

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Joint Research Centre

Agenda

- Who we are and what we do
- Eurovoc Thesaurus
- Automatic assignment of thesaurus descriptors to text
 - Training Phase
 - Assignment Phase
- Document Similarity Calculation and Translation Identification
- Application Areas of the Technology

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Joint Research Centre

Goal of JRC's Language Technology work

IDoRA System: *Intelligent Document Retrieval and Analysis*

- **Retrieval** of potentially relevant texts
- **Text analysis** and extraction of information from texts
- **Visualisation** of the contents

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Focus of JRC's Language Technology work

Joint Research Centre

- Multilingual and **cross-lingual** applications
- Also for languages of EU **Candidate Countries**
- Many languages; few human resources
 - Applications using more **statistics** and less language-specific resources

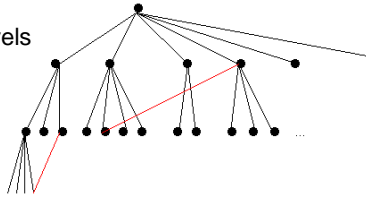
EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Eurovoc Thesaurus

<http://europa.eu.int/celex/eurovoc>

Joint Research Centre

- **Multilingual list of terms** about many different subject areas (wide coverage)
- Developed by the European Parliament (EP) and others
- Actively used **to index (catalogue) and retrieve documents** in large collections (fine-grained classification and cataloguing system)
- **Hierarchically organised** into a maximum of 8 levels
 - top level: 21 fields
 - next level: 127 micro-thesauri
 - total: 5933 descriptors (version 3.0)
 - 5877 reciprocal relations (BT, NT)
 - 2730 **reciprocal associations (RT)**



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Eurovoc (Top Level and Detail)

Joint Research Centre

<ul style="list-style-type: none"> 04 Politics 08 International Relations 10 European Communities 12 Law 16 Economics 20 Trade 24 Finance <li style="color: blue;">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 	<div style="color: blue; font-weight: bold; margin-bottom: 5px;">28 SOCIAL QUESTIONS</div> <ul style="list-style-type: none"> 2806 family 2811 migration 2816 demography and population 2821 social framework 2826 social affairs <li style="color: blue;">2831 culture and religion <li style="padding-left: 20px;">arts <li style="padding-left: 20px;">cultural policy <li style="color: blue;">culture <li style="padding-left: 20px;">acculturation <li style="padding-left: 20px;">civilization <li style="padding-left: 20px;">cultural difference <li style="color: blue;">cultural identity <li style="padding-left: 20px;">RT: protection of minorities (1236) <li style="padding-left: 20px;">RT: socio-cultural group (2821) <li style="padding-left: 20px;">cultural pluralism <li style="padding-left: 20px;">popular culture <li style="padding-left: 20px;">regional culture <li style="padding-left: 20px;">religion 2836 social protection 2841 health 2846 construction and town planning
---	--

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Eurovoc Users

Joint Research Centre

Documentation Centres and Libraries of:

- European Parliament
- DG OPOCE
- Belgium:
 - Senate
 - La Chambre
- Portugal: Assembleia da Republica
- Sweden: Riksdag
- Spain:
 - El Senado
 - Congreso de los Diputados
- Switzerland: Assemblée Fédérale

- Czech Republic
 - Chamber of Deputies
 - Euro Info Centre
 - European Documentation Centre
 - Info Centre of the EU
 - Supreme Audit Office
 - Parliamentary Library
- Lithuanian Seimas
- Polish Sejm
- Slovenian Državni zbor
- Romanian Camera Deputatilor
- Russian Duma
- Albanian Parliament
- Croatia
- Ukraine



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Eurovoc Languages


Joint Research Centre

- Used by the EP and DG OPOCE for **all 11 official EU languages**

- Also exists for:
Albanian, Czech, Croatian, Hungarian, Latvian, Lithuanian, Polish, Romanian, Russian, Slovak, Slovenian

