### **Scrambling for Metadata**

# Using Topic Modeling and Word2Vec to explore the Archives of the EC



Seth van Hooland, Mathias Coeckelbergs, Ettore Rizza and Simon Hengchen Brussels, September 2017



	LA COMMISSION DES NAUTES EUROPEENNES	Bruxelles , le				
COMMUN		JUR/3023/72 – JA/mh SECRÉTARIAT GÉNÉRAL				
		ARRIVE				
		LE 从卫在1972 8855				
	NOTE à l'attenti	on du Monsieur le Secrétaire Général				
	mutuelle sés dans	tion d'une directive du Conseil concernant l'assistance e en matière de recouvrement des montants indûment ver- s le cadre de la politique agricole commune, des prélè- agricoles et des droits de douane.				
	Réf. : Note GUD	0/8 n° 22959 du 29.11.72				
		55				
		and the second se				
		ce juridique propose d'apporter au projet de directive e de référence les modifications suivantes :				
	annexe a la note	de reierence les modifications suivantes :				
	1. Article	5 \$ 5				
	-	cième phrase un nouvel alinéa et compléter celui-ci par				
1		le texte suivant :				
	· la demande de re le titre permett	loivent intervenir dans un délai d'un mois à compter de ecouvrement. Ils ne peuvent être refusés, dès lors que cant l'exécution est régulier en la forme".				
	2. Article	10				
	A ajouter in fin	le : "Ce refus motivé est communiqué à la Commission".				
	off					
	Ji <sup>St</sup>	Jean Aufland J. AMPHOUX				
		lean of t				
		J. AMPROUX				
		•				
		)				
	c. M. Pingel (GU					
	c. M. Pingel (GU)					
	c. M. Pingel (GU					
	c. M. Pingel (GU					



Quick search: insert free text, CELEX number or descriptors. Use "" for exact matches.

Advanced search

?

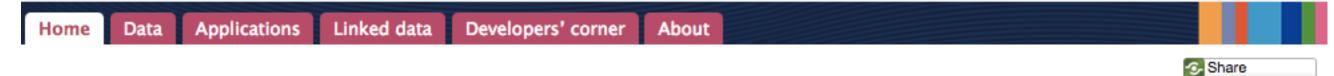
\$

٩

EUROPA	> EU law and	publications > EUR-Lex > Types of docum	nents in EUR-Lex					
Home	Official Jour	nal EU law and related documents	National law	Legislative procedures	More			
Types	of docum	ents in EUR-Lex				ʔ 🝌 🚔 🙀 Share	My EUR-Lex	
	ou can use this table to build CELEX numbers. The Celex number is the unique identifier of each document in EUR-Lex, regardless of language. Read more or Register - or Re							
	Il documents (except from sectors 0 and 7) may have one or more corrigenda. A corrigendum has the same Celex number as the original act, with a suffix R(xx)", where "xx" is the sequential number of the corrigendum.							
Sector	1: Treatie	S						
Sector	Descriptor	Type of document					So Webservice registration	
1	К	Treaty establishing the European Coa	l and Steel Comm	nunity (ECSC Treaty) (1951	)		🖳 Online tutorials	
	A	Treaty establishing the European Ator 2012)	mic Energy Comm	nunity (EURATOM Treaty)	1957); EURATOM Treaty consol	idated versions (2010,	III Statistics	
	E Treaty establishing the European Economic Community (EEC Treaty) (1957); EC Treaty Maastricht consolidated version 1992; EC Treaty Amsterdam consolidated version 1997; EC Treaty Nice consolidated version 2002; EC Treaty Athens consolidated version 2006; Treaty on							
	$\odot$	EU <b>Open Data</b> Po	ortal					

Access to European Union open data

EUROPA > EU Open Data Portal > Home





The European Union Open Data Portal (EU ODP) is your single point of access to open data produced by EU institutions and bodies. All the data you can find via this catalogue are free to use and reuse for commercial or non-commercial purposes.

