

## Using Thesauri for Automatic Indexing and for the Visualisation of Multilingual Document Collections

**Sozopol, Bulgaria, 8 September 2000**

**Ralf Steinberger, Johan Hagman, Stefan Scheer**

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)**

- Who we are and what we do
- Automatic indexing (keyword assignment)
  - monolingual approach
  - controlled vocabulary multilingual indexing
  - how we do it
  - limitations of our approach
- Usage for multilingual document clustering and visualisation
  - Document Profiles (one document)
  - Document Maps (many documents)
- Summary and Outlook

- EC's JRC Ispra site in Italy (~ 1800 employees)
- Mission: give technological support to the EC
- LE activities of the AIM Sector:
  - multilingual **retrieval** of potentially relevant documents
  - **extraction of different information aspects** from these documents; language-neutral representation where possible
  - **visualisation** of contents of single documents and of document collections
- **Goal:** give cross-language access to information 'hidden' in large amounts of multilingual text (fight the '*information overflow*')
- **Basic condition:** multilinguality (applications eventually for 11 official EU languages)
  - use statistical methods in order to be able to apply algorithm to other languages

## 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**. Is The **Commission** aware of this matter and, if so, when were the **Commission** and its services, and other European agencies, informed of it? Can the **Commission** say whether the **Joint Research Centre** in **Karlsruhe** was involved, what services it provided for the **German** police, when it provided them, when the **plutonium** was seized, and when it was handed over to the **Joint Research Centre**?

2 -- Answer given by **Mr Papoutsis** on behalf of the **Commission** (10 January 1996)

The Commission would refer the Honourable Member to its earlier replies to questions about this incident (Written questions 1489/95[(1)] OJ C 213, 17.8. 1995] and 1508/95[(2) OJ C 230, 4.9.1995] by **Mrs Breyer**). 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**. In accordance with formal agreements between the **Commission** and the **German** government this information was immediately passed by phone to the **European institute for transuranium elements** (TUI) at **Karlsruhe** to ensure that preparations were made to receive any **material** seized. The **seizure** was made by the **German** police, and the TUI was not involved. Its activities that night were limited to receiving the closed **suitcase** at its premises in **Karlsruhe**. Subsequently, the TUI performed a precise analysis of the material found inside the **suitcase**, to support the investigations carried out by Member State authorities and to determine as far as possible the source and history of the nuclear **material**.

**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](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

- Problem of the presented method:
  - no compounds (e.g. *'power plant'*)
  - monolinguality
  - inconsistency ('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.

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)

- Method:
  - Create lists of associated general language lemmas for each descriptor ('associates')
  - Check new texts whether any of these associated terms exist
  - Count how many pointers there are for each descriptor
  - Formula: add log of associate weights and divide by text length
  - Ranked list of descriptors for the new text

# Sample lists of 'Associates'

## 'Fishery\_Management' & 'Democracy'

fishery	2751.07
fish	1743.80
stock	1653.37
fishing	1191.11
conservation	826.47
management	731.24
vessel	720.05
flag	533.36
organization	525.05
agreement	493.99
migratory	424.20
subregional	422.25
catch	390.41
mediterranean	323.22
sea	320.55
highly	312.76
session	263.72
resource	258.71
arrangement	252.56
fly	250.37
fleet	214.19
gfcms	202.66
fisherman	198.93
regulation	181.7

...

human	1007.52
right	939.07
democracy	892.03
operation	450.15
democratic	408.99
ombudsman	359.25
freedom	270.69
fundamental	245.70
cuba	211.33
principle	192.35
russia	185.05
consolidate	184.68
political	182.20
cooperation	177.99
respect	174.74
country	144.50
situation	130.41
turkey	129.87
general	127.28
finance	110.42
headquarters	103.17
relation	100.35
election	98.75
subsidiarity	96.82

...

### Score Descriptor Associates and their weight

47	<b>nuclear safety</b>	research (2 * 4) + euratom (1 * 6) + reply (1 * 3) + commission (7 * 3) + source (1 * 4) + plutonium (3 * 6) + nuclear (1 * 8) + schulz (1 * 3) + question (2 * 4) + breyer (1 * 3) + safeguard (1 * 4) + material (4 * 6) + munich (2 * 4)
46	<b>radioactive waste</b>	euratom (1 * 4) + aware (1 * 3) + commission (7 * 4) + incident (1 * 3) + german (4 * 3) + plutonium (3 * 5) + nuclear (1 * 7) + question (2 * 5) + germany (2 * 3) + element (1 * 3) + material (4 * 4)
43	<b>plutonium</b>	euratom (1 * 4) + reply (1 * 4) + seizure (2 * 3) + commission (7 * 3) + plutonium (3 * 6) + nuclear (1 * 6) + schulz (1 * 3) + question (2 * 4) + element (1 * 3) + breyer (1 * 3) + material (4 * 5) + munich (2 * 3)
...	...	...

