

Analysis and Visualisation of Multilingual Document Collections

Brussels, 20 October 2000

Ralf Steinberger

European Commission – Joint Research Centre
Institute for Systems, Informatics and Safety (ISIS)
Risk Management and Decision Support Unit (RMDS)
Anti-fraud Information Management Sector (AIM)

T.P. 361

I - 21020 Ispra (VA)

Tel: +39 - 0332 786271

Ralf.Steinberger@jrc.it

- Who we are and what we do
- How Language Engineering can support the fight against fraud
- Brief general presentation of our Language Engineering work
- Focus on our efforts
 - to give multilingual and cross-lingual access to information contained in texts
 - to visualise the contents of multilingual text collections
- Work in progress !
 - Presentation of a mixture of working software, prototypes and plans
- Summary

- EC's JRC Ispra site in Italy (~ 1800 employees, see <http://www.jrc.it/>)
- JRC's Mission: give technological support to the EC
- AIM's main customer: OLAF
- *Language Technology* can help anti-fraud agencies and others
 - **to fight the 'information overflow'**
by providing tools which give quick access to information 'hidden' in large amounts of texts, written in a variety of languages
 - **to keep abreast of developments**
 - by monitoring the internet or intranets continuously,
 - by pointing out that some new relevant information is available,
 - by analysing this new information automatically and
 - by producing summarising reports automatically.

- A system which carries out
 - multilingual **retrieval** of potentially relevant documents (e.g. OSILIA project)
 - **extraction of different information aspects** from these documents;
language-neutral representation where possible
 - **visualisation** of contents of single documents and of document collections
- **Basic condition:** multilinguality
(applications should eventually be usable for all 11 official EU languages)
 - use statistical methods in order to be able to apply algorithm to other languages

Title E-3083/95 by Martin Schulz (PSE) **Seizure of plutonium at Munich airport**
Retrieval Date 03.05.1999
Creation Date 27.03.1996
Language(s) English (97% probability)
Source http://cnnfn.com/digitaljam/wires/9906/13/plutonium_eu.html
Display Language English (En, Fr, De, Es, It, Pt, Da, Fi, He, NI, Sv)

Free Indexing Terms

TUI, Commission, Karlsruhe, seizure, OJ, plutonium, suitcase, German, material

Eurovoc Indexing Terms

import, Federal Republic of Germany, plutonium, illicit trade, fraud, EAEC Joint Research Centre, airport

Names

Organisations: Commission, European Institute for Transuranium Materials (TUI), Joint Research Centre, PSE

People: Martin Schulz, Mrs. Breyer, Mr. Papoutsis

Geographical Profile

Relevance: 70%

Germany: 100%
Germany, German, München, Karlsruhe

Others: 0%

Combined Nomenclature Product Groups

CN 2844: "radioactive chemical elements and radioactive isotopes, incl. their fissile or fertile chemical elements and isotopes, and their compounds; mixtures and residues containing these products" (**plutonium**, 3)

CN 4204: "Trunks, **suit**, vanity, executive, brief, spectacle, binocular, camera, musical instrument, gun **cases**, holsters and similar; travelling, toilet bags, rucksacks, handbags, school satchels, shopping-bags, wallets, purses, map, cigarette cases" (**suitcase**, 3)

Document Summary

E-3083/95 by Martin Schulz (PSE)

Seizure of plutonium at Munich airport

In the summer of 1994 a suitcase containing plutonium illegally imported into Germany was seized in sensational circumstances at Munich airport in the Federal Republic of Germany. The Commission (Euratom safeguards directorate) was alerted by the German authorities in the early afternoon of 10 August, 1994, that some material might be seized.

<u>KEYWORD</u>	<u>KEYNESS</u>
TUI	65.31
Commission	62.27
Karlsruhe	57.55
seizure	55.84
OJ	42.21
plutonium	39.78
suitcase	38.44
German	29.49
material	28.51
Munich	23.60
Breyer	22.52
airport	17.80

<u>KEYWORD</u>	<u>KEYNESS</u>
PSE	17.06
Schulz	16.46
Euratom	15.99
Joint	14.11
Germany	12.83
authority	11.79
directorate	11.78
answer	11.58
question	11.56
safeguards	11.05
sensational	11.04
alert	10.98

- Restrictions of the presented method:
 - no compounds (e.g. *'power plant'*)
 - monolingual !
 - dependent on wording in text ('bread' vs. 'toast' vs. 'bakery products')
- Human indexers often use 'controlled vocabulary' such as *Eurovoc* for consistency and multilinguality
- We assign such keywords automatically after training our system on manually indexed documents (EP's *Eurovoc*).