Show results with: all of these words | 
any of these words | 
the exact phrase

Search for metadata using our SPARQL endpoint query editor or access the application programming

### Eurovoc

### **Content language:** (en) English

٤

▶

### nuclear safety

- Simple search
- Advanced search

### Browse

 Browse the subjectoriented version

### Download

- By domain
- Permuted alphabetical
- Multilingual list
- Alphabetical index
- EuroVoc SKOS/RDF
- EuroVoc XML

### Your proposals

- Contribute
- New approved concepts

UF Euratom inspection Euratom safeguards IAEA inspection nuclear control reactor safety safety of nuclear installations safety of nuclear power stations

### 66 ENERGY

MT 6621 electrical and nuclear industries BT1 nuclear industry NT1 nuclear accident NT1 radiation protection

### NT1 radioactivity

RT ecology movement [0431] health risk [2841] International Atomic Energy Agency [7621] nuclear law [6606] peaceful use of energy [6606] radioactive effluent [5216] radioactive pollution [5216] thermal discharge [5216]

### URI http://eurovoc.europa.eu/4042

### Has Exact Match

Nuclear safety Nuclear safety (ECLAS) Nuclear safety (STW) nuclear safety (UNBIS)

### LANGUAGE EQUIVALENTS

- BG ядрена безопасност
- ES seguridad nuclear
- CS jaderná bezpečnost
- DA nuklear sikkerhed
- DE nukleare Sicherheit
- ET tuumaohutus
- EL πυρηνική ασφάλεια
- EN nuclear safety
- FR sécurité nucléaire
- nuclear safety (under translation) GA
- HR nuklearna sigurnost
- IT sicurezza nucleare
- LV kodoldrošība
- LT branduolinė sauga
- HU nukleáris biztonság
- MT sikurezza nukleari
- NL nucleaire veiligheid
- PL bezpieczeństwo jądrowe
- PT segurança nuclear
- RO securitate nucleară
- SK jadrová bezpečnosť
- SL jedrska varnost
- FI ydinturvallisuus
- SV nukleär säkerhet
- SR нуклеарна безбедност
- MK нуклеарна безбедност
- SQ siguri bërthamore



University Institute	HISTORICAL ARCHIVES DF THE EUROPEAN UNION		٩	Search the database.		۰
		Home	Holdings	Record Creators	Oral History	Search



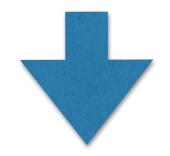
	A-Z Index   Sitemap   About this site   FAQ   What's New   Legal notice   Cookies	Contact   Search English (EN)
	HISTORICAL ARCHIVES	
European Commission		
European Commission > Historica	al Archives > COM files	
What we offer	ARCHISplus: searching COM files	📥 🛕 🏹 😏 Share
Archival fonds		
Collections	COM documents matching this search criteria: 214	Live web support
Availability Rules	Expressions searched : Title contains : (%agriculture%)	
Searching	Modify search	Who are we?
Archival Fonds		
Speeches		"Remembering the past, to build a better future"
Minutes of meetings	Image: Market Marke	
COM files		Links
ECSC inventory	12 Ref 11 Title 11 Date	
Visiting the archives	COM NOTE préparée par le Groupe VI. contenant des	Archives — EU
FAQ	(1958) 23 propositions relatives à l'organisation de la Division de l'Agriculture (Div. VI)	institutions
		Historical Archives of the
	COM         NOTE DE M. MANSHOLT, Président du Groupe         18/06/1958           (1958) 125         Agriculture, relative à l'entretien du 13 juin 1958	European Union in Florence (HAEU)
	avec une délégation danoise, conduite par M. KRAG, Ministre, au sujet de certains problèmes concernant	
	l' <mark>agriculture</mark> .	Archives Portal Europe
	COM NOTE RELATIVE A L'ENTRETIEN DU GROUPE 04/09/1958	Other links
	(1958) 183 AGRICULTURE DE LA COMMISSION AVEC UNE DELEGATION DANOISE CONDUITE PAR Monsieur le	
	Ministre KRAG, AU SUJET DE CERTAINS PROBLEMES	U.E.
	CONCERNANT L'AGRICULTURE BRUXELLES, le 2 Septembre 1958 (présentée par Monsieur MANSHOLT)	
	COM Organigramme de la DIRECTION GENERALE DE 24/09/1958	The Estonian presidency of the Council of the EU:
	(1958) 204 L'AGRICULTURE (adopté par la Commission à sa réunion du 24/9/1958)	<u>1 July - 31 December</u> 2017
	COM PLAN DE TRAVAIL POUR LA DIRECTION GENERALE DE 24/09/1958	

