

# WEB CLEF 2005

Erfahrungen bei der Bearbeitung des ersten mehrsprachigen Web-Korpus EuroGOV

## Gliederung

1. WebCLEF 2005
2. EuroGOV Korpus
3. Topicgenerierung
4. WebCLEFSearch Prozess (MIMOR)
5. Ergebnisse
6. Ausblick



Niels Jensen  
4. HIER Workshop 2005

## WebCLEF 2005

WebCLEF is about evaluating cross-language retrieval systems in a web setting. To facilitate these evaluation efforts, a corpus consisting of a crawl of governmental sites in Europe has been built, and WebCLEF 2005 participants have developed 575 known-item topics.

<http://ips.science.uva.nl/WebCLEF/index.html> verifiziert am 19.07.2005



Niels Jensen  
4. HIER Workshop 2005

## EuroGOV Korpus 1/3

- Europäischen Regierungsseiten
  - Regierungsportale
  - Websites aller Ministerien
- 27 Domains sind vertreten
- 13 Main Domains mit 131 Ministerien

EuroGOV Collection Domains			
Main domains		Additional domains	
Domain	Country	Domain	Country
.cz	Czech Republic	.at	Austria
.de	Germany	.be	Belgium
.es	Spain	.cy	Cyprus
.eu.int	European Union	.dk	Denmark
.fi	Finland	.ee	Estonia
.fr	France	.gr	Greece
.hu	Hungary	.ie	Ireland
.it	Italy	.lt	Lithuania
.nl	The Netherlands	.lu	Luxembourg
.pt	Portugal	.lv	Latvia
.ru	Russia	.mt	Malta
.se	Sweden	.pl	Poland
.uk	United Kingdom	.si	Slovenia
		.sk	Slovakia

(De Rijke & Mänttä, 2008:10)



Niels Jensen  
4. HIER Workshop 2005

## EuroGOV Korpus 2/3

- 3,6 Mio Webseiten
- 11GB komprimiert
- 157 Dateien zu je 25.000 Seiten
- Den höchsten Anteil haben
  - Finnland 660.000 Seiten
  - Germany 450.000 Seiten
  - EU 375.000 Seiten
  - Hungary 330.000 Seiten
  - Czech Republic 320.000 Seiten

EuroGOV Collection	
Language	Percentage
finnish	20.28%
german	18.20%
english	10.16%
latvian	8.80%
french	6.98%
swedish	5.32%
portuguese	3.85%
dutch	3.01%
polish	2.14%
italian	1.70%
spanish	1.30%
czech-lao850_2	1.13%
slovak-windova1250	0.80%
russian-windova1251	0.60%
danish	0.46%
estonian	0.39%
russian-ko8r_2	0.30%
slovak-ko8r	0.27%
greek-lao850_7	0.27%
lithuanian	0.19%
irish	0.05%
welsh	0.01%

(De Rijke & Mänttä, 2008:11)



Niels Jensen  
4. HIER Workshop 2005

## EuroGOV Korpus 3/3

```
<EuroGOV:bin domain="se" id="001">
<EuroGOV:doc
url="http://www.regeringen.se/"
id="Ese-001-35"
md5="659b462005b40f04bde5946b2beaad71"
fetchDate="Wed Sep 22 10:57:39 WEST 2004"
contentType="text/html">
<EuroGOV:content>
<![CDATA[
<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN">
<html lang="sv">
<head>
<title>Regeringen och Regeringskansliet</title>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<meta http-equiv="Content-Script-Type" content="text/javascript">
<meta http-equiv="Content-Style-Type" content="text/css">
<script language="javascript" type="text/javascript" src="/js/popup.js"></script>
<script language="javascript" type="text/javascript" src="/js/validationTexts.sv">
<script language="javascript" type="text/javascript" src="/js/formFunctions.js">
<link rel="stylesheet" type="text/css" href="/css/deprecatedstyle.css">

```



Niels Jensen  
4. HIER Workshop 2005

## Topicgenerierung 1/2

- Topicgenerierung
  - German 15 named page topics
  - German 15 home page topics
- Schwierigkeiten bei der Generierung
  - Frames
  - Flash & Grafiken



Niels Jensen  
4. HIER Workshop 2005

## Topicgenerierung 2/2

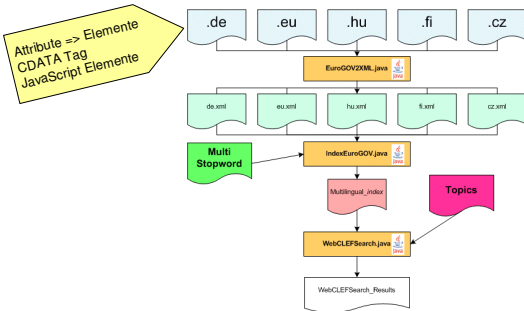
```

1 <topic>
2 <num>WC0008</num>
3 <title>LKW Maut Verordnung</title>
4 <metadata>
5 <topicprofile>
6 <language language="DE" />
7 <translation language="EN">German Truck Toll Regulation</translation>
8 </topicprofile>
9 <targetprofile>
10 <language language="DE" />
11 <domain domain="de"/>
12 </targetprofile>
13 <userprofile>
14 <native language="DE" />
15 <active language="EN" />
16 <passive language="NL" />
17 <passive language="ES" />
18 <countryofbirth country="DE" />
19 <countryofresidence country="DE" />
20 </userprofile>
21 </metadata>
22 </topic>
    
```