First experiments have yielded rather good results.

- Developed by the European Parliament, for usage by the EP
- Controlled Vocabulary
- Multilingual (exists in all 11 official EU languages) !
- Hierarchically organised
 - 21 fields
 - 127 micro-thesauri
 - 5933 descriptors
 - 5877 reciprocal relations (BT, NT), 2730 reciprocal associations (RT)
- We have access to large amounts of training material (manually indexed texts)

SCORE	Top 40 DESCRIPTORS
92	community programme
84	young person
80	<u>cultural policy</u>
79	ceec
77	european union
76	continuing education
68	integration into employment
66	<u>rights of minorities</u>
65	<u>minority language</u>
65	cultural identity
64	education policy
64	vocational training
64	education
63	cultural heritage
63	new technology
61	<u>regional culture</u>
61	dissemination of culture
59	socrates
55	multilingualism
55	community action
54	european citizenship
54	efta countries

SCORE	Top 40 DESCRIPTORS
54	annual report
54	action programme
53	accession to the community
52	information network
52	cultural cooperation
52	translation
51	student mobility
51	<u>linguistic group</u>
51	cultural pluralism
51	community policy
50	information technology
49	language teaching
48	human rights
48	community financial instrument
47	cyprus
47	leonardo
47	telecommunications
47	regional language
...	...
BLUE:	manually assigned descriptors
GREEN:	further ‘reasonable’ descriptors
RED:	obviously wrong

Document Clustering

Measuring Document Similarity

document name	node+attraction	node#	docs	word#1	word#2	word#3	...
agricultural_policy_h.....\		7	1	consumer	restoration	encephalopathy	...
	53.\	333	2	consumer	labelling	spongiform	...
consumer_movement_h...../		69	1	consumer	labelling	transparency	...
	43.\	379	3	consumer	spongiform	encephalopathy	...
investment_aid_h...../		166	1	processing	encephalopathy	spongiform	...
	29-\	449	4	consumer	encephalopathy	bovine	...
community_control_h...../		43	1	monitoring	ban	bovine	...
	22----\	471	7	bovine	bse	consumer	...
goat_h.....\		143	1	scrapie	infect	scientific	...
	62.\	304	2	scientific	scrapie	veterinary	...
press_h...../		196	1	scientific	bovine	veterinary	...
	42..../	387	3	scrapie	scientific	infect	...
cosmetic_product_h...../		74	1	scrapie	encephalopathy	infect	...