## Corpus

- 24.787 pdf documents, representing I38,3 GB
- Period 1958 1982, with documents in French, Dutch, German, Italian, Danish, English and Greek
- Tombstone metadata
- Conversion to .txt and split per language => 205.370 . txt files, representing 7,4 GB
- 835.717.292 words or 1.671.434 pages

			_	_	-								-	_	
	A	< > D	E	F	G	Н	1	J	к	L	м	N	0	Р	Q
1	Topic Number	Salient tokens													
2		0 amounts	cereals	compensatory	products	rice	wheat	applicable	regulation	amount	eec	refund	falling	weight	referred
3		1 fund	decision	council	assistance	social	article	european	operations	february	training	regulation	persons	application	financial
4		2 article	member	provisions	states	paragraph	commission	council	directive	measures	proposal	laid	referred	state	provided
5		3 commission	council	committee	decision	european	community	member	states	draft	proposal	economic	opinion	concerning	treaty
6		4 securities	quantitative	textiles	authorities	arrangement	products	community	textile	limits	units	zealand	quotation	ciuts	agreement
7		5 directive	products	substances	product	annex	sea	medicinal	foodstuffs	acid	mediterranean	solution	health	laws	content
8		6 products	customs	weight	tariff	fruit	duties	ii	sugar	exceeding	falling	heading	common	content	net
9		7 tax	project	taxable	will	rate	amount	oil	services	taxation	system	supplies	production	excise	equipment
10		8 including	products	heading	articles	machines	parts	apparatus	metal	falling	ii	example	kind	paper	machinery
11		9 product	vehicles	tha	produot	agreement	motor	froa	bureaux	territory	produota	republic	based	including	nr
12		10 contract	credit	contracts	insurer	amount	guarantee	payment	insured	grouping	export	loss	works	bank	commercial
13		11 agreement	community	european	trade	economic	products	republic	negotiations	imports	article	council	cooperation	parties	commission
14		12 contracting	party	parties	agreement	agent	operating	executive	committee	task	agency	will	participants	article	operation
15		13 test	gas	pressure	directive	equipment	water	appliance	type	tests	measuring	inspection	requirements	appliances	conditions
16		14 increase	community	production	countries	prices	year	situation	will	market	world	total	level	price	increased
17		15 description	products	de	member	port	airport	fabrics	destination	european	communities	address	state	handicrafts	tariff
18		16 member	states	state	directive	law	national	undertakings	community	legal	provisions	rights	laws	territory	authorities
19		17 convention	protocol	agreement	article	treaty	council	force	states	commerce	accession	government	provisions	member	international
20		18 holdings	income	farm	labour	ua	ha	family	survey	pigs	land	alu	returning	poultry	paid
21		19 regulation	eec	article	regard	council	commission	community	european	products	amended	states	treaty	applicable	member
22		20 project	subsidy	aid	guidance	agricultural	regulation	eec	fund	granted	will	article	application	eaggf	construction
23		21 products	countries	community	tariff	states	preferences	member	islands	article	imports	quotas	duties	territories	commission
24		22 member	quota	states	community	tariff	share	state	shares	tons	reserve	metric	initial	commission	article
25		23 coal	steel	decision	ecsc	treaty	will	market	undertakings	article	commission	industry	aid	iron	production
26		24 fishing	fish	staff	community	fishery	species	annex	sea	allowance	servants	catch	officials	conditions	ii
27		25 products	originating	protocol	manufacture	certificate	article	goods	movement	customs	community	exceed	product	joint	conditions
28		26 tho	bo	cf	tha	ir	ar	ii	ia	ho	io	en	ti	cr	er
29		27 amounts	meat	compensatory	beef	animals	veal	eggs	amount	regulation	monetary	bovine	eec	weight	poultry
30		28 price	prices	market	year	marketing	intervention	sugar	production	wine	community	fixed	products	common	quality
31		29 aid	milk	food	powder	skimmed	tons	countries	community	metric	programme	commission	supply	council	butteroil
32		30 european	data	projects	industry	programme	community	development	processing	will	aircraft	statistics	study	national	market
33		31 social	workers	employment	training	work	benefits	working	women	health	labour	action	migrant	security	family
34		32 tender	intervention	invitation	butter	tenders	export	agency	decision	market	minimum	levy	article	price	tenderer
	+	EU_en_Pres1-50 v	EU_en_Pres1-1	00 - EU_en_P	Pres2-100 - EU	J_en_Pres2-50 ▼	EU_en_Pres3-	100 - EU_en_F	Pres3-50 🔻						

