



Cross-lingual keyword assignment and other Language Technology activities at the JRC

Ralf Steinberger

European Commission – Joint Research Centre (JRC)

Institute for the Protection and Security of the Citizen (IPSC)
Cybersecurity and New Technologies for Combating Fraud
Anti-fraud Information Management Sector (AIM)
<http://www.jrc.it/langtech>



Agenda




- Introduction: Who we are and what we do
 - JRC
 - Sector: Anti-fraud Information Management (AIM)
 - Language Technology in AIM
- Monolingual keyword assignment
- Assignment of descriptors of the multilingual thesaurus Eurovoc to texts
- Remarks on clustering, document similarity and visualisation
- Summary




Joint Research Centre



- Directorate General of the European Commission (DG JRC)
- > 1500 scientists and technicians, ca. 2500 people
- 8 institutes
- Placed in Ispra (I), Karlsruhe (D), Sevilla (E), Petten (NL), Geel (B), Headquarters in Brussels
- Ispra: ca. 1800 people
- Scientific research and scientific services for DGs
- Wide range of subjects:
 - Nuclear safety
 - Environment (alternatives to animal testing, recognition of adulterated wine, (non-)biological food, ...)
 - Security of food and chemical products
 - Dependability of information systems
 - ...




JRC Mission




The mission of the JRC is to provide customer-driven scientific and technical support for the conception, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

To carry out this mission, the JRC has a unique combination of facilities and expertise transcending national boundaries. Moreover, through its networks it stimulates collaborative research and broadens its knowledge.



Institute for the Protection and Security of the Citizen

Anti-fraud Information Management Sector



JOINT RESEARCH CENTRE
EUROPEAN COMMISSION

- Part of:
 - Institute for the Protection and Security of the Citizen (IPSC)
 - Cybersecurity and New Technologies for Combating Fraud Unit
- 8-15 people
- We do not fight fraud directly, but
- Main client is OLAF (Organisation de la Lutte Anti-Fraude:
misuse of subsidies,
veiling the origin of meat,
adulterating butter or wine, etc.)
- Also other clients, European SCA projects, etc.
- Analyses for OLAF, technological overview, feasibility studies, EDMS,
work-flow management, software prototypes, etc.
- Multidisciplinary



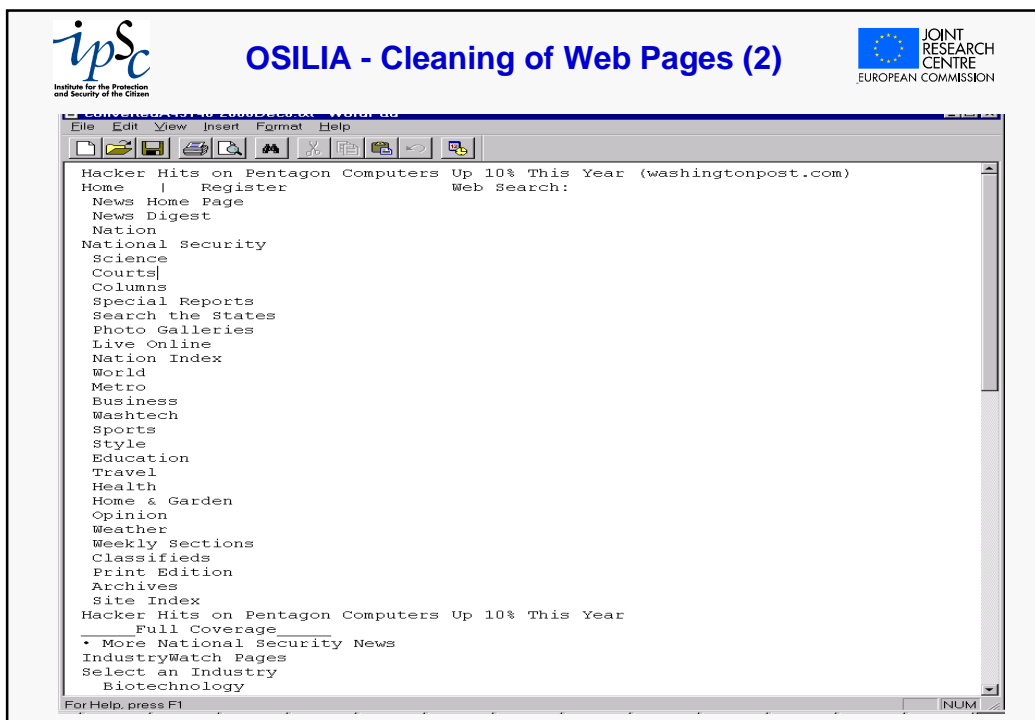
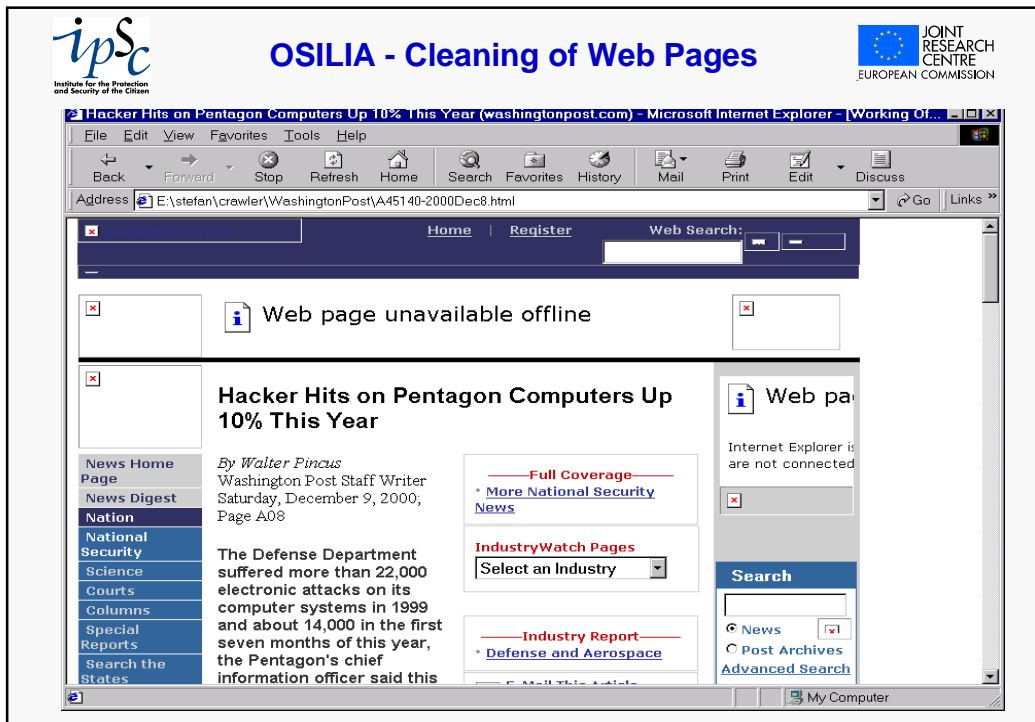
Institute for the Protection and Security of the Citizen