- Consider using Eurovoc: **Armenia, Bosnia-Herzegovina, Bulgaria, Estonia, France, Georgia, Iceland, Macedonia, Turkey**

- Most multilingual thesaurus in existence? (currently **22 languages**)



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Automatic Indexing: Challenge


Joint Research Centre

- Descriptors are mostly abstract multi-word concepts, e.g.
 - PROTECTION OF MINORITIES
 - FISHERY MANAGEMENT
 - CONSTRUCTION AND TOWN PLANNING
 - SIMPLIFICATION OF FORMALITIES
 - PLUTONIUM
 - FRANCE

→ Searching for descriptors (baseline) in text is not a solution:

Maximum recall	~ 30%,
Maximum precision	~ 7%

→ **Keyword Assignment** as opposed to keyword extraction



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre


JRC's Statistical / Associative Approach

Joint Research Centre

1. **Training Phase:** Identify many (statistically or semantically) **related words (associates)**
2. **Assignment phase:** Assign descriptor if many of its *associates* are present in text.

FISHERY MANAGEMENT

f fishery_resource	54.4721542368385
f fishing	49.111563204862
f fish	46.196436023147
f common_fishery_policy	44.6741845971235
f fishery	44.1911518447189
f fishing_activity	43.3777671334009
f fly_the_flag	42.8744724542378
f aquaculture	39.2749719215554
f conservation	38.3480454820621
f vessel	37.911138722495
f fishing_vessel	37.8343365844963
f catch	36.8503034704154
f fish_stock	34.5283935973103
f tacs	34.388453583343
f allowable_catch	33.2880590561664
f catch_quota	32.2683540654092
f control_system	31.1753892078216
f fish_for	29.8386698340017
f nautical_mile	29.541061528168
f fishing_right	29.1916760888221
f centimetre	28.7167313169535
f control_measure	28.0527345432075
f gross_tonnage	28.0043616725124
f fishing_zone	27.8678836557192




EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Training: Text Normalisation

Joint Research Centre

- Linguistic pre-processing = normalisation of the text
 - **Lemmatisation** (base-form reduction of words) and lower-casing:
Transporting → transport
 - Mark-up of **multi-word expressions**
'plant' → 'green_plant' vs. 'power_plant'
 - **Stop word lists** to avoid words that are not content-bearing
general: are, they, having, in spite of, interesting,
domain-specific: question, answer, commission, article




Training: Produce Associate Lists

- Using a large collection of manually indexed documents (training corpus)
- For each descriptor D_1 , take all documents indexed with D_1
- identify the **statistically salient words** in each of these texts
- **join these lists** of statistically salient words, e.g. **RADIOACTIVE MATERIALS**

radioactive ukraine resolution plutonium deuterium parliament nuclear blotnitz ...	+	plutonium deuterium assembly nuclear schmidt radioactive korea iaea ...	+	Illegal_traffic chernobyl radioactive ukrainian plutonium lithium dangerous mox ...	=	radioactive (3) plutonium (3) nuclear (2) deuterium (2) Illegal_traffic (1) chernobyl (1) ...
--	---	---	---	--	---	---

- **Normalise the weight** according to a number of different criteria.
- **Result of Training:** Weighed associate lists for all descriptors



Associate List: RADIOACTIVE MATERIALS

→	deuterium	35.7836791092845
→	lithium	33.0805724769899
→	thorium	32.560703225522
→	tritium	32.0826451843048
→	nuclear_material	13.79399100837
→	radioactive_material	7.84970673161556
→	plutonium	6.72955494180221
→	radioactive_substance	6.43422856440347
→	nuclear	5.851612117697
→	undine_uta_bloch_von_blotnitz	5.53278869694883
→	radioactive	4.89399300382035
→	nuala_ahern	4.04706620369489
→	radon	4.03336435560442
→	mox	3.5654196472221
→	uranium	3.33954480260962
→	illegal_traffic	3.03072833135354




