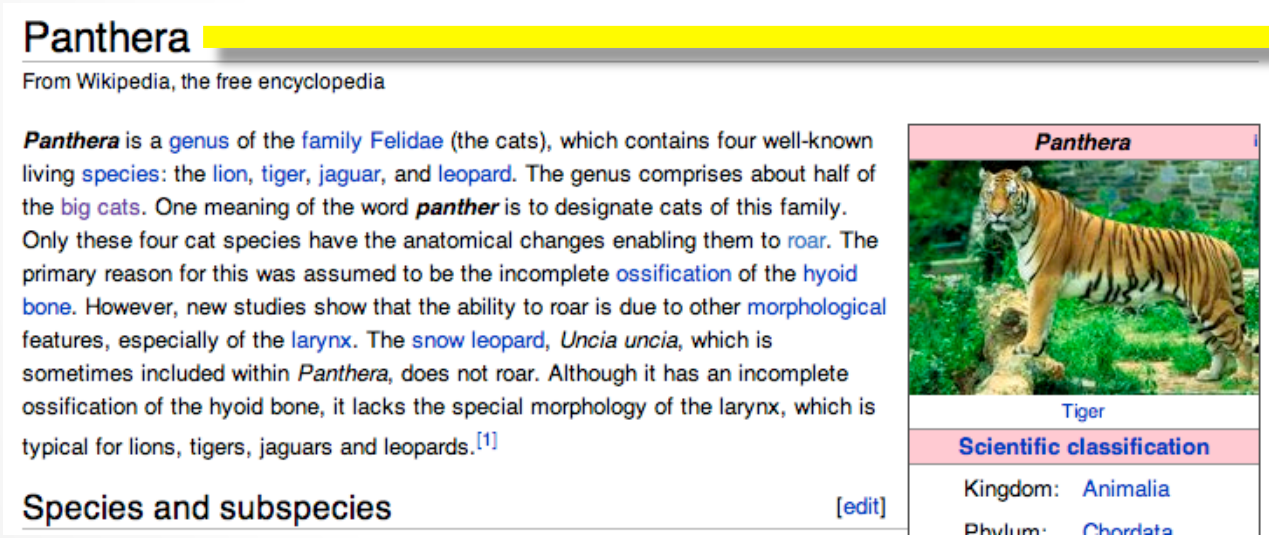


ESA	Concept 1	...	Concept n
Term 1	TF-IDF	TF-IDF	TF-IDF
...	TF-IDF	TF-IDF	TF-IDF
Term k	TF-IDF	TF-IDF	TF-IDF

- 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.[1]

Species and subspecies [edit]

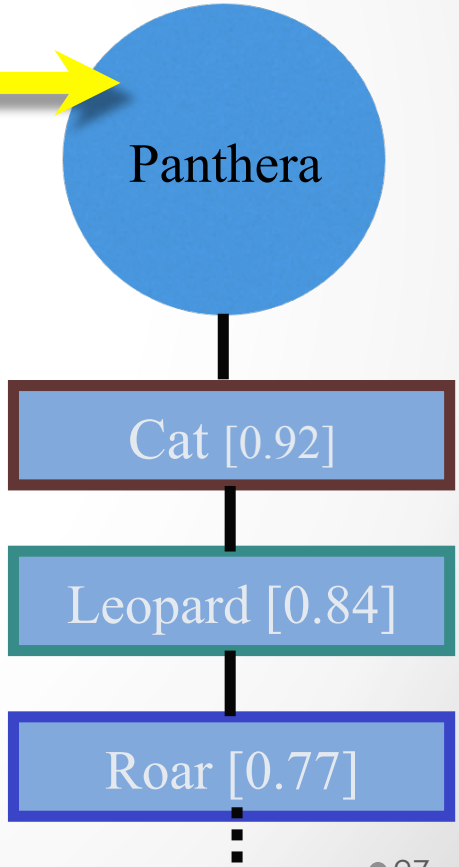
Panthera

Tiger

Scientific classification

Kingdom: Animalia

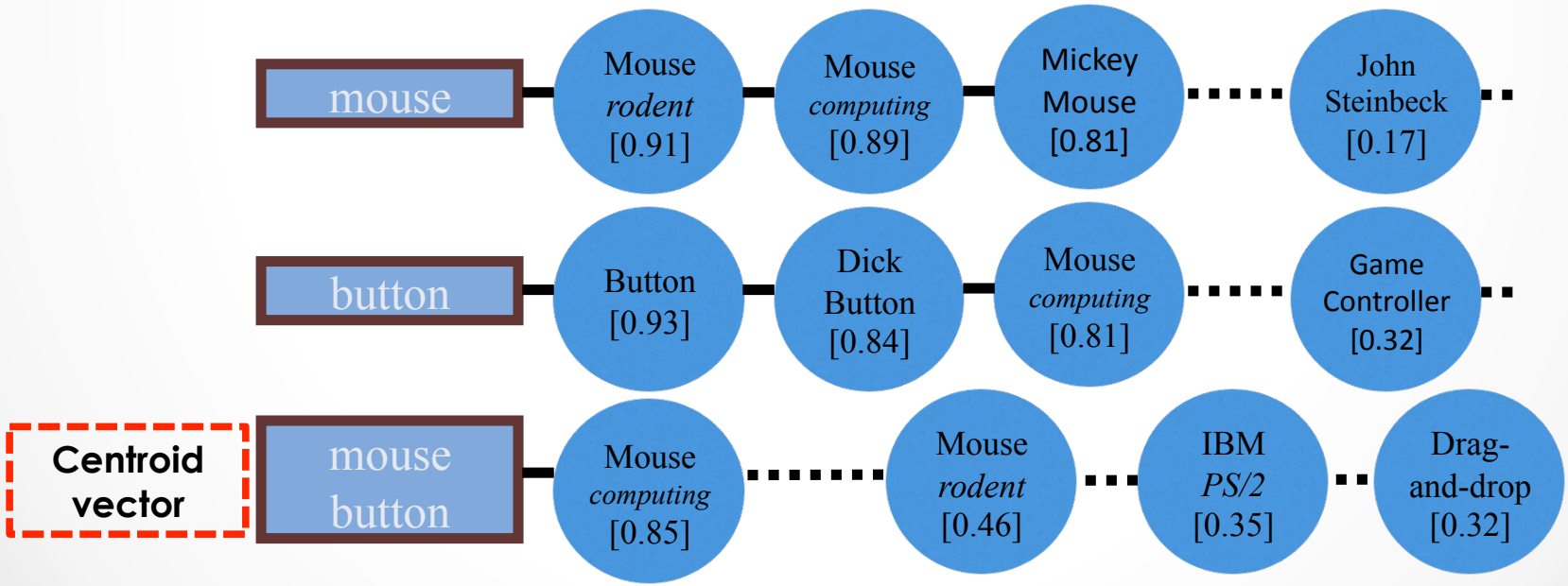
Phylum: Chordata



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

ESA: Text Fragment Interpretation

The **sematic 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 servicei in the ranked list, and N is the number of the services correctly retrieved **with the configuration**.

- Motivation
- CroSeR
- Experimental Evaluation
- Conclusions

Catalog	a@1	a@3	a@5	a@10	a@20	a@30
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