Goal of our Language Technology Work



JOINT RESEARCH CENTRE
EUROPEAN COMMISSION


- **Retrieval** of potentially relevant texts in a variety of languages (OSILIA project)
- **Extraction** of a variety of information aspects from these texts;
when possible: language-independent representation of the contents
 - **key words (monolingual and cross-lingual)**
 - language of texts
 - subject domains
 - references to geographical places
 - references to people, to products, etc.
 - summary
 - Calculation of the similarity of documents
 - clustering and classification of documents
- **Visualisation of the contents**
 - of individual documents in *document profiles*
 - of whole text collections in *document maps*





Institute for the Protection
and Security of the Citizen


OSILIA - Cleaning of Web Pages (3)



JOINT
RESEARCH
CENTRE
EUROPEAN COMMISSION


WashingtonPost-A45140-2000Dec09.txt - WordPad

Hacker Hits on Pentagon Computers Up 10% This Year (washingtonpost.com)
 Hacker Hits on Pentagon Computers Up 10% This Year
 IndustryWatch Pages
 Select an Industry
 • Defense and Aerospace
 By Walter Pincus
 Washington Post Staff Writer
 Saturday, December 9, 2000; Page A08
 The Defense Department suffered more than 22,000 electronic attacks on its computer systems in
 The vast majority of those attacks were either harmless or caused only petty harassment, but i
 Pentagon officials said that, to the best of their knowledge, the Department of Defense's clas
 The department was able to make an accurate count of the number of attacks for the first time
 In 1999, the Pentagon detected 22,144 attempts to probe, scan, hack into, infect with viruses
 So far this year, officials said, the number of attacks is up approximately 10 percent, and th
 In an interview, Money predicted that the number of attacks is only "going to increase."
 "A majority of the attacks [that cause damage] come through vulnerabilities in existing softwa
 Although the Pentagon is "putting more and more effort into testing" off-the-shelf software ar
 "On a lot of these [programs], we don't know where the code is written," he said.
 Many of the vulnerabilities are unintentional, but some appear to be "trapdoors" deliberately
 As a result, the official added, "we are not buying such off-the-shelf products in our most se
 The Pentagon's cybersecurity problem is enormous. The Defense Department has roughly 10,000 cc
 In August, Congress put an additional \$163 million for computer security into the fiscal 2001
 The "seminal event" that awakened the Pentagon to its computer security problems occurred in E
 Those attacks, which came as preparations were underway for a possible military operation agai
 Military computer administrators had been warned about the weakness that the California hacke
 © 2000 The Washington Post Company




Institute for the Protection
and Security of the Citizen

Goal of our Language Technology Work (2)




JOINT
RESEARCH
CENTRE
EUROPEAN COMMISSION

- **Goal:** give cross-language access to information 'hidden' in large multilingual document collections ("fight the information overflow", "overcome the language barrier")
- Purchase of existing commercial software
Development of applications not available commercially
- **Focus:** multilinguality and cross-lingual applications
- For political and practical reasons: all 11 official EU-languages
- Usage of mainly **statistical methods**
 - less labour-intensive
 - developed methods can easily be adopted to further languages
- Team: Johan Hagman (johan.hagman@jrc.it)
Bruno Pouliquen (bruno.pouliquen@jrc.it)
Ralf Steinberger (ralf.steinberger@jrc.it)



Sample Text




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.



Document Profile



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)

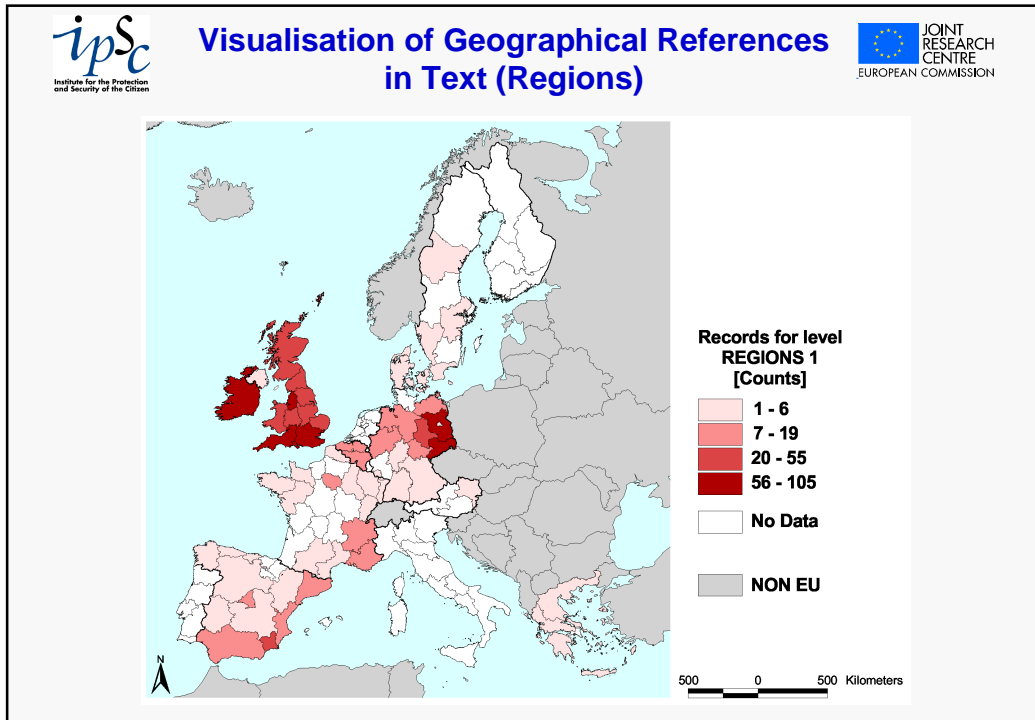
CN 4204: "Trunks, suit, vanity, executive, brief, spectacle, binocular, camera, musical instrument, gun cases, holsters and similar; travelling, toilet bags, rucksacks, handbags, school satchels, shopping-bags, wallets, purses, map, cigarette cases" (suitcase, 3)