Human input - topic	EuroVoc URI	Label	Salient tokens		
coal and steel industry aid	http://eurovoc.europa.eu/852	ECSC aid	coal	steel	aid
ree trade	http://eurovoc.europa.eu/3584	trade restriction	tax	trade	duty
aducation	http://eurovoc.europa.eu/668	education	education	training	social
employment policies	http://eurovoc.europa.eu/217190	employment policy	employment	workers	work
energy production	http://eurovoc.europa.eu/2715	energy production	energy	oil	coal
ood trade	http://eurovoc.europa.eu/2446	food policy	gatt	tariff	cheese
agricultural aid	http://eurovoc.europa.eu/2965	aid to agriculture	aid	agricultural	measures



### ECSC aid

### 16 ECONOMICS

MT 1606 economic policy

BT1 EU aid

BT2 economic support

RT ECSC loan [ 1021]

URI http://eurovoc.europa.eu/852

### LANGUAGE EQUIVALENTS

помощ на Европейско обединение за въглища и стомана BG

RDF/XML TURTLE CONTRIBUTE MAP

- ES ayuda CECA
- pomoc Evropského společenství uhlí a oceli CS
- DA EKSF-støtte
- DE EGKS-Beihilfe
- ET ESTÜ abi
- EL ενισχύσεις ΕΚΑΧ
- EN ECSC aid
- FR aide CECA
- ECSC aid (under translation) potpora EZUČ-a GA
- HR
- IT aiuto CECA
- EOTK palīdzība LV
- EAPB pagalba LT
- HU ESZAK-segély
- ghajnuna KEFA MT
- NL EGKS-steun PL
- pomoc EWWS PT auxílio CECA
- RO
- ajutor CECO pomoc ESUO
- SK SL pomoč ESPJ
- FI EHTY:n tuki
- SV EKSG-stöd



### Topic labeling ?

- "More of an art than a science" (Chang et al, 2009)
- Traditionally, two automated approaches :
  - deriving labels from the top tokens of the topics
  - label with concept from a controlled vocabulary

## Topic labeling ?

- Hulpus et al (2013) & Allahyaria and Kochuta (2015) use the graph structure of DBPedia to rank the different label candidates
- But graph structure of DBPedia as a knowledge structure is not terribly coherent ...
- Our approach : use pre-trained Word2Vec to exclude in an iterative manner the "outliers" from the tokens of a topic and match the remaining token with Eurovoc

