

Proposal for an ISIS Exploratory Research Project

Proposed by: Ralf Steinberger

Duration: 1 year for first phase

Requirements: 1 person year +
20 KEuro for specific credits

JRC - ISIS, 20 February 2001

Aim / Application

- Automatic cross-lingual indexing (keyword assignment)
- Eurovoc Thesaurus (<http://europa.eu.int/celex/eurovoc/>)
 - developed by the European Parliament (EP) in collaboration with OPOCE and many national governments, over many years
 - to index all their documents (manually) and to allow cross-lingual retrieval in the document database
 - It is a hierarchically organised, closed list of terms describing the main interests of the EP
 - It exists in exact (one-to-one) translations in all eleven official European Union languages !!!
- ➔ **Identifying Eurovoc terms for a text in one language allows viewing the keyword list in all eleven languages**
- Novelty:
 - Indexing software exists, but not cross-lingual and not in 11 languages
 - Producing a resource like Eurovoc would take many years

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://cnnfm.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.

Aim / Application

- Automatic cross-lingual indexing (keyword assignment)
- Eurovoc Thesaurus (<http://europa.eu.int/celex/eurovoc/>)
 - developed by the European Parliament (EP) in collaboration with OPOCE and many national governments, over many years
 - to index all their documents (manually) and to allow cross-lingual retrieval in the document database
 - It is a hierarchically organised, closed list of terms describing the main interests of the EP
 - It exists in exact (one-to-one) translations in all eleven official European Union languages !!!
- ➔ **Identifying Eurovoc terms for a text in one language allows viewing the keyword list in all eleven languages**
- Novelty:
 - Indexing software exists, but not cross-lingual and not in 11 languages
 - Producing a resource like Eurovoc would take many years

04 Politics
08 International Relations
10 European Communities
12 Law
16 Economics
20 Trade
24 Finance
28 Social Questions
32 Education and Competition
36 Science
40 Business and Competition
44 Employment and Working Conditions
48 Transport
52 Environment
56 Agriculture, Forestry and Fisheries
60 Agri-Foodstuffs
64 Production, Technology and Research
66 Energy
68 Industry
72 Geography
76 International Organisations

28 SOCIAL QUESTIONS

2806 family
2811 migration
2816 demography and population
2821 social framework
2826 social affairs

2831 culture and religion

arts

cultural policy

culture

acculturation

civilization

cultural difference

cultural identity

RT: protection of minorities (1236)

RT: socio-cultural group (2821)