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.




 **Monolingual Keyword Identification** 

- Statistical approach
 - Minimal linguistic input
 - lemmatisation (base form reduction of words)
 - extensive stop word lists
 - mark-up of multi-word terms (MWU)
 - Comparison of the text lemma (word) frequency list with the lemma frequency list of a reference corpus (e.g. BNC, several years of newspaper text)
 - Comparison of the frequency tables using the *log-likelihood* (or *chi-square*) tests
 - Result: a list of keywords and their *keyness* (weight; relevance for the document contents)
 - Easy to extend to other languages, only requires lemmatiser and reference corpus for this language




Monolingual Keyword Identification Example




Document: *Question to the European Parliament regarding a case of smuggled plutonium, seized at Munich airport. Plutonium was analysed by JRC-TUI in Karlsruhe*


KEYWORD	KEYNESS	KEYWORD	KEYNESS
TUI (3/5)	65.31	PSE	17.06
Commission (7/9484)	62.27	Schulz	16.46
Karlsruhe (3/22)	57.55	Euratom	15.99
seizure	55.84	Joint	14.11
OJ	42.21	Germany	12.83
plutonium	39.78	authority	11.79
suitcase	38.44	directorate	11.78
German	29.49	answer	11.58
material	28.51	question	11.56
Munich	23.60	safeguards	11.05
Breyer	22.52	sensational	11.04
airport	17.80	alert	10.98




Advantages and Limitations of this Method



- **Advantage:**
 - To extend to other languages, only lemmatiser (or stemmer) and reference corpus are needed
- **Limitations of this keyword identification procedure**
 - No compounds apart from the closed list of MWUs (z.B. 'power plant')
 - Monolinguality (multi-monolinguality)
 - Lack of abstraction and consistency ('bread' vs. 'toast' vs. 'bakery products')
- Professional organisations use people to assign controlled vocabulary keywords from a multilingual thesaurus (e.g. Eurovoc)
- ➔ We want to assign Eurovoc descriptors automatically



Eurovoc Thesaurus




- Developed by the European Parliament (EP) and the EC's Publications Office (OPOCE), together with several national organisations
- Controlled Vocabulary
- Multilingual (exists in all 11 official EU languages) !
- We have access to large amounts of training material (manually indexed texts)
- Hierarchically organised into max. 8 levels
 - 21 fields (*politics; law; economics; social questions; environment; industry; geography; energy; agri-foodstuffs; agriculture, forestry and fisheries; international organisations; etc.*)
 - 127 micro-thesauri
 - 5933 descriptors
 - 5877 reciprocal relations (BT, NT), 2730 reciprocal associations (RT)
- Challenge: Descriptor terms like 'DEMOGRAPHY AND POPULATION' or 'CONSTRUCTION AND TOWN PLANNING' are unlikely to occur as such in a text




Eurovoc (Top Level and Detail)



<ul style="list-style-type: none"> 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 	<ul style="list-style-type: none"> 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
--	---



Assignment of Eurovoc Descriptors




● Training phase

- Produce, for each descriptor, lists of general language lemmas (words) which are 'associated' with this descriptor (*associates*) by
 - compiling *metatexts* containing all documents which were manually indexed with a descriptor
 - using the monolingual keyword assignment tool to identify the most pertinent words (keywords, associates) of this metatext, plus their *weight* (keyness, association strength)


● Assignment phase

- Compare the lemma frequency list of a new text with all descriptor associate lists
- Use a statistical algorithm to calculate which descriptor list is most relevant to the text

➔ **Result:** a ranked list of the most suitable descriptors for this text




Examples for '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
gfcml	202.66	relation	100.35
fisherman	198.93	election	98.75
regulation	181.7	subsidiarity	96.82
...		...	



Calculation of the Descriptor Score


Document title: "Seizure of Plutonium at Munich Airport"




Initial, intuitive algorithm: Multiply the text lemma frequency with the log of the keyness of each descriptor and add the result to the score of the descriptor; divide the final score by the text length

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)
...



Further Assignment Algorithms Tried Out



TF.IDF (Salton, G. & C. Buckley, 1988: *Term Weighting Approaches in automatic text retrieval*. Information Processing and Management, vol. 1, 24, N° 5, pp. 513-523)

Cosine (Salton, G, 1989: *Automatic Text Processing: the Transformation, Analysis and Retrieval of Information by Computer*. Reading, Mass., Addison-Wesley)


Okapi (Robertson, S. E., S. Walker, M. Hancock-Beaulieu & M. Gatford, 1994: *Okapi at TREC-3*, Text Retrieval Conference TREC-3, U.S. National Institute of Standards and Technology, Gaithersburg, USA. NIST Special Publication 500-225, pp. 109-126)

Expected Mutual Information Measure (EMI) (Church K. & P. Hanks, 1989: *Word association norms, mutual information, and lexicography*. In ACL Proceedings, 27th Annual Meeting, Vancouver, pp. 76-83)


Home-grown: e.g. Descriptor Score += Sum of all (text lemma count * absolute frequency of associate / total frequency of lemma in training collection)

Mixed algorithms

...