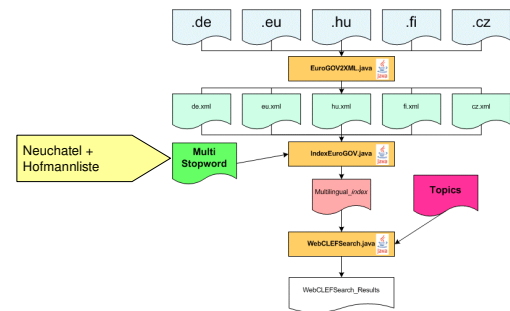
Niels Jensen  
4. HIER Workshop 2005

## WebCLEFSearch Prozess



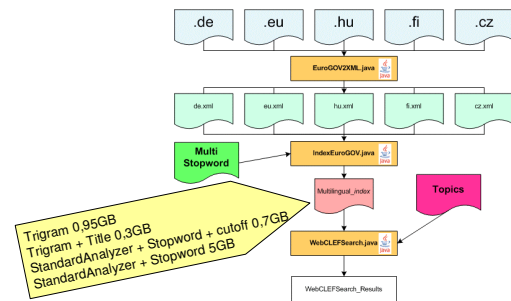
Niels Jensen  
4. HIER Workshop 2005

## WebCLEFSearch Prozess



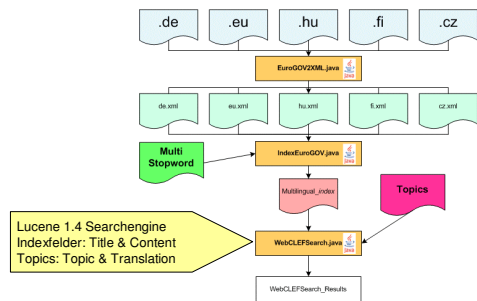
Niels Jensen  
4. HIER Workshop 2005

## WebCLEFSearch Prozess



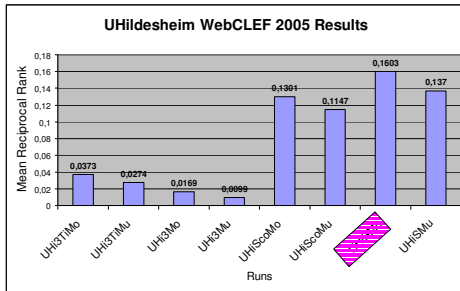
Niels Jensen  
4. HIER Workshop 2005

## WebCLEFSearch Prozess



Niels Jensen  
4. HIER Workshop 2005

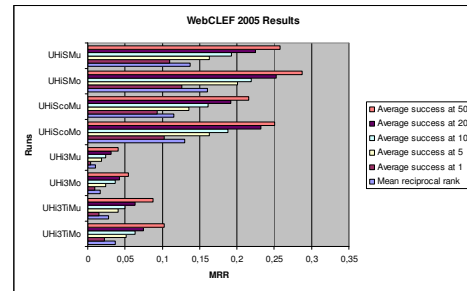
## Ergebnisse 1/2



WEB 2005

Niels Jensen  
4. HIER Workshop 2005

## Ergebnisse 2/2



WEB 2005

Niels Jensen  
4. HIER Workshop 2005

## Ausblick

- Hohes Verbesserungspotential bei der Indexierung
  - ein Index pro Sprache
  - Fehlerfreie XML Umformung
- Nutzen der Topicmetadaten

WEB 2005

Niels Jensen  
4. HIER Workshop 2005

## Topicmetadaten

```

1 <topic>
2 <num>W0008</num>
3 <title>...</title>
4 <metadata>
5 <topicprofile>
6 <language language="DE" />
7 <...>
8 </topicprofile>
9 <targetprofile>
10 <...>
11 </targetprofile>
12 </targetprofile>
13 <userprofile>
14 <...>
15 <...>
16 <...>
17 <...>
18 <countryofbirth country="DE" />
19 <countryofresidence country="DE" />
20 </userprofile>
21 </metadata>
22 </topic>
    
```

WEB 2005

Niels Jensen  
4. HIER Workshop 2005

Vielen Dank!

WEB 2005

Niels Jensen  
4. HIER Workshop 2005