>>> word\_vectors.doesnt\_match("amounts meat compensatory beef animals veal eggs amount regulation monetary bovine eec weight poultry conditions prices competent imports authorities] ".split()) 'eec' >>> word vectors.doesnt match("amounts meat compensatory beef animals veal eggs amount regulation monetary bovine weight poultry conditions prices competent imports authorities".sp lit()) 'competent' >>> word\_vectors.doesnt\_match("amounts meat compensatory beef animals veal eggs amount regulation monetary bovine weight poultry conditions prices imports authorities".split()) 'compensatory' >>> word\_vectors.doesnt\_match("amounts meat beef animals veal eggs amount regulation monetary bovine weight poultry conditions prices imports authorities".split()) 'regulation' >>> word\_vectors.doesnt\_match("amounts meat beef animals veal eggs amount monetary bovine weight poultry conditions prices imports authorities".split()) 'conditions' >>> word\_vectors.doesnt\_match("amounts meat beef animals veal eggs amount monetary bovine weight poultry prices imports authorities".split()) 'authorities' >>> word\_vectors.doesnt\_match("amounts meat beef animals veal eggs amount monetary bovine weight poultry prices imports".split()) 'weight' >>> word vectors.doesnt match("amounts meat beef animals veal eqgs amount monetary bovine poultry prices imports".split()) 'monetary' >>> word\_vectors.doesnt\_match("amounts meat beef animals veal eggs amount bovine poultry prices imports".split()) 'amount' >>> word vectors.doesnt match("amounts meat beef animals veal eggs bovine poultry prices imports".split()) 'amounts' >>> word\_vectors.doesnt\_match("meat beef animals veal eggs bovine poultry prices imports".split()) 'prices' >>> word vectors.doesnt match("meat beef animals veal eqgs bovine poultry imports".split()) 'imports' >>> word\_vectors.doesnt\_match("meat beef animals veal eggs bovine poultry".split()) 'animals' >>> word\_vectors.doesnt\_match("meat beef veal eggs bovine poultry".split()) 'bovine' >>> word\_vectors.doesnt\_match("meat beef veal eggs poultry".split()) 'eggs' >>> word\_vectors.doesnt\_match("meat beef veal poultry".split()) 'poultry' >>> word\_vectors.doesnt\_match("meat beef veal".split()) 'veal' >>> word\_vectors.doesnt\_match("meat beef".split()) 'beef' >>> word\_vectors.doesnt\_match("meat".split()) 'meat' >>>

Topic label from Word2Vec	Topic Number	Salient tokens								
barley		0 amounts	cereals	compensatory	products	rice	wheat	applicable	regulation	amount
December	4 7	1 fund	decision	council	assistance	social	article	european	operations	february
council			member	provisions	states	paragraph	commission	council	directive	measures
committee			council	committee	decision	european	community	member	states	draft
ciuts/textile			quantitative	textiles	authorities	arrangement	products	community	textile	limits
labelling			products	substances	product	annex	sea	medicinal	foodstuffs	acid
sugar			customs	weight	tariff	fruit	duties	11	sugar	exceeding
tax			project	taxable	will	rate	amount	oil	services	taxation
apparatus			products	heading	articles	machines	parts	apparatus	metal	falling
german/productversions		· •	vehicles	tha	produot	agreement	motor	froa	bureaux	territory
guarantee			credit	contracts	insurer	amount	guarantee	payment	insured	grouping
agreements			community	european	trade	economic	products	republic	negotiations	imports
•				parties			operating	executive	committee	task
party			party		agreement	agent				tests
appliances			gas	pressure	directive	equipment	water	appliance	type will	
prices			community	production	countries	prices	year	situation		market
handicrafts			products	de	member	port	airport	fabrics	destination	european
laws		-	states	state	directive	law	national	undertakings	community	legal
ratification			protocol	agreement	article	treaty	council	force	states	commerce
labour			income	farm	labour	ua	ha	family	survey	pigs
council			eec	article	regard	council	commission	community	european	products
eaggf/project			subsidy	aid	guidance	agricultural	regulation	eec	fund	granted
states			countries	community	tariff	states	preferences	member	islands	article
tariff			quota	states	community	tariff	share	state	shares	tons
ecsc/market			steel	decision	ecsc	treaty	will	market	undertakings	article
fishery			fish	staff	community	fishery	species	annex	sea	allowance
products			originating	protocol	manufacture	certificate	article	goods	movement	customs
1		6 tho	bo	cf	tha	ir	ar	ii	ia	ho
meat	27	7 amounts	meat	compensatory	beef	animals	veal	eggs	amount	regulation
prices			prices	market	year	marketing	intervention	sugar	production	wine
programme/butteroil			milk	food	powder	skimmed	tons	countries	community	metric
programme			data	projects	industry	programme	community	development	processing	will
labour/programme			workers	employment	training	work	benefits	working	women	health
tender		2 tender	intervention	invitation	butter	tenders	export	agency	decision	market
oil		3 energy	oil	community	power	supply	electricity	production	nuclear	stations
import			customs	duties	processing	authorities	import	transit	origin	country
programme			environment	water	protection	environmental	waste	quality	states	action
vehicles			approval	vehicle	directive	eec	vehicles	member	requirements	light
wood			paper	products	community	imports	ceilings	commission	member	wood
appropriations		8 expenditure	financial	budget	appropriations	year	appropriation	staff	article	commission
vehicles			road	traffic	carriage	system	goods	vehicles	regulation	inland
food		0 country	recipient	community	european	agreement	economic	delivery	aid	goods
policies			policy	member	community	monetary	states	policies	measures	will
agriculture			areas	regions	agricultural	development	directive	farming	region	population
prices			market	products	article	undertakings	prices	competition	commission	trade
development			developing	states	cooperation	community	trade		aid	associated
			will	commission		action	problems	general	international	
proposals programme/irc					development			general		policy
programme/jrc			nuclear	materials	project	material	technical	reactor	energy	activities
luxembourg/nederland		7 kingdom	ireland	united	member	italy	states	rate	italian	netherlands
fibre			cotton	fabrics	fibres	man	articles	flax	knitted	textile
companies	49	9 company	board	european	companies	group	employees	article	management	supervisory