Comparison of Manual and Automatic Assignment Results




- Keyword assignment is an abstract, highly conceptual task
- There are no clear rules which say that a certain descriptor is suitable or not
- Assignment results differ from one person to the next, sometimes even from one day to the next for the same person, and in any case they differ over time because the historical view changes.
- 80%, 60%, 30% overlapping results between people in different experiments

➔ Judging the automatic results comparing them to the manually assigned descriptors is an approximation.

➔ The manual results are not an absolute criterion for the assignment quality.




Automatically Assigned Descriptors




Top 23 English Eurovoc descriptors and their score assigned automatically to the Spanish policy document *Postura de la Unión Europea frente al descubrimiento del contrabando de plutonio (Attitude of the European Union towards the discovery of plutonium smuggling)* (383 words long)

Score	Descriptor (En)	Score	Descriptor (En)
97	<u>NUCLEAR SAFETY</u>	16	<u>EUROPOL</u>
62	<u>NUCLEAR NON-PROLIFERATION</u>	14	NUCLEAR ACCIDENT
43	<u>NUCLEAR FUEL</u>	14	BUDGETARY DISCHARGE
42	<u>NUCLEAR POWER STATION</u>	13	<u>UKRAINE</u>
38	<u>NUCLEAR TEST</u>	13	<u>CIS COUNTRIES</u>
34	<u>IAEA</u>	12	TRANSPORT OF DANGEROUS GOODS
32	<u>RADIOACTIVE WASTE</u>	12	RESEARCH AND DEVELOPMENT
29	<u>RADIOACTIVE MATERIALS</u>	12	EC GENERAL BUDGET
28	<u>NUCLEAR ENERGY</u>	11	<u>POLICE COOPERATION</u>
25	<u>ILLICIT TRADE</u>
21	<u>DECOMMISSIONING OF POWER STATIONS</u>		<u>Underlined:</u> manually assigned descriptors
18	<u>EABC</u>		<i>Italics:</i> further 'reasonable' descriptors
18	<u>ORGANIZED CRIME</u>		Normal: wrong, but semantically related
17	<u>CIS</u>		Strikethrough: wrong descriptors, semantically not related




Automatically Assigned Descriptors (Currently Best Results)




Top 23 English Eurovoc descriptors and their score assigned automatically to the Spanish policy document *Postura de la Unión Europea frente al descubrimiento del contrabando de plutonio* (*Attitude of the European Union towards the discovery of plutonium smuggling*) (383 words long; 11 manually assigned descriptors)

Score	Descriptor (En)	Score	Descriptor (En)
84	<u>IAEA</u>	39	<u>NUCLEAR POWER STATION</u>
82	<u>PLUTONIUM</u> (NT to <u>RADIOACTIVE MATERIALS</u>)	37	<u>DECOMMISSIONING OF POWER STATIONS</u>
79	<u>NUCLEAR FUEL</u>	37	<u>FUEL REPROCESSING</u> (RT to <u>NUCLEAR FUEL</u>)
72	<u>RADIOACTIVE MATERIALS</u>	37	<u>RADIOACTIVE WASTE</u>
71	<u>ILLICIT TRADE</u>	37	<u>NUCLEAR ENERGY</u>
67	<u>NUCLEAR SAFETY</u>	35	<u>EUROPOL</u> (Rank 20)
65	<u>NUCLEAR NON-PROLIFERATION</u>	35	<u>EAEC</u>
59	<u>POLICE COOPERATION</u>	35	TECHNOLOGICAL CHANGE
55	<u>CIS COUNTRIES</u>	34	<u>CS</u>
44	<u>PEACEFUL USE OF ENERGY</u> (RT to <u>NUCLEAR SAFETY</u>)...
43	<u>ACTION PROGRAMME</u>		Underlined: manually assigned descriptors
41	<u>ORGANISED CRIME</u> (Rank 12)		Italics: further 'reasonable' descriptors
40	<u>COMMUNITY PROGRAMME</u>		Normal: wrong, but semantically related
40	<u>NUCLEAR TEST</u>		Strike through: wrong descriptors, semantically not related



Sample Text (Es)



PREGUNTA ORAL B4-0034/94 con debate formulada por los diputados Noël MAMERE , Jaak VANDEMEULEBROUCKE al Consejo. Postura de la Unión Europea frente al descubrimiento del contrabando de **plutonio**


PREGUNTA ORAL O-0063/94 de conformidad con el artículo 40 del Reglamento de Noël Mamère y Jaak Vandemeulebroucke, en nombre del Grupo ARE al Consejo

Asunto: Postura de la Unión Europea frente al descubrimiento del contrabando de plutonio

Además de los casos recientes de tráfico ilegal de **plutonio**, desde 1991 se han descubierto más de 267 casos idénticos. Esta situación es preocupante porque el **plutonio** es un material altamente tóxico, así como una materia prima fundamental para la producción de bombas atómicas. ¿Puede comunicar el Consejo:

1. Cuál va a ser la contribución de la Unión Europea a los esfuerzos internacionales destinados a poner fin a este tráfico ilegal?
2. Si está dispuesto a poner a disposición de la Agencia **EUROPOL** los medios financieros y científicos necesarios para que ésta coordine la lucha contra dicho contrabando?
3. Qué iniciativas piensa desarrollar a fin de elaborar programas de asistencia técnica destinados a crear sistemas de control y de gestión de las existencias de materias fisibles, en particular en los países de la Europa central y del Este y de la **CEI**? A este respecto, ¿puede comprometerse a restablecer, e incluso incrementar, el esfuerzo financiero realizado desde 1993 por iniciativa del Parlamento Europeo que permitía desarrollar intercambios de conocimientos técnicos por lo que a la seguridad se refiere con los expertos nucleares de dichos países (partida B4-2001)?
4. Si la Unión Europea piensa fortalecer la cooperación con el **OIEA** a fin de crear un fichero internacional de huellas digitales nucleares (nuclear fingerprints) que facilitará la identificación de los materiales nucleares descubiertos e incrementará la eficacia de la lucha contra dicho contrabando?
5. Si va a desarrollar las iniciativas necesarias para que, con motivo de la próxima revisión del Tratado sobre la no proliferación nuclear en 1995, los Estados signatarios se comprometan más firmemente de cara a un estricto control de las disposiciones del Tratado?