Associate List: FISHERY MANAGEMENT

Joint Research Centre

fishery-related

management-related

fishery_resource	54.4721542368385
fishing	49.111563204862
fish	46.196436023147
common_fishery_policy	44.6741845971235
fishery	44.1911518447189
fishing_activity	43.3777671334009
fly_the_flag	42.8744724542378
aquaculture	39.2749719215554
conservation	38.3480454820621
vessel	37.911138722495
fishing_vessel	37.8343365844963
catch	36.8503034704154
fish_stock	34.5283935973103
tacs	34.388453583343
allowable_catch	33.2880590561664
catch_quota	32.2683540654092
control_system	31.1753892078216
fish_for	29.8386698340017
nautical_mile	29.541061528168
fishing_right	29.1916760888221
centimetre	28.7167313169535
control_measure	28.0527345432075
gross_tonnage	28.0043616725124
fishing_zone	27.8678836557192



Assignment Phase

Joint Research Centre

- Normalise new document (lemmatise, multi-word mark-up)
- Produce lemma frequency list (excluding stop words)
- Calculate similarity between lemma frequency list and descriptor associate lists, using statistical formulae

Word	Freq.
convention	6
agreement	5
europaan_community	3
exchange_of_letter	3
goods	3
kingdom_of_norway	3
kingdom_of_sweden	3
recommendation	3
republic_of_austria	3
republic_of_finland	3
republic_of_iceland	3
simplification_of_formality	3
swiss_confederation	3
trade_in	3
approve	2
community	2
empower	2
joint_committee	2

STOP WORDS

SIMPLIFICATION OF FORMALITIES

...

56.4721542368385	fishery_resource
49.111563204862	fishing
46.196436023147	fish
44.6741845971235	common_fishery_policy
44.1911518447189	fishery
43.3777671334009	fishing_activity
42.8744724542378	fly_the_flag
39.2749719215554	aquaculture
38.3480454820621	conservation
37.911138722495	vessel
37.8343365844963	fishing_vessel
36.8503034704154	catch
34.5283935973103	fish_stock
34.388453583343	tacs
33.2880590561664	allowable_catch
32.2683540654092	catch_quota
31.1753892078216	control_system
29.8386698340017	fish_for
29.541061528168	nautical_mile
29.1916760888221	fishing_right
28.7167313169535	centimetre
28.0527345432075	control_measure
28.0043616725124	gross_tonnage
27.8678836557192	fishing_zone

56.4721542368385	fishery_resource
49.111563204862	fishing
46.196436023147	fish
44.6741845971235	common_fishery_policy
44.1911518447189	fishery
43.3777671334009	fishing_activity
42.8744724542378	fly_the_flag
39.2749719215554	aquaculture
38.3480454820621	conservation
37.911138722495	vessel
37.8343365844963	fishing_vessel
36.8503034704154	catch
34.5283935973103	fish_stock
34.388453583343	tacs
33.2880590561664	allowable_catch
32.2683540654092	catch_quota
31.1753892078216	control_system
29.8386698340017	fish_for
29.541061528168	nautical_mile
29.1916760888221	fishing_right
28.7167313169535	centimetre
28.0527345432075	control_measure
28.0043616725124	gross_tonnage
27.8678836557192	fishing_zone

56.4721542368385	fishery_resource
49.111563204862	fishing
46.196436023147	fish
44.6741845971235	common_fishery_policy
44.1911518447189	fishery
43.3777671334009	fishing_activity
42.8744724542378	fly_the_flag
39.2749719215554	aquaculture
38.3480454820621	conservation
37.911138722495	vessel
37.8343365844963	fishing_vessel
36.8503034704154	catch
34.5283935973103	fish_stock
34.388453583343	tacs
33.2880590561664	allowable_catch
32.2683540654092	catch_quota
31.1753892078216	control_system
29.8386698340017	fish_for
29.541061528168	nautical_mile
29.1916760888221	fishing_right
28.7167313169535	centimetre
28.0527345432075	control_measure
28.0043616725124	gross_tonnage
27.8678836557192	fishing_zone