### This site is part of Ewropa



Europa > EuroVoc homepage > Domains and MT > barley

Content language: (en) English 🕈	barley	
Simple search  Simple search  Advanced search  Browse Browse the subject- oriented version  Download  By domain Permuted alphabetical Multilingual list Alphabetical index	60 AGRI-FOODSTUFFS MT 6006 plant product BT1 cereals RT malt [6026] URI http://eurovoc.europa.eu/2193 Has Exact Match Barley (AGROVOC) Barley (ECLAS) Barley (STW) barley (UNBIS)	ВG ечемик ES cebada CS ječmen DA byg DE Gerste ET oder EL кріθάρι EN barley FR orge GA barley (under translation) HR ječam IT orzo LV mieži LT miežiai HU árpa
<ul> <li>EuroVoc SKOS/RDF</li> <li>EuroVoc XML</li> <li>Your proposals</li> <li>Contribute</li> <li>New approved concepts</li> </ul>		MT xgħir NL gerst PL jęczmień PT cevada RO orz SK jačmeň SL ječmen FI ohra SV korn SR јечам MK јачмен SQ elb

### This site is part of Ewropa

### EUFOVOC Multilingual Thesaurus of the European Union

### Europa > EuroVoc homepage > Search > council

Content language:		
(en) English	ŧ	
Simple search		
council		►

### Advanced search

### Browse

 Browse the subjectoriented version

### Download

- By domain
- Permuted alphabetical
- Multilingual list
- Alphabetical index
- EuroVoc SKOS/RDF
- EuroVoc XML

### Your proposals

- Contribute
- New approved concepts

Page 1 of 4		
MT 0416 electoral procedure and voting		
election		
local election	UF	council election
MT 0436 executive power and public service		
executive body		
Council of Ministers		
public administration		
local government	UF	county <b>council</b>
local government	UF	town council
MT 0806 international affairs		
international instrument		
European convention	UF	convention of the Council of Europe
MT 1006 EU institutions and European civil service		
EU institution		
Council configuration		
Council of the European Union	UF	Council of European Ministers
Council of the European Union	UF	Council of the European Communities
Council of the European Union	UF	Council of the Union
Council of the European Union	UF	EC Council
Council of the European Union	UF	EU Council
Council of the European Union	UF	European Union Council
EU Council Presidency	UF	EC Council Presidency
EU Council committee	UF	EC Council committee
Ecofin	UF	Ecofin <b>Council</b>
Ecofin	UF	Economic and Financial Affairs Council
European Council		
President of the European Council		

### Other scenario's ?

- Use pre-trained Word2Vec to bring down the amount of terms per topic to for example 3
- Roll out Word2Vec trained on the EC corpus in order to identify the term most closely associated with those 3
- But : number of parameters to configure / play with exponentially increases the overhead of manual evaluation