Presentación: 09.09.1994 Transmisión: 12.09.1994 Plazo límite: 03.10.1994



Institute for the Protection
and Security of the Citizen

Sample Text (En)



JOINT
RESEARCH
CENTRE
EUROPEAN COMMISSION

ORAL QUESTION B4-0034/94 with debate by Noël MAMERE , Jaak VANDEMEULEBROUCKE to the Council.
Attitude of the European Union towards the discovery of **plutonium** smuggling


ORAL QUESTION O-0063/94 pursuant to Rule 40 of the Rules of Procedure by Noël Mamère and Jaak Vandemeulebroucke, on behalf of the ARE Group to the Council

Subject: **Attitude of the European Union towards the discovery of plutonium smuggling**

In addition to the recent instances of **plutonium** smuggling, more than 267 identical cases have been discovered since 1991. This is an alarming situation given that **plutonium** is a particularly toxic material and a raw material for the production of nuclear weapons.


1. What contribution will the European Union be making to the international efforts to put an end to this illegal trade?
2. Will the Council provide the **EUROPOL** Agency with the financial and scientific resources to coordinate action against this trade?
3. What action will the Council take to draw up technical assistance programmes for the development of management and control systems for fissile material stocks, particularly within the countries of central and eastern Europe and the **CIS**? Will it undertake to reinstate, or even increase, the funding provided since 1993 at the initiative of the European Parliament allowing know-how on safety to be exchanged with nuclear experts from those countries (line B4-2001)?
4. Will the European Union be stepping up cooperation with the **IAEA** to establish an international 'nuclear fingerprints' register to help identify the nuclear materials discovered and increase the effectiveness of the fight against this trade?
5. Will the Council take the necessary measures to ensure that, in the next review of the Nuclear Non-Proliferation Treaty in 1995, the signatory states commit themselves more firmly to strict monitoring of the provisions of the Treaty?

Tabled: 09.09.1994 Forwarded: 12.09.1994 Deadline for reply: 03.10.1994




Institute for the Protection
and Security of the Citizen

Assignment Results on Training Collection (Spanish)




JOINT
RESEARCH
CENTRE
EUROPEAN COMMISSION

MaxRank	First Experiments 28 June 2001		Latest Experiments 4 September 2001	
	Recall CT	Precision CT	Recall CT	Precision CT
1	9%	62%	13%	89%
2	15%	54%	23%	81%
3	21%	47%	32%	74%
5	28%	39%	45%	61%
7	34%	33%	53%	52%
10	40%	27%	62%	42%
15	47%	21%	70%	32%
20	52%	17%	75%	26%
25	56%	15%	79%	22%
30	58%	13%	82%	19%
50	66%	9%	88%	12%
100	73%	5%	93%	7%
Formula:	TF * log (keyness)		0.61 Cosine + 0.21 Okapi + 0.18 SumTfidf	
Descriptors:	2870		3643	
Precision with arbitrary distribution:	0.24%		0.19%	
Training material / descriptor	> 10KB		> 2KB	
Aver. No. of manually assigned descr.	6.91		6.91	
Documents in test collection	3743		3955	



Formula of the Currently Best Experiments



$$\Phi = 0.61 \frac{COSINE}{\max(COSINE)} + 0.21 \frac{Okapi}{\max(Okapi)} + 0.18 \frac{SumTfidf}{\max(SumTfidf)}$$

$$TFIDF_{l,d} = TF_{l,d} \cdot \left(\log_2 \frac{N}{DF_l} + 1 \right)$$


$$COSINE(d,t) = \frac{\sum_{l \in d \cap t} TFIDF_{l,d} \cdot TFIDF_{l,t}}{\sqrt{\left(\sum_{l \in d} TFIDF_{l,d}^2 \right) \cdot \left(\sum_{l \in t} TFIDF_{l,t}^2 \right)}}$$

$$SumTfidf(d,t) = \sum_{l \in d \cap t} TFIDF_{l,d} \cdot TFIDF_{l,t}$$


TFIDF = Term Frequency,
Inverse Document Frequency

l = lemma,
d = Eurovoc descriptor
|d| = number of associates in descriptor (size)
M = average size of descriptors
t = new text
N = number of descriptors used
DF = document frequency (n° of descriptors for which the lemma is an associate)



$$Okapi_{t,d} = \sum_{l \in t \cap d} \log\left(\frac{N - DF_l}{DF_l}\right) \frac{TF_{l,d}}{TF_{l,d} + \frac{|d|}{M}}$$





Challenges and Difficulties



- **There is not enough training material for all descriptors**
Training material for < 3700 Es, En and De descriptors > 2KB (½ page)
- **Training material with very different text sizes** ranging from short titles to 20-page texts
- **The training material is very much biased by the interests of the EP**
 - e.g. many 'associates' of the descriptor MAURITANIA pertain to the semantic field 'fishery'
- Some descriptors were used thousands of times, others never
 - very different lengths of 'associate' lists
 - very different assignment likelihood

		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	...			