cultural pluralism

popular culture

regional culture

religion

2836 social protection

2841 health

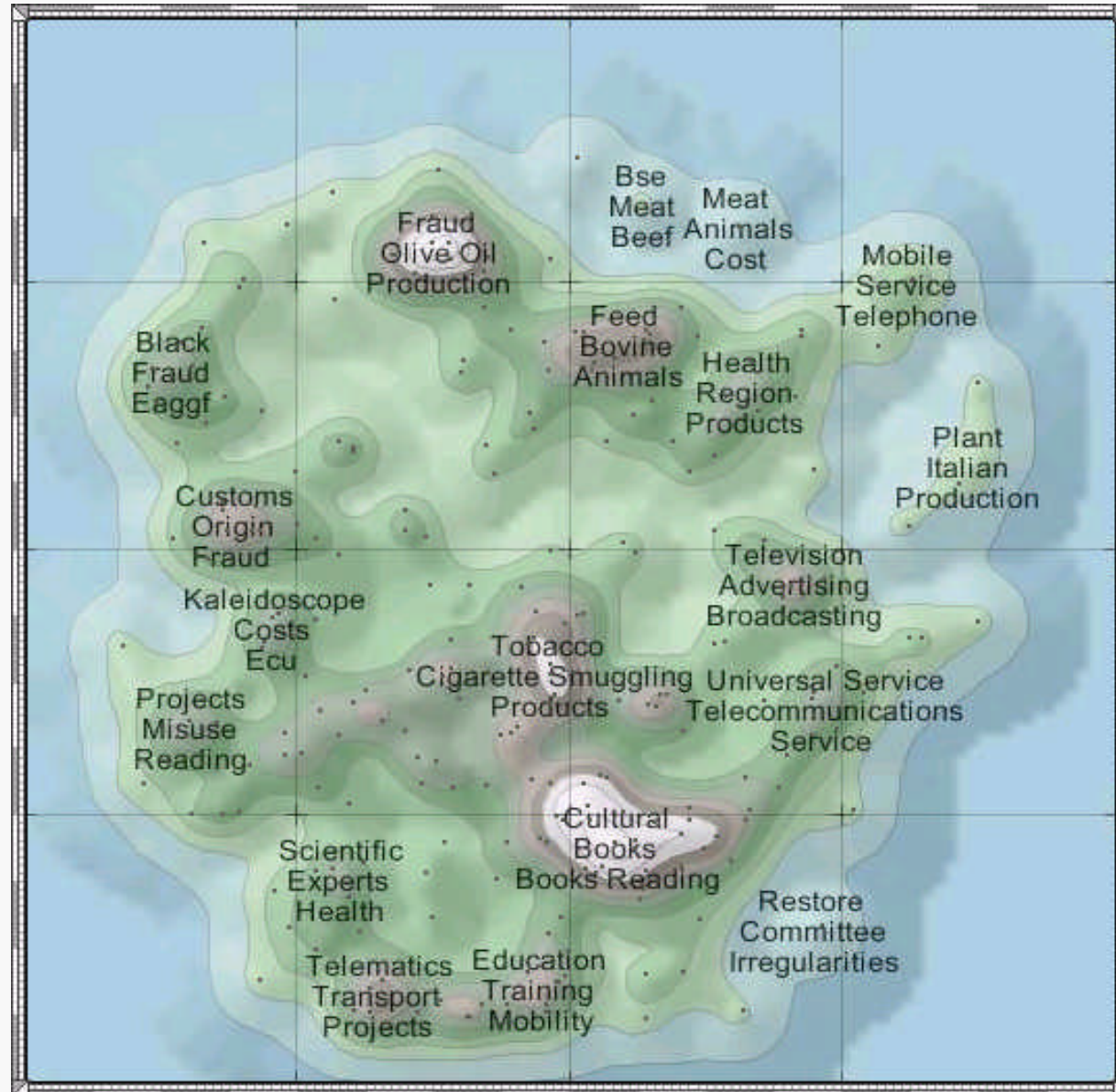
2846 construction and town planning

Aim / Application

- Automatic cross-lingual indexing (keyword assignment)
- Eurovoc Thesaurus (<http://europa.eu.int/celex/eurovoc/>)
 - developed by the European Parliament (EP) in collaboration with OPOCE and many national governments, over many years
 - to index all their documents (manually) and to allow cross-lingual retrieval in the document database
 - It is a hierarchically organised, closed list of terms describing the main interests of the EP
 - It exists in exact (one-to-one) translations in all eleven official European Union languages !!!
- ➔ **Identifying Eurovoc terms for a text in one language allows viewing the keyword list in all eleven languages**
- Novelty:
 - Indexing software exists, but not cross-lingual and not in 11 languages
 - Producing a resource like Eurovoc would take many years

- Allow cross-language access to document contents
- Allow cross-language comparison of texts (find similar documents even in different languages)
- Allow visualisation of multilingual document collections, e.g. using document maps
- Interested parties (customers)
 - **OLAF**
 - **Documentation Centres of the EP, OPOCE and the Swedish Parliament**
 - Nuclear Safeguards Unit (Knowledge Centre for non-proliferation project)
 - ...
- Partners
 - **Document Management Partners (Antwerp, Belgium; www.DMPartners.be)**
 - IBM in Stuttgart; LexiQuest in Paris
 - Cross-fertilisation with other ISIS sectors like *Web Technologies* and *Information Dependability*