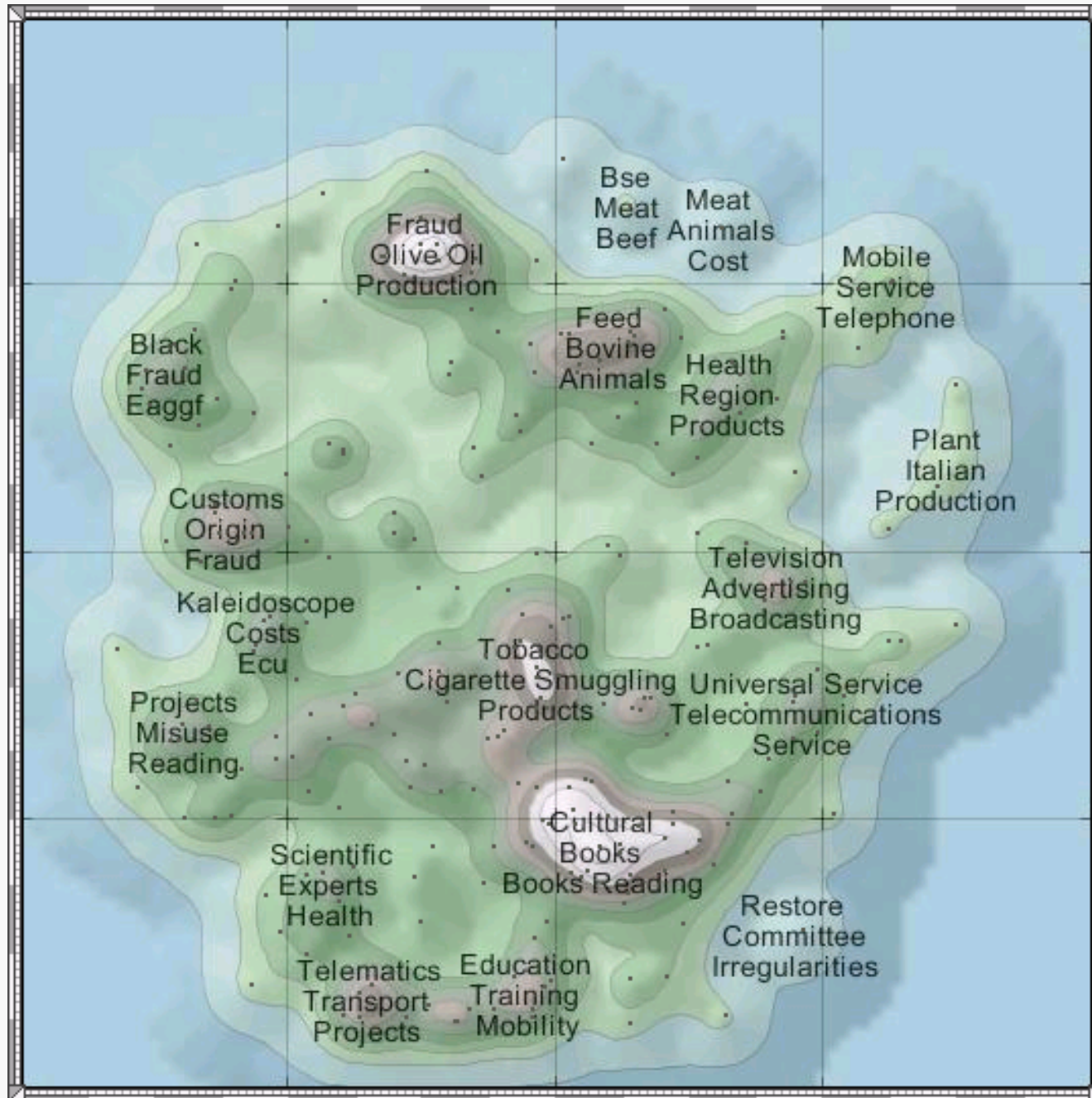
Small sample cluster of seven documents and the first three of a ranked list of indexing words for each document.

The system also calculates the most representative indexing words for each document cluster.

Clustering of multilingual document collections by using language-independent Eurovoc descriptors as input

- Document clustering according to similarity
 - Documents sharing words / keywords are similar
 - Eurovoc indexing allows cross-language comparison
 - Possibility to show a ranked list of all related documents (in different languages!)
- Visualisation of document collections
 - Two-dimensional representation of multi-dimensional document space
 - Kohonen maps (neural networks, JRC implementation)
 - JRC document charts
 - document maps using *Cartia* product **ThemeScape**
(see <http://www.cartia.com> and
<http://demo.cartia.com/jrcdescriptors/map1024.html> for our own data)

Document Map (ThemeScape)



Document Map – 2

(Search word 'olive' + documents in area)

ThemeScape Map Viewer: JRC Full Text - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://demo.cartia.com/JRCfulltext/map1024.html> Go