		Ferber's Approach (1997)			
<p>Reginald Ferber (1997). <i>Automated Indexing with Thesaurus Descriptors: A co-occurrence Based Approach to Multilingual Retrieval</i>. In Peters Carol & Thanos Costantino (eds.) <i>Research and Advanced Technology for Digital Libraries</i>. 1st European Conference (ECDL'97). Springer Lecture Notes, Berlin, pp. 232-255.</p>					
<ul style="list-style-type: none"> ● Goal: indexing for cross-lingual document retrieval ● Corpus is very homogeneous: <ul style="list-style-type: none"> ● 80.000 manually indexed <i>titles</i> of publications ● OECD thesaurus, ca. 4000 descriptors in 4 languages ● Uses the absolute word frequency (instead of the 'keyness') ● Calculates the descriptor-word association with a variation of the <i>Expected Mutual Information Measure</i> (EMIM): $p(i\&j)/p(i)^x \cdot p(j)^y$ (Church & Hanks, 1989: <i>Word association norms, mutual information, and lexicography</i>) ● No consideration of the hierarchical structure (BTs and NTs) and of RTs ● Good results 					

Outlook - To Do

- Improve performance
 - Improve performance by identifying optimal algorithm
 - Test and tune with parallel texts (texts and their translations)
 - Test on documents that are *not* part of the training corpus


- Apply to real world scenarios
 - Apply to further languages (currently trained for En, Es and De)
 - Incorporate with other tools such as crawler and extraction software

Text Clustering and Visualisation

- Calculation of document similarity by comparing keyword lists of documents
 - Criterion: number of keywords in common, weighed according to inverted frequency (and other algorithms)
 - Eurovoc indexing allows cross-lingual document comparison
optimal assignment results are less important than assignment consistency (e.g. 'fishery' and MAURITANIA)

- Clustering of documents (hierarchical clustering), using similarity
 - Representation as hierarchical clustering tree, or display of the X most similar documents to a given one


- Multilingual visualisation of document collections via Eurovoc indexing
 - Two-dimensional representation of the multidimensional document space
 - Kohonen maps (neural networks)
 - JRC document charts (Hagman et al., 2000)
 - Document maps with *Cartia's product ThemeScape* (Steinberger et al., 2000) (<http://www.cartia.com> and <http://demo.cartia.com/jrcdescriptors/map1024.html> for JRC data)



ipsc
Institute for the Protection
and Security of the Citizen

Document Clustering

Measuring Document Similarity




**JOINT
RESEARCH
CENTRE**
EUROPEAN COMMISSION

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

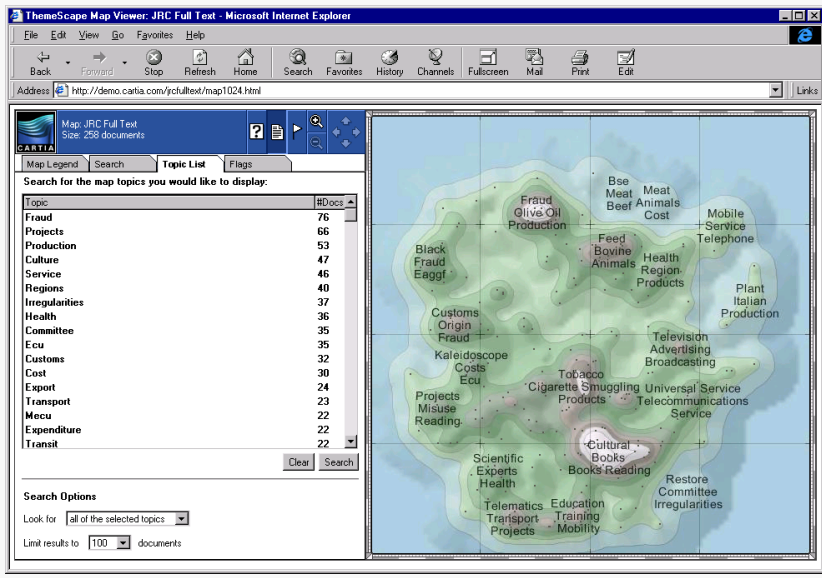


ipsc
Institute for the Protection
and Security of the Citizen

Document Map (ThemeScape)




**JOINT
RESEARCH
CENTRE**
EUROPEAN COMMISSION



The screenshot shows the ThemeScape Map Viewer interface. On the left, there is a sidebar with a 'Topic List' containing various categories and their corresponding document counts. The main area displays a map where these topics are represented as colored regions. The browser window title is 'ThemeScape Map Viewer: JRC Full Text - Microsoft Internet Explorer' and the address bar shows 'http://demo.cartia.com/jrcfulltext/map1024.html'.


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



ipsc
Institute for the Protection
and Security of the Citizen

Document Map – 2

(Search word 'olive' + documents in area)



**JOINT
RESEARCH
CENTRE**
EUROPEAN COMMISSION

ThemeScape Map Viewer: JRC Full Text - Microsoft Internet Explorer

Address: http://demo.cartia.com/JRCFullText/map1024.html

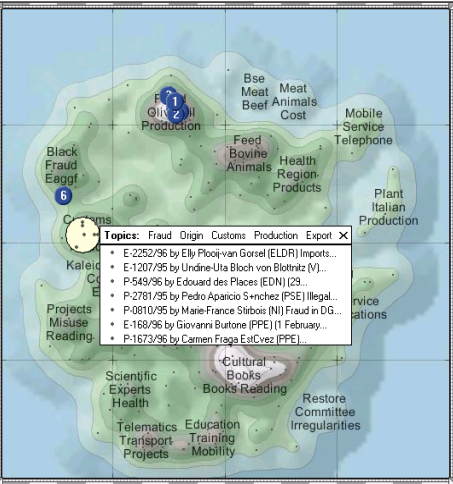
Map: JRC Full Text
Size: 259 documents

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 JesPounds CabezCenin Alonso (PSE) and Juan Colino
- 5 E-2913/96 by Salvador Garriga Polledo (PPE) COM in olive
- 6 E-0190/96 by Iseccionigo MCndez de Vigo (PPE) Commission investigations

New Search 1-6 of 6 < Back Next >




Topics: Fraud Origin Customs Production Export

- E-2252/96 by Ely Flooi-van Gorzel (ELDR) Imports...
- E-1207/96 by Undine-Uta Bloch von Blotnitz (V...
- P-549/96 by Edouard des Places (EDN) [23...
- P-2781/96 by Pedro Apaicio S-nchez (PSE) Illegal...
- P-0810/96 by Marie-France Stibois (NI) Fraud in DG...
- E-168/96 by Giovanni Buttone (PPE) (1 February...
- P-1673/96 by Carmen Flaiga EstCvez (PPE)...



