

DutchParl 1.0

A corpus of parliamentary documents in Dutch

Anne Schuth and Maarten Marx



Objective

- The aim of DutchParl is to create a corpus containing all digitally available parliamentary documents written in the Dutch language.

Coverage

- **Spatial:** Belgium, Flanders, The Netherlands (EU, Suriname forthcoming)
- **Temporal:** B: from 1999, V: from 1971, NL: from 1917.

Rich Metadata

- Metadata is described in a uniform way for all sub-collections using the 15 Dublin Core properties.

```
<meta>
  <dc:contributor>http://www.politicalmashup.nl</dc:contributor>
  <dc:coverage>
    <country dcterms:ISO3166="NL">Netherlands</country>
  </dc:coverage>
  <dc:creator>http://www.politicalmashup.nl</dc:creator>
  <dc:date>2002-09-05</dc:date>
  <dc:description>
    Handelingen 2001-2002, nr. 95, Tweede Kamer, pag. 5609-5621
  </dc:description>
  <dc:format>text/xml</dc:format>
  <dc:identifier>http://polidocs.nl/text/HAN7442A02.pm.xml</dc:identifier>
  <dc:language>nl</dc:language>
  <dc:publisher>http://parlando.sdu.nl</dc:publisher>
  <dc:relation>
  </dc:relation>
  <pm:dossiers>
    <pm:entity>28072</pm:entity>
  </pm:dossiers>
  <pm:person/>
  </dc:relation>
  <dc:rights>Tweede Kamer der Staten-Generaal</dc:rights>
  <dc:source>
    <pm:textsource>http://cdn.ikregoor.nl/pdf/HAN7442A02.pdf</pm:textsource>
    <pm:metasource>http://polidocs.nl/meta/HAN7442A02.meta.xml</pm:metasource>
  </dc:source>
  <dc:subject>
  </pm:keywords>
    <pm:entity>Gerechtelijk vooronderzoek</pm:entity>
    <pm:entity>Opsporingen</pm:entity>
  </pm:keywords>
  </pm:categories>
  </pm:entity>
    Strafrecht en strafprocesrecht (Gerechtelijk vooronderzoek)
  </pm:entity>
  </pm:entity>
    Strafrecht en strafprocesrecht (Opsporingsonderzoek)
  </pm:entity>
  </pm:categories>
  <dc:abstract/>
  </dc:subject>
  <dc:title>
    Behandeling van het wetsvoorstel Wijziging van de regeling van het DNA-onderzoek in
    strafzaken in verband met het vaststellen van uiterlijk waarneembare persoonskenmerken uit
    celmateriaal (28072)
  </dc:title>
  <dc:type>Verbatim Proceedings</dc:type>
</meta>
```

Size

Subcorpus	Mbyte text	# Documents	# Pages	# Tokens
Belgian Federal	800	3.901	216.522	129.085.483
Flanders	454	5.470	161.881	72.958.408
Netherlands	4.331	198.433	1.594.845	684.932.669
Flanders OCR	146	1.018	34.867	23.924.567
Netherlands OCR	7.043	328.722	1.701.130	1.003.555.596

Token counts

	NL DIGITAL	NL SCAN	Flanders DIGITAL	Flanders SCAN	BE federal
Total number of words	102870201	329540359	38629223	17120704	41152224
Unique words	353677	1963712	258304	184945	245447
Words occurring just once	149719	1311243	118992	91889	102093
Words occurring more than once	203958	652469	139312	93056	143354
Words occurring at least 4 times	130008	370932	88518	57277	90911
Words occurring at least 20 times	55054	134735	36413	22945	37250

Future Work

- Expand with all EU parliaments
- Search Engine



PoliticalMashup