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

Joint Research Centre

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Formulae tested for descriptor assignment

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

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

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

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

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

Term Frequency, Inverse Document Frequency Considers occurrence frequency of lemma (l) in meta-text (TF_{l,t}) and number of descriptors (d) for which the lemma is an associate (DF_l)

Cosine uses TF.IDF; computes the angle of two multi-dimensional vectors (of the document (t) and of the descriptor associate list)

Okapi considers occurrence frequency of lemma as an associate (DF_l); the number of associates in the associate list (size, |d|); the average size of descriptor associate lists (M); the total number of descriptors used (N)

'Scalar Product' adds product of TF.IDF values of associates and text lemmas

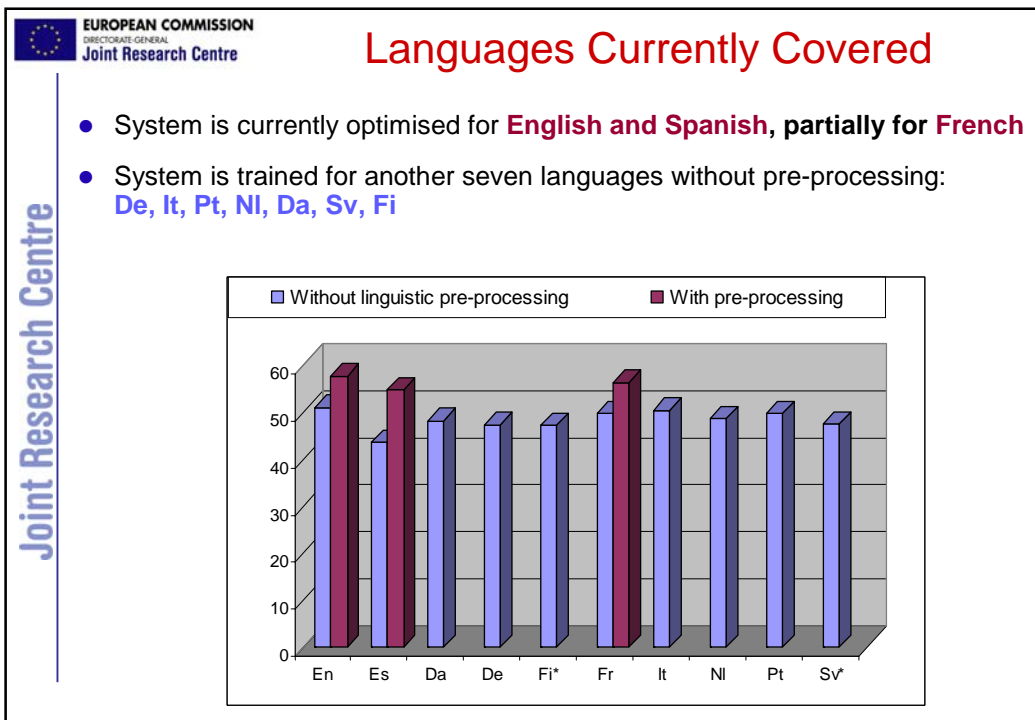
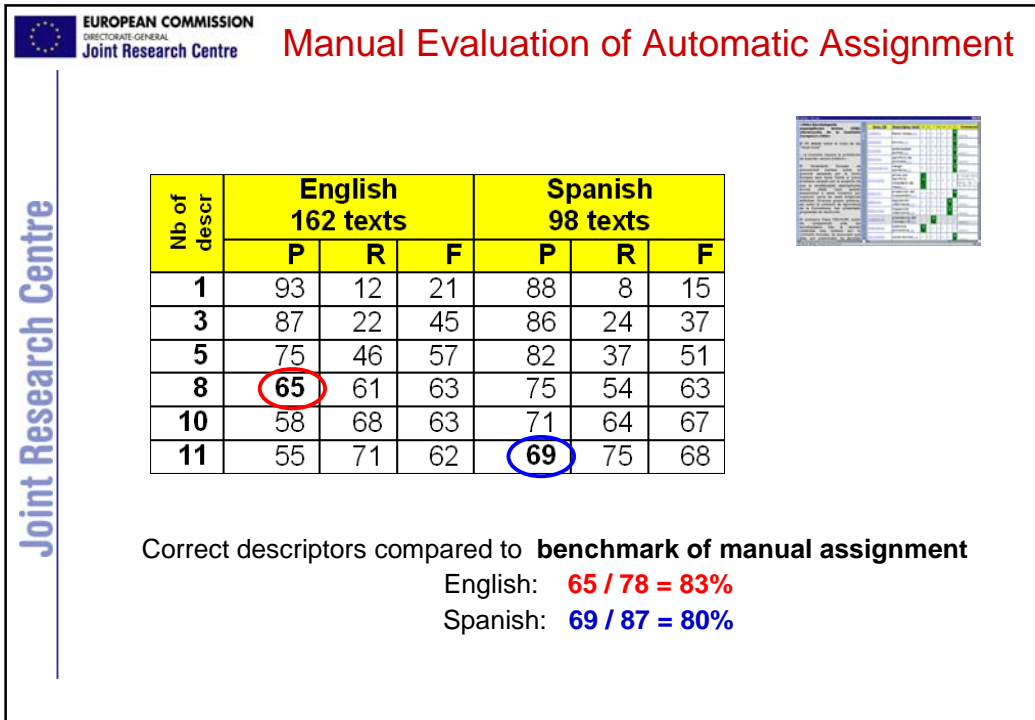
'622' mixed formula, uses all of the above