Document Map (ThemeScape)



- Allow cross-language access to document contents
- Allow cross-language comparison of texts (find similar documents even in different languages)
- Allow visualisation of multilingual document collections, e.g. using document maps
- Interested parties (customers)
 - **OLAF**
 - **Documentation Centres of the EP, OPOCE and the Swedish Parliament**
 - Nuclear Safeguards Unit (*Knowledge Centre for non-proliferation* project)
 - ...
- Partners
 - **Document Management Partners (Antwerp, Belgium; www.DMPpartners.be)**
 - IBM in Stuttgart; LexiQuest in Paris
 - Cross-fertilisation with other ISIS sectors like *Web Technologies* and *Information Dependability*

- Machine-learning on the basis of manually indexed document collections (training material from EP and OPOCE)
 - Identify specific features of the texts indexed with a particular Eurovoc term (ranked word lists; associated words)
 - look for these 'associated words' in new texts and produce a ranked list of most likely Eurovoc terms, based on an assignment
- **Challenge / Difficulties** ® **Research**
 - **How to make use of the structure of Eurovoc? (hierarchy and 'related terms')**
 - **Optimisation of the keyword assignment algorithm**
 - Uneven usage of Eurovoc descriptor terms in training material
 - Uneven length of manually indexed texts

Sample lists of 'Associates'

'Fishery_Management' & 'Democracy'

fishery	2751.07	human	1007.52
fish	1743.80	right	939.07
stock	1653.37	democracy	892.03
fishing	1191.11	operation	450.15
conservation	826.47	democratic	408.99
management	731.24	ombudsman	359.25
vessel	720.05	freedom	270.69
flag	533.36	fundamental	245.70
organization	525.05	cuba	211.33
agreement	493.99	principle	192.35
migratory	424.20	russia	185.05
subregional	422.25	consolidate	184.68
catch	390.41	political	182.20
mediterranean	323.22	cooperation	177.99
sea	320.55	respect	174.74
highly	312.76	country	144.50
session	263.72	situation	130.41
resource	258.71	turkey	129.87
arrangement	252.56	general	127.28
fly	250.37	finance	110.42
fleet	214.19	headquarters	103.17
gfcms	202.66	relation	100.35
fisherman	198.93	election	98.75
regulation	181.7	subsidiarity	96.82
...		...	

- Machine-learning on the basis of manually indexed document collections (training material from EP and OPOCE)
 - Identify specific features of the texts indexed with a particular Eurovoc term (ranked word lists; associated words)
 - look for these 'associated words' in new texts and produce a ranked list of most likely Eurovoc terms
- **Challenge / Difficulties** ® **Research**
 - **How to make use of the structure of Eurovoc? (hierarchy and 'related terms')**
 - **Optimisation of the keyword assignment algorithm**
 - Uneven usage of Eurovoc descriptor terms in training material
 - Uneven length of manually indexed texts

Proposal for an ISIS Exploratory Research Project

Proposed by: Ralf Steinberger

Duration: 1 year for first phase

Requirements: 1 person year +
20 KEuro for specific credits

JRC - ISIS, 20 February 2001

Limitations of the Approach

- It is very difficult to extend the automatic Eurovoc keyword assignment to further languages (no training material)
- Bias of the training material towards EP interests
 - e.g. many associates of the descriptor 'Mauritania' have to do with fishery
- Areas covered are limited to those of the Eurovoc thesaurus e.g. 'fraud'

customs fraud	
electoral fraud	
elimination of fraud	(USE fraud)
European Anti-fraud Office	(USE OLAF)
fight against fraud	(USE fraud)
fraud	
fraud against the Community	
fraud prevention	(USE fraud)
fraudulent trade	(USE illicit trade)

Score Descriptor Associates and their weight

47	nuclear safety	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	radioactive waste	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	plutonium	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)
...

Automatically assigned indexing terms

“Resolution on linguistic and cultural minority in the European Community”

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