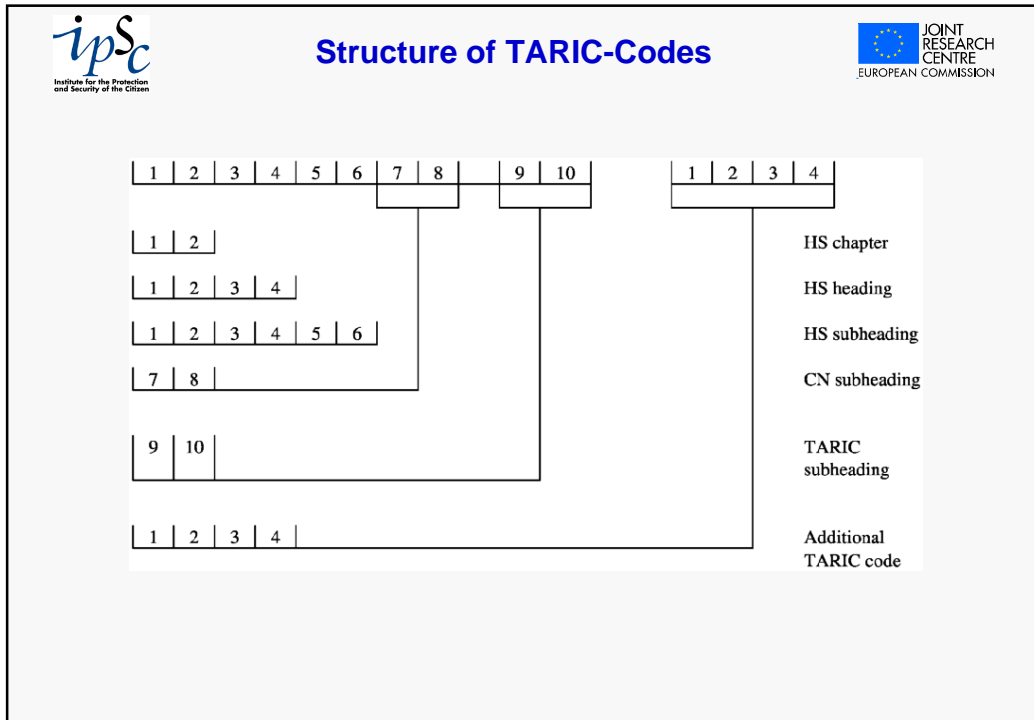
ipsc
Institute for the Protection
and Security of the Citizen

Recognition of Product Groups




**JOINT
RESEARCH
CENTRE**
EUROPEAN COMMISSION

- Use the multilingual product nomenclature: *Customs Tariff Code* TARIC
- Exists in all 11 official EU languages
- Hierarchically organised:
 - TARIC uses codes with 10 digits
 - based on *Combined Nomenclature CN*, with 8 digits
 - based on the *Harmonised System HS*, with 6 digits (developed by the Brussels *Customs Council*, now *World Customs Organisation WCO*)
- Consists of a numerical code and its description in different languages
- ➔ We want to produce lists of lemmas (and their weight) for each product or each product group (collaboration with the CL department of the University of Munich (CIS at LMU))




Some German TARIC Entries (1)


DE 010000000 80	LEBENDE TIERE
DE 010100000 80	Pferde, Esel, Maultiere und Maulesel, lebend
DE 010110000 10	Pferde
DE 010110000 80	reinrassige Zuchttiere
DE 0101190000 80	andere
DE 0101191000 80	zum Schlachten
DE 0101199000 80	andere
DE 0101200000 80	Esel, Maultiere und Maulesel
DE 0101201000 80	Esel
DE 0101209000 80	Maultiere und Maulesel
DE 0102000000 80	Rinder, lebend
DE 0102100000 80	reinrassige Zuchttiere
DE 0102101000 80	Färsen (weibliche Rinder, die noch nicht gekalbt haben)
DE 0102103000 80	Kühe
DE 0102109000 80	andere
DE 0102900000 80	andere
DE 0102900500 10	Hausrinder
DE 0102900500 80	mit einem Gewicht von 80 kg oder weniger
DE 0102900510 80	männliche Jungrinder zum Mästen
DE 0102900520 80	Färsen, nicht zum Schlachten, der Höhenrassen Grauvieh, Braunvieh, Gelbvieh und Pinzgauer Fleckvieh
DE 0102900530 80	Färsen, nicht zum Schlachten, der Schwyzer und Freiburger Rasse
DE 0102900540 80	Färsen, nicht zum Schlachten, der Simmentaler Fleckvieh Rasse
DE 0102900550 80	Stiere, nicht zum Schlachten, der Schwyzer, Simmentaler Fleckvieh oder Freiburger Rasse
DE 0102900590 80	andere
DE 0102902100 10	mit einem Gewicht von mehr als 80 kg bis 160 kg
DE 0102902100 80	zum Schlachten
DE 0102902910 80	männliche Jungrinder zum Mästen
DE 0102902920 80	Färsen der Höhenrassen Grauvieh, Braunvieh, Gelbvieh und Pinzgauer Fleckvieh
DE 0102902930 80	Färsen der Schwyzer und Freiburger Rasse
DE 0102902940 80	Färsen der Simmentaler Fleckvieh Rasse
DE 0102902950 80	Stiere der Schwyzer, Simmentaler Fleckvieh oder Freiburger Rasse
DE 0102902990 80	andere
DE 0102904100 10	mit einem Gewicht von mehr als 160kg bis 300kg
DE 0102904100 80	zum Schlachten
DE 0102904910 80	männliche Jungrinder zum Mästen
DE 0102904920 80	Färsen und Kühe der Höhenrassen Grauvieh, Braunvieh, Gelbvieh und Pinzgauer Fleckvieh
DE 0102904930 80	Färsen und Kühe der Schwyzer und Freiburger Rasse
DE 0102904940 80	Färsen und Kühe der Simmentaler Fleckvieh Rasse
DE 0102904950 80	Stiere der Schwyzer, Simmentaler Fleckvieh oder Freiburger Rasse