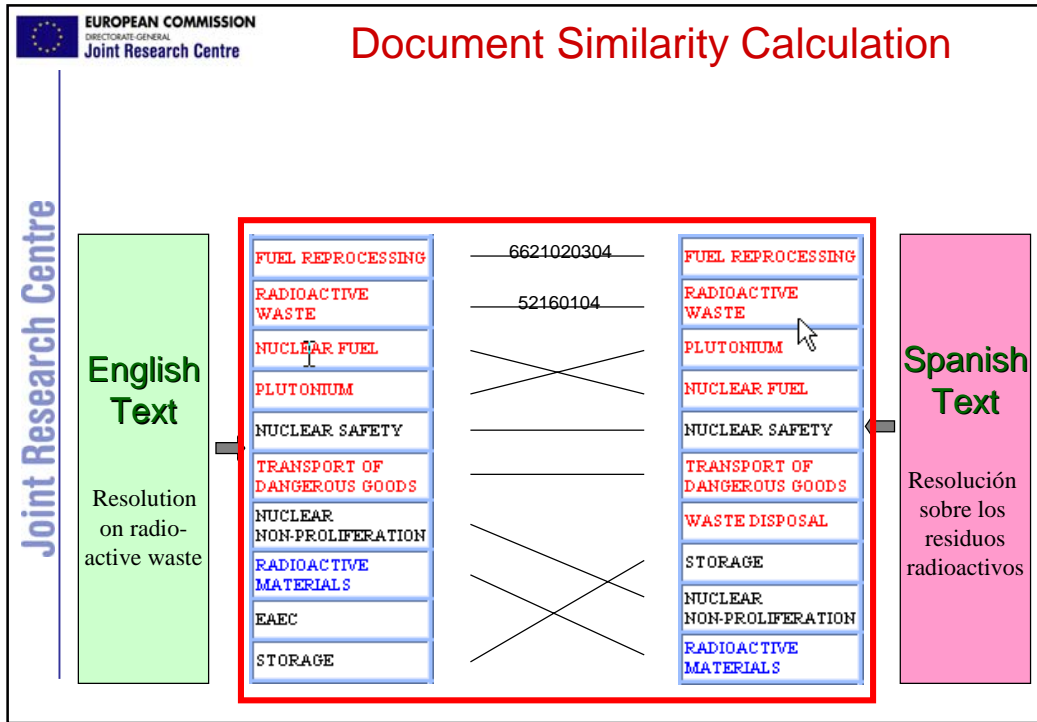
Joint Research Centre

EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Manual Evaluation of the Assignment

