EuroVoc and thesauri from EU institutions and agencies Interoperability and perspectives for collaborative thesaurus management

Publications Office of the EU
Christine LAABOUDI-SPOIDEN

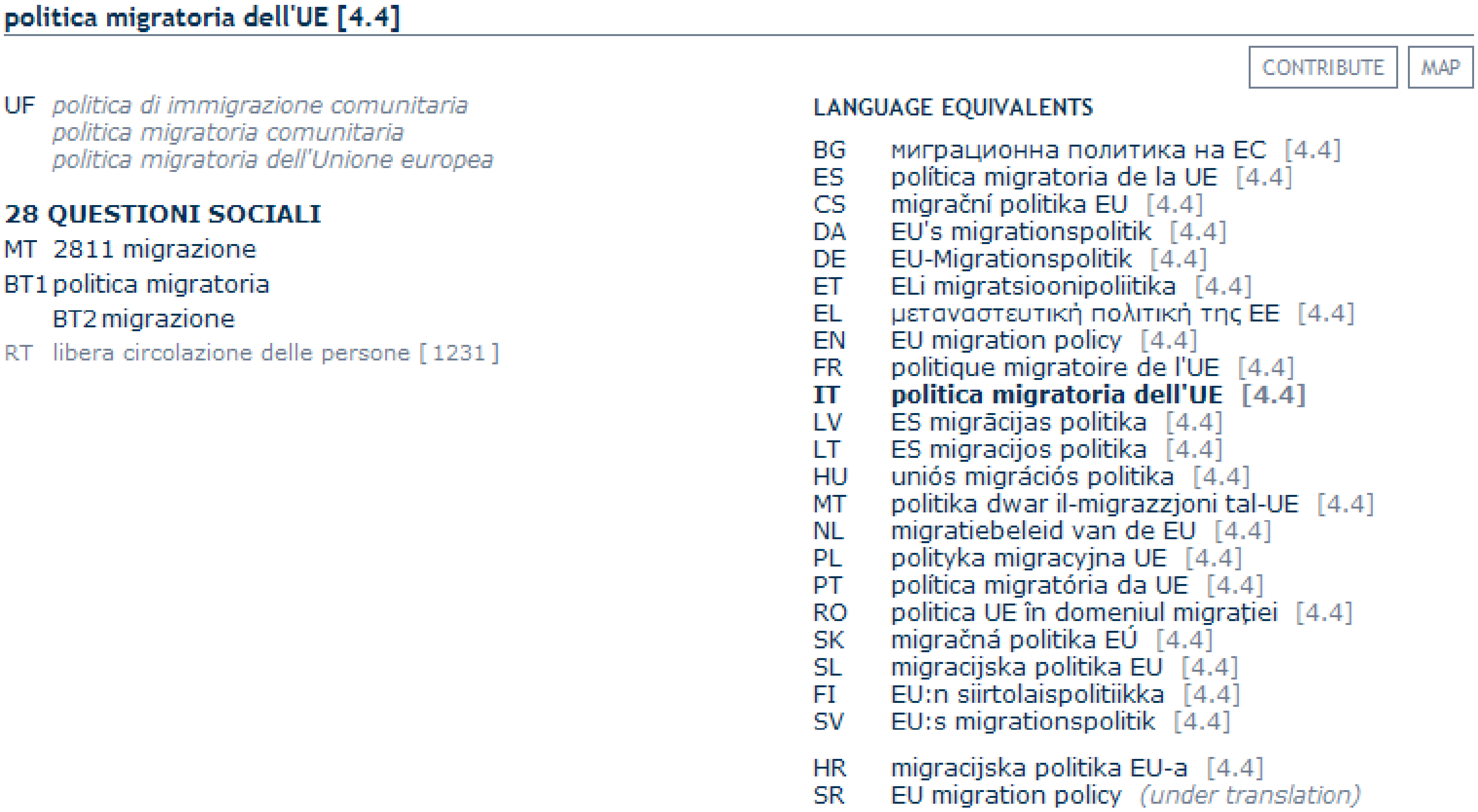
NKOS Workshop, 26 July 2013, Malta

EuroVoc, multilingual thesaurus of the EU

- Maintained by the Publications Office of the EU
- Version 4.4 : 6 800 concepts, 21 Domains, 120 Microthesauri
  - Published on 15 December 2012
- Multilingual
  - 24 languages
  - 23 EU official languages + Serbian
  - Links to other languages
    - Basque, Catalan, Macedonian
- Multidisciplinary thesaurus
  - Strong coverage
    - Parliamentary activities
    - European Union, EU legislation, EU activities, EU policies, EU Institutions
    - EU regions
- Exact equivalence between concepts
  - Symmetry
    - No coverage for national or regional concepts

EuroVoc, multilingual thesaurus of the EU

- A concept represented in every language
  http://eurovoc.europa.eu/6541



politica migratoria dell'UE [4.4]

UF *politica di immigrazione comunitaria*
politica migratoria comunitaria
politica migratoria dell'Unione europea

28 QUESTIONI SOCIALI