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

- **Not enough training material** for all descriptors (< 3000 En and De texts for each descriptor > 10KB)
- Very **different text sizes** (one-line titles vs. 20 page documents)
- **Bias of the training material towards EP** interests
  - e.g. many associates of the descriptor 'Mauritania' have to do with fishery
- However, this is an extreme case and the bias of the associates is the same for all languages
  - ***document comparison is unimpaired***

# Associates of 'Mauritania'

## Bias towards EP interests

mauritania	1424.89	cod-end	115.54
vessel	1035.01	tonnage	114.99
mauritanian	966.70	catch	102.19
fishing	619.02	by-catch	101.16
fish	405.61	mile	100.34
observer	376.97	pelagic	89.93
licence	367.47	tonnage/fees	89.00
shipowner	344.26	category	88.51
islamic	289.74	republic	87.84
master	267.03	exchange_of_letter	80.58
agreement	264.30	mesh	80.04
seaman	228.14	low-water	79.34
chapter	216.49	scientific	79.18
ministry	215.96	gear	77.26
board	197.78	copy	77.10
zone	167.78	maximum	76.29
licences	161.14	port	75.86
latitude	149.54	biological	75.70
log	146.97	cephalopod	75.15
datasheet	139.25	period	74.72
inspection	127.39	surveillance	73.24
fishery	124.15	sea	72.14
authorize	117.16	sheet	69.30
nautical	116.86	...	

- Document clustering according to similarity
  - Documents sharing descriptors are similar
  - Eurovoc indexing allows cross-language comparison
  - Possibility to show a ranked list of all related documents
- Visualisation of document collections
  - Two-dimensional representation of multi-dimensional document space
    - Kohonen maps (neural networks)
    - 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)

ThemeScape Map Viewer: JRC Full Text - Microsoft Internet Explorer

File Edit View Go Favorites Help

Back Forward Stop Refresh Home Search Favorites History Channels Fullscreen Mail Print Edit

Address <http://demo.cartia.com/jrcfulltext/map1024.html> Links

Map: JRC Full Text  
Size: 258 documents

Map Legend Search Topic List Flags

Search for the map topics you would like to display:

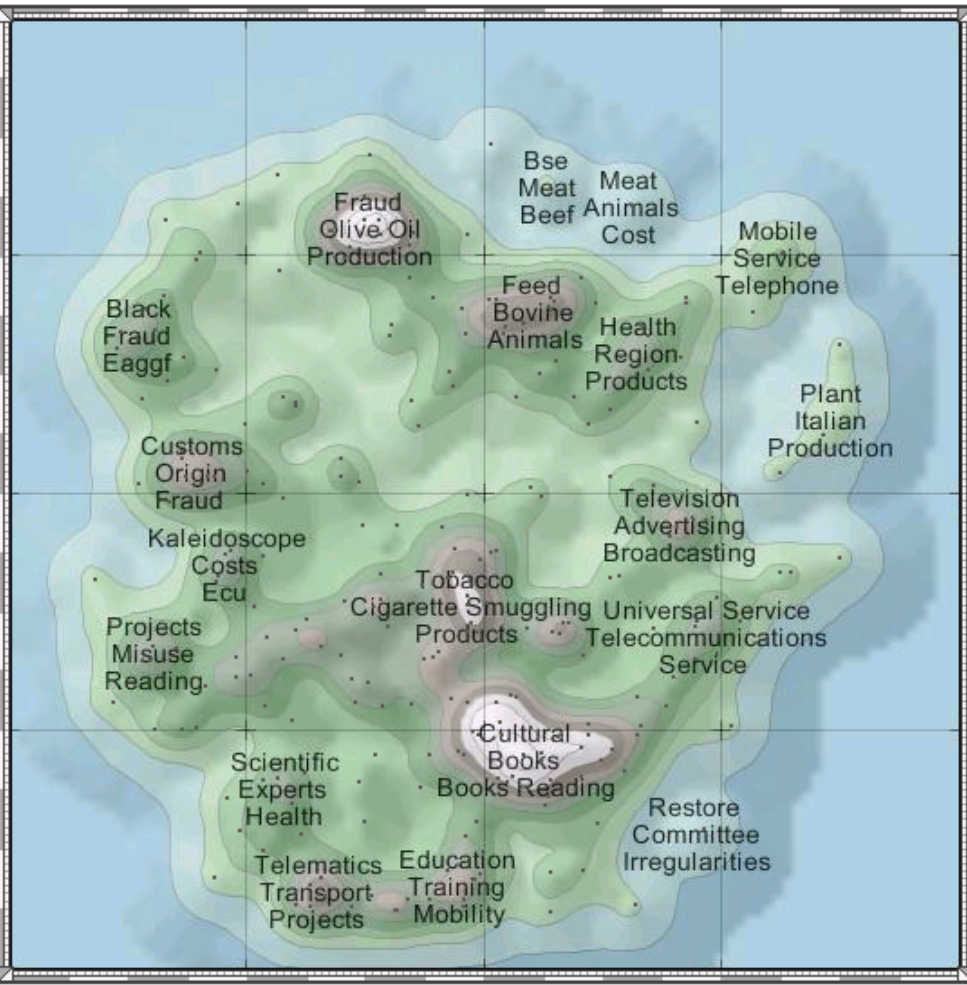
Topic	#Docs
Fraud	76
Projects	66
Production	53
Culture	47
Service	46
Regions	40
Irregularities	37
Health	36
Committee	35
Ecu	35
Customs	32
Cost	30
Export	24
Transport	23
Mecu	22
Expenditure	22
Transit	22