Desc ID	Descriptor text	B	S	?	bt	nt	G	V	Comment
72060411	Reino Unido(26)	c	c	c	c	c	c	c	victoria
56260205	bovino(25)	c	c	c	c	c	c	c	victoria
56310601	enfermedad animal(22)	c	c	c	c	c	c	c	victoria
60360104	sacrificio de animales(21)	c	c	c	c	c	c	c	victoria
284104041103	riesgo sanitario(17)	c	c	c	c	c	c	c	victoria
5611010501	prima por sacrificio voluntario de reses(17)	c	c	c	c	c	c	c	I think this unsuitable. unusual way to 08-APR-03 vict
20260102	protección del consumidor(17)	c	c	c	c	c	c	c	victoria
56060104	legislación veterinaria(17)	c	c	c	c	c	c	c	victoria
5606010401	inspección veterinaria(15)	c	c	c	c	c	c	c	victoria
1006020102	presidencia del Consejo CE(15)	c	c	c	c	c	c	c	victoria
2841040403	medicina preventiva(14)	c	c	c	c	c	c	c	victoria
6011010604	carne bovina(14)	c	c	c	c	c	c	c	victoria






EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Results for Similarity Calculation and Translation Spotting

Joint Research Centre

Task: find Spanish translations of English source document in a parallel text collection

	Search Space	Without length factor	With length factor
1) Simple document similarity (DS)	820 Es	90.61%	96.83%
2) DS considering the length of documents	820 Es	00.12%	01.71%
3) Different text type	795 Es	84.28%	90.31%
4) Mixed-language search space	410 Es + 410 En	69.68%	81.91%
5) DS correcting mono-lingual bias (83%)	410 Es + 410 En	92.91%	96.82%



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Is there a Translation?

- Setting a threshold; juggling precision and recall


Test bed	Average similarity	Threshold	Recall	Noise (1-Precision)
Set T1 (820)	0.82	0.70	90%	2.2%
Set T2 (795)	0.79	0.70	76.5%	5%

- Searching for a translation where there is none:
Searching in T2 for documents of T1
→ 4.15% noise

→ Best threshold depends on:

- Document set
- Requirement: high recall or high precision

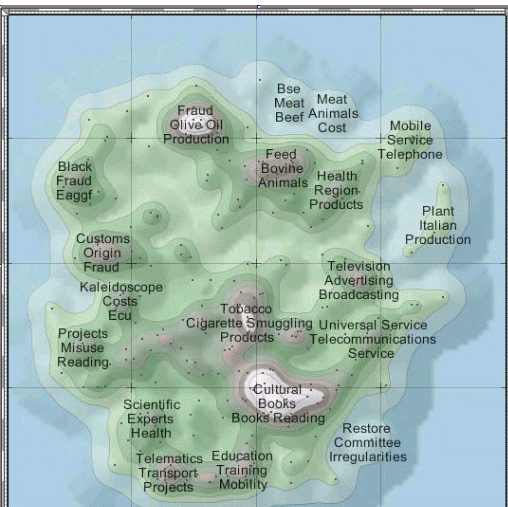
Joint Research Centre



EUROPEAN COMMISSION
DIRECTORATE-GENERAL
Joint Research Centre

Application Areas

- Translation Spotting, e.g. to produce a parallel corpus
- Finding similar documents to a given text, independent of language
- Identification of cross-lingual document plagiarism
- Cross-lingual classification and clustering
- Multilingual document maps



Map produced with ThemeScape, by CARTIA Inc.

Joint Research Centre