Some German TARIC Entries (2)



<p>DE 0303800000 80 Fischlebern, Fischrogen und Fischmilch</p> <p>DE 0303801000 80 Fischrogen und Fischmilch, zum Herstellen von Desoxyribonucleinsäure oder Protaminsulfat</p> <p>DE 0303809000 80 andere</p> <p>DE 0303809010 10 Rogen</p> <p>DE 0303809010 80 von Heringen (Clupea harengus, Clupea pallasii)</p> <p>DE 0303809019 80 andere</p> <p>DE 0303809091 10 andere</p> <p>DE 0303809091 80 von Heringen (Clupea harengus, Clupea pallasii)</p> <p>DE 0303809099 80 andere</p> <p>DE 0304000000 80 Fischfilets und anderes Fischfleisch (auch fein zerkleinert), frisch, gekühlt oder gefroren</p> <p>DE 0304100000 80 frisch oder gekühlt</p> <p>DE 0304101100 10 Filets</p> <p>DE 0304101100 20 von Süßwasserfischen</p> <p>DE 0304101100 80 von Forellen der Arten Salmo trutta, Oncorhynchus mykiss, Oncorhynchus clarki, Oncorhynchus aguabonita und Oncorhynchus gilae</p> <p>DE 0304101110 80 Forellen (Oncorhynchus mykiss)</p> <p>DE 0304101190 80 andere</p> <p>DE 0304101300 80 vom Pazifischen Lachs (Oncorhynchus nerka, Oncorhynchus gorbuscha, Oncorhynchus keta, Oncorhynchus tshawytscha, Oncorhynchus kisutch, Oncorhynchus masou und Oncorhynchus rhodurus), Atlantischen Lachs (Salmo salar) und Donaulachs (Hucho hucho)</p>	<p>DE 0304101311 10 Atlantischer Lachs (Salmo salar)</p> <p>DE 0304101311 80 wilde</p> <p>DE 0304101321 10 andere</p> <p>DE 0304101321 80 ganze Fischfilets, mit einem Gewicht von mehr als 300 g/Stück</p> <p>DE 0304101329 80 andere Fischfilets oder Filetteile, mit einem Gewicht von 300 g/Stück oder weniger</p> <p>DE 0304101390 80 andere</p> <p>DE 0304101900 80 von anderen Süßwasserfischen</p> <p>DE 0304101920 80 von Aalen (Anguilla-Arten)</p> <p>DE 0304101930 80 von Karpfen</p> <p>DE 0304101990 80 andere</p> <p>DE 0304103100 10 andere</p> <p>DE 0304103100 80 vom Kabeljau (Gadus morhua, Gadus ogac, Gadus macrocephalus) und von Fischen der Art Boreogadus saida</p> <p>DE 0304103110 80 der Art Gadus morhua</p> <p>DE 0304103190 80 andere</p> <p>DE 0304103300 80 vom Köhler (Pollachius virens)</p> <p>DE 0304103500 80 vom Rotbarsch, Goldbarsch oder Tiefenbarsch (Sebastes-Arten)</p> <p>DE 0304103800 80 andere</p> <p>DE 0304103810 80 vom Schellfisch (Melanogrammus aeglefinus)</p> <p>DE 0304103820 80 vom Schwarzen Heilbutt (Reinhardtius hippoglossoides) und vom Atlantischen Heilbutt (Hippoglossus hippoglossus)</p>
--	---



Summary



- **Aim:** system carrying out document retrieval, information extraction and visualisation of the extracted information
- Focus on **cross-lingual applications**, achieved by combining texts with thesauri and nomenclatures
 - e.g. *Eurovoc* and
 - *Customs Tariff Code* TARIC or *Combined Nomenclature* CN
- **Approach to Eurovoc indexing:** usage of training corpus to produce long lists of associated lemmas ('associates'), which point to the descriptors
- **Result:** a ranked list of language-independent descriptors for each text
- **Usage:** visualisation of multilingual document collections
 - single documents in 'document profiles'
 - multilingual 'document maps'
 - a ranked list of the most similar documents, even in different languages



Working at the JRC (www.jrc.it)



- 1/3/5-year contracts (<http://www.cordis.lu/nppr-candidature> for AT3)
- Grants: post-doc and Ph.D. (payment and conditions similar to the conditions for Marie-Curie fellowships: <http://www.cordis.lu/improving/fellowships/home.htm>). More info: http://www.jrc.org/what_we_offer/research_fellowships.htm
- Visiting scientist
- Seconded national expert
- Internships ('stagiaire': currently we are not allowed to pay!)
 - apartment (50 Euro + ca. 70 Euro for bills, inclusive washing of the bedding)
 - lunch (mensa food costs: 1-2 Euro/meal)
 - free bus service to and from the JRC
 - outdoor leisure activities (mountains, lakes) and over 30 clubs:

Sports activities include Aikido, alpinism (trekking, skiing, climbing), badminton, basketball, billiard, bowling, bridge, chess, cricket, cycling, dance, diving, football, fitness training, gymnastics, golf, handball, horse riding, hockey, ice skating, judo, karate, rugby, shooting, skiing, square dance, swimming, table tennis, tennis, volleyball and wind surfing. Furthermore there are clubs for **cultural and other activities** covering such diverse areas as amateur radio, ceramics, choir singing, computers, gardening, mineralogy, philately, photography, television and ufology.