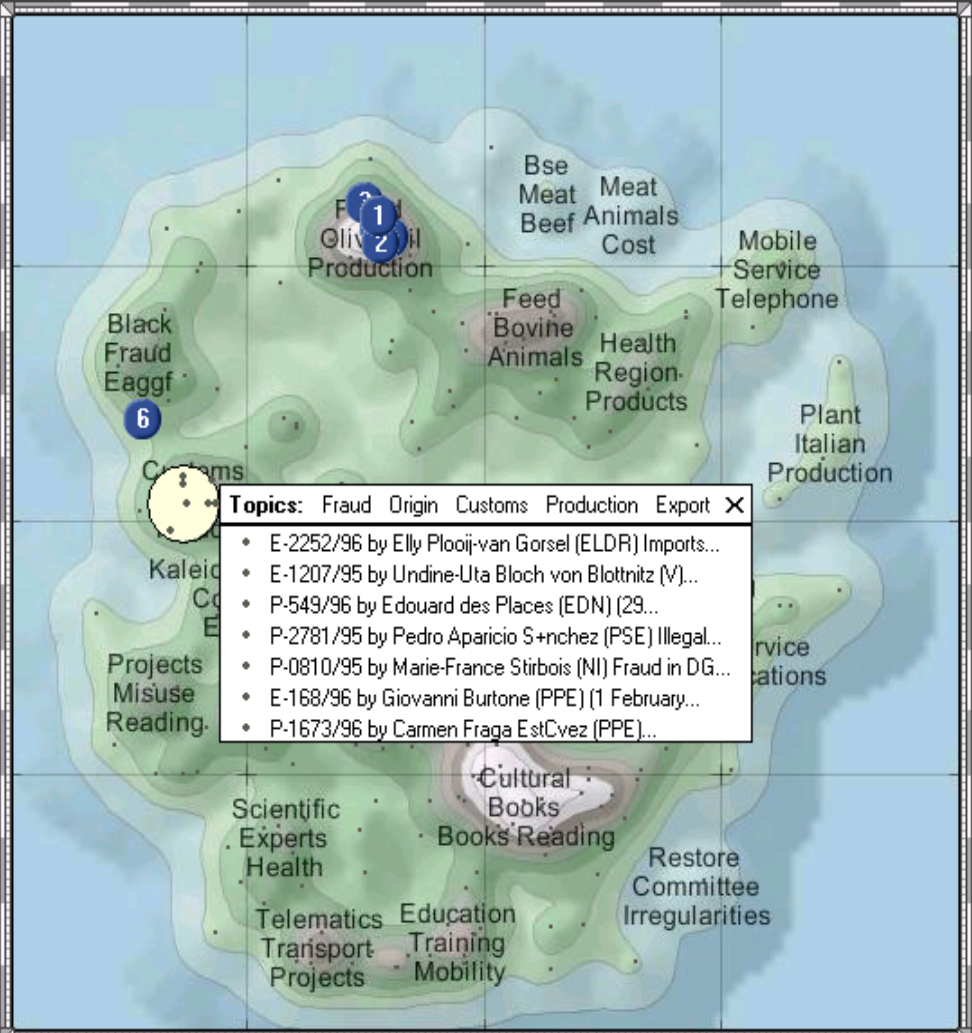
Map: JRC Full Text
Size: 258 documents

Map Legend Search Topic List Flags

6 of 6 results for the search "olive" were returned.

1	P-0633/97 by Joan Colom i Naval (PSE) Fraud in
2	E-3387/96 by Salvador JovC Peres (GUE/NGL) Operation of the
3	E-2908/96 by Alexandros Alavanos (GUE/NGL) Fraud involving Community subsidies
4	E-2963/96 by JesPoundss CabezCentn Alonso (PSE) and Juan Colino
5	E-2913/96 by Salvador Garriga Polledo (PPE) COM in olive
6	E-0190/96 by Isectionigo MCndez de Vigo (PPE) Commission investigations