Clear Search

Search Options

Look for

Limit results to  documents



# 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

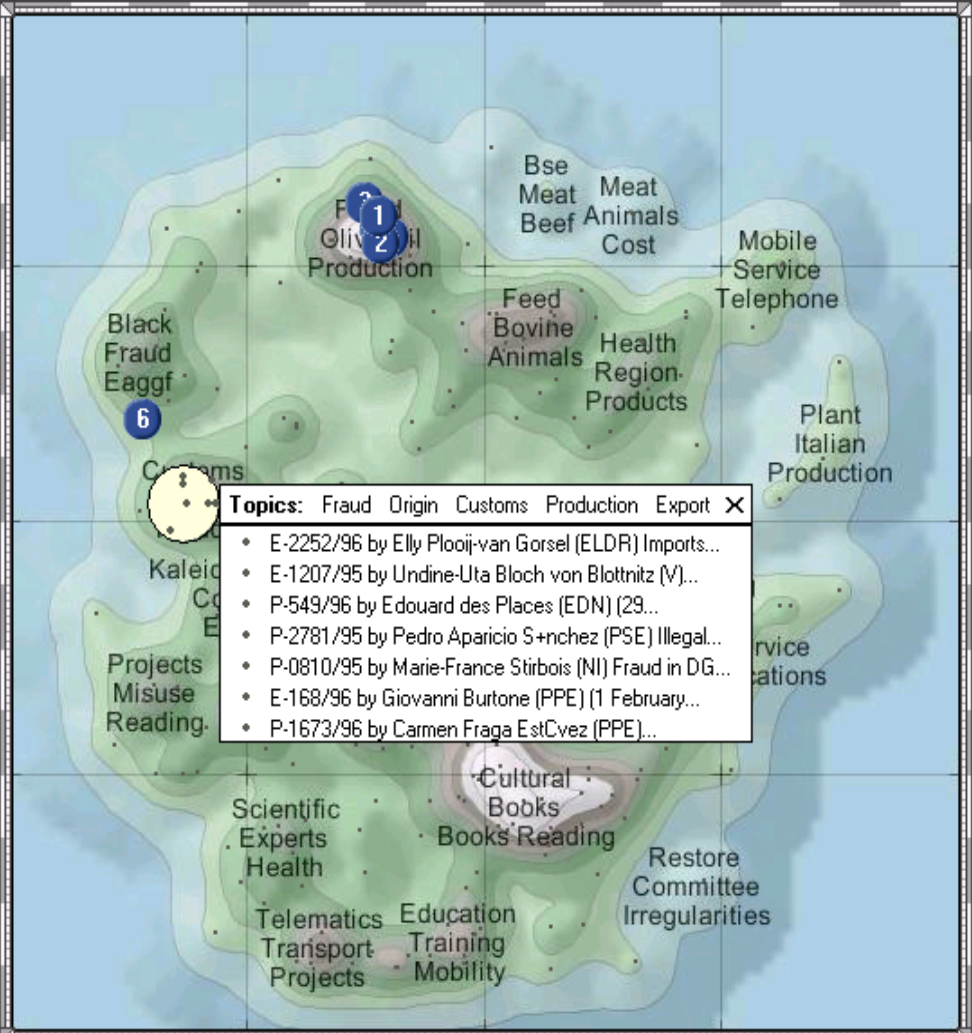
Map: JRC Full Text  
Size: 258 documents

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)...

Internet

- **Goal:** System for multilingual document retrieval, information extraction and information visualisation
- e.g. **Controlled vocabulary indexing** with Eurovoc descriptors
- **Means:** using training data to identify ‘associated lemmas’ which point to descriptors
- **Result:** ranked list of language-independent descriptors for each document
- **Aim:** visualisation of *multilingual* document collection
  - document profile
  - ranked list of similar documents in any language
  - multilingual document maps

# To Do

- Large scale **evaluation**
- Compare to manual indexing results
- Compare results for identical translations
- Use descriptor co-occurrence statistics to improve the results

# Pause for laughter