# Keyword/Term Extraction and Named Entity Resolution for PoliticalMashup
# USFD contribution to Talk of Europe 3

**Wim Peters and Adam Funk**
**University of Sheffield, UK**
**(w.peters|a.funk)@sheffield ac.uk**

During this event we enriched the English and Dutch parliamentary documents available from political mashup speeches (http://search.politicalmashup.nl/) with the following content metadata below.
Once integrated into the PoliticalMashup data base and interface functionality, this material will allow access to important conceptual elements within speeches, and will enable users to perform research on the data beyond keyword search.
The UK data are divided up into two sets, each representing a political party (Labour/Tory). This is useful for contrastive research into language use, topic/issue coverage and formulation.

The DbPedia links for English material will enable semantic normalization for named entities, and opens up conceptual expansion to additional information about the entities from this source.

In order to create the data we adjusted workflows from the Talk of Europe 2 event to accommodate the provenance and multilinguality (UK/NL) of the extracted material.
We also applied a new tool for English named entity resolution.

## Data

- **Terminology**: Dutch and English terms extracted from the texts by means of the TermRaider tool (http://www.dcs.shef.ac.uk/~wim/termraider.html).

This tool automatically provides domain-specific noun phrase term candidates from a text corpus on the basis of statistically derived termhood scores.
Possible terms are filtered by means of a multi-word-unit grammar that defines the possible sequences of part of speech tags constituting noun phrases.

- **Named Entities** (only for English)

Named entities from the English speeches were resolved against DbPedia by means of the YODIE tool (https://gate.ac.uk/applications/yodie.html).

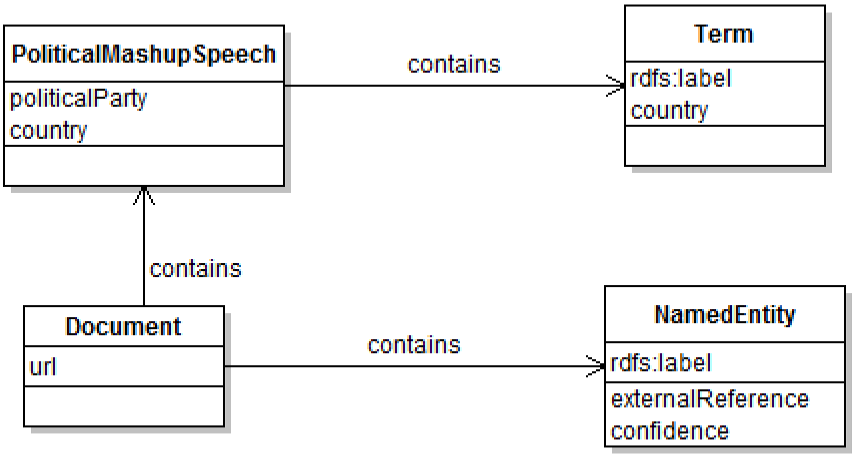
Reference:
Gorrell, Genevieve, Johann Petrak, and Kalina Bontcheva, Using @Twitter Conventions to Improve LOD-based Named Entity Disambiguation.
In: The Semantic Web. Latest Advances and New Domains, pp. 171-186. Springer International Publishing, 2015.

- **Document Selection**

We selected speeches within documents that match the pattern immigrat*/immigrant* (these patterns work for both EN and NL), and created the following subsets:

- UK Labour (4659 documents; 9316 speeches)
- UK Conservative speeches (4306 documents; 8610 speeches)
- NL parliamentary speeches without party distinction (9993 documents; 9998 speeches). Party information is available through integration of this data set with PoliticalMashup.

The data comes in RDF format and conforms to the data model on http://www.gate.ac.uk/ns/ontologies/toe-data-model.owl.



The RDF data is available as the five files below:

https://gate.ac.uk/ns/ontologies/TOE3-NL.rdf.gz (NL documents, speeches, terms)
https://gate.ac.uk/ns/ontologies/TOE3-UK-Conservative.rdf.gz (UK documents, Conservative speeches, terms)
https://gate.ac.uk/ns/ontologies/TOE3-UK-Labour.rdf.gz (UK documents, Labour speeches, terms)
https://gate.ac.uk/ns/ontologies/TOE3-UK-NE-Conservative.rdf.gz (UK documents, Conservative named entities)
https://gate.ac.uk/ns/ontologies/TOE3-UK-NE-Labour.rdf.gz (UK documents, Labour named entities)

**Statistics**:

| | |
|---|---|
| PoliticalMashup Document NL | 9988 |
| PoliticalMashup Document UK Labour | 4659 |
| PoliticalMashup Document UK Conservative | 4306 |
| Speech UK Labour | 4658 |
| Speech UK Conservative | 4305 |
| Speech NL | 9993 |
| Terms UK Labour | 7471 |
| Terms UK Conservative | 6684 |
| Terms NL | 19481 |
| Named Entities UK | 75533 |
| Triples | 3111574 |

Regretfully we had to stop short at providing the same information for the English and Dutch TalkOfEurope speeches because these were at the time unavailable through PoliticalMashup.

The data can be queried through Sparql.
Example 1 below selects terms and their labels from UK parliamentary Labour speeches.

```
PREFIX toedat:<http://www.gate.ac.uk/ns/ontologies/toe-data-model.owl#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

select distinct ?term ?label
where {
?speechuk rdf:type toedat:PoliticalMashupSpeech .
?speechuk toedat:contains ?term .
?speechuk toedat:politicalParty ?p .
?term rdf:type toedat:Term .
?term rdfs:label ?label .
filter(str(?p)="Labour")
}
```

Example 2 selects the named entity labels that have a DbPedia link with a confidence > 0.99

```
PREFIX toedat:<http://www.gate.ac.uk/ns/ontologies/toe-data-model.owl#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>

select ?l ?u ?v
where {
?s rdf:type toedat:NamedEntity .
?s toedat:externalReference ?u .
?s toedat:confidence ?v .
?s rdfs:label ?l .
filter(?v>0.99)
}
```