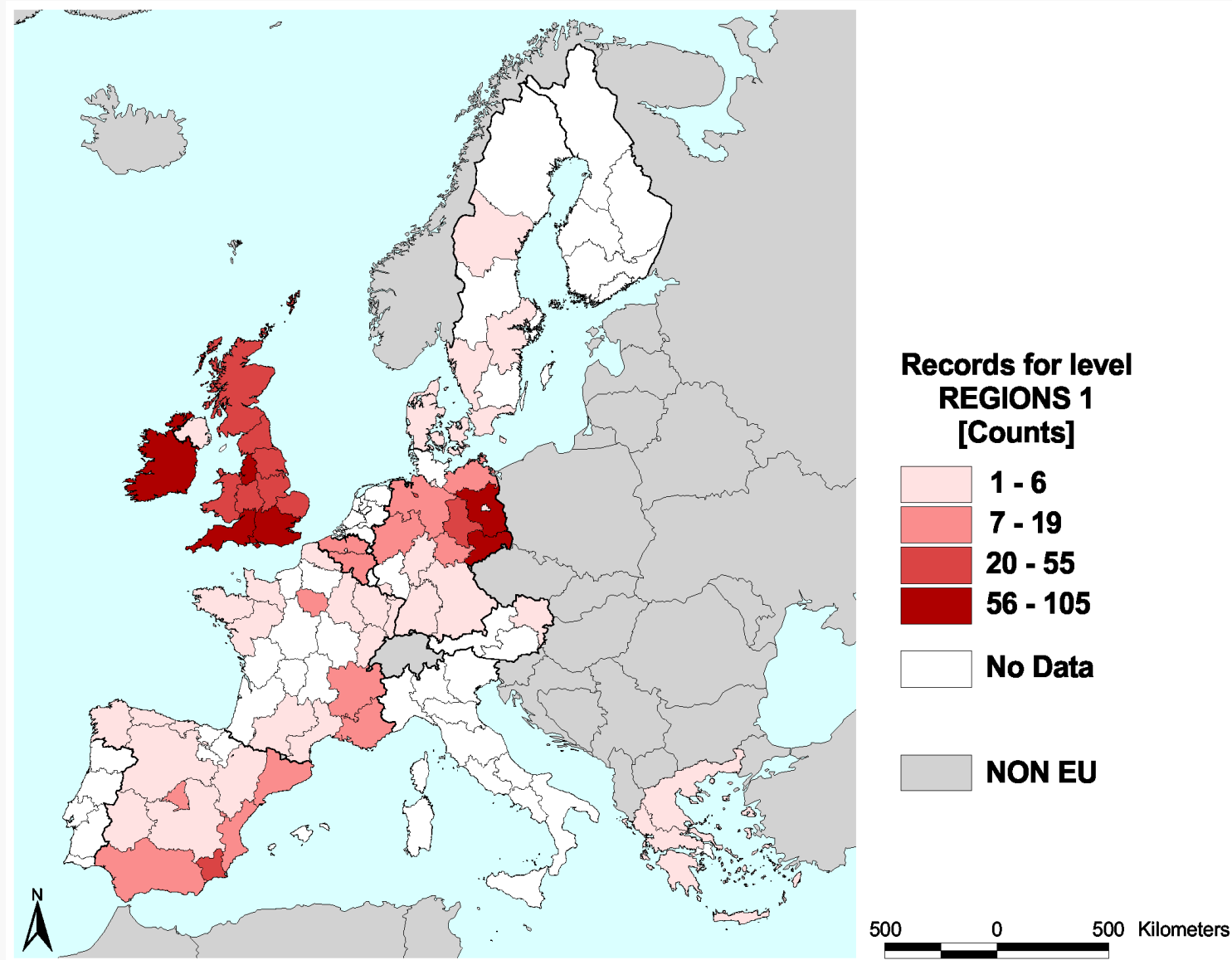
New Search 1-6 of 6 Back Next



Topics: Fraud Origin Customs Production Export X

- E-2252/96 by Elly Plooij-van Gorsel (ELDR) Imports...
- E-1207/95 by Undine-Uta Bloch von Blottnitz (V)...
- P-549/96 by Edouard des Places (EDN) (29...
- P-2781/95 by Pedro Aparicio S+nchez (PSE) Illegal...
- P-0810/95 by Marie-France Stirbois (NI) Fraud in DG...
- E-168/96 by Giovanni Burtone (PPE) (1 February...
- P-1673/96 by Carmen Fraga EstCvez (PPE)...

Visualisation of Geographical References in Text (Regions)



- **Goal:** System for multilingual document retrieval, information extraction and information visualisation
- **Means:** collection of independent tools carrying out different tasks
- **Multilinguality** achieved by linking documents to language-independent representations (Eurovoc thesaurus, Customs Tariff code TARIC)
- **Aim:** giving access to information in *multilingual* document collections
 - document profile
 - ranked list of similar documents even if written in different languages
 - multilingual document maps
 - representation of geographical references made in texts
- **Warning:** All linguistic applications are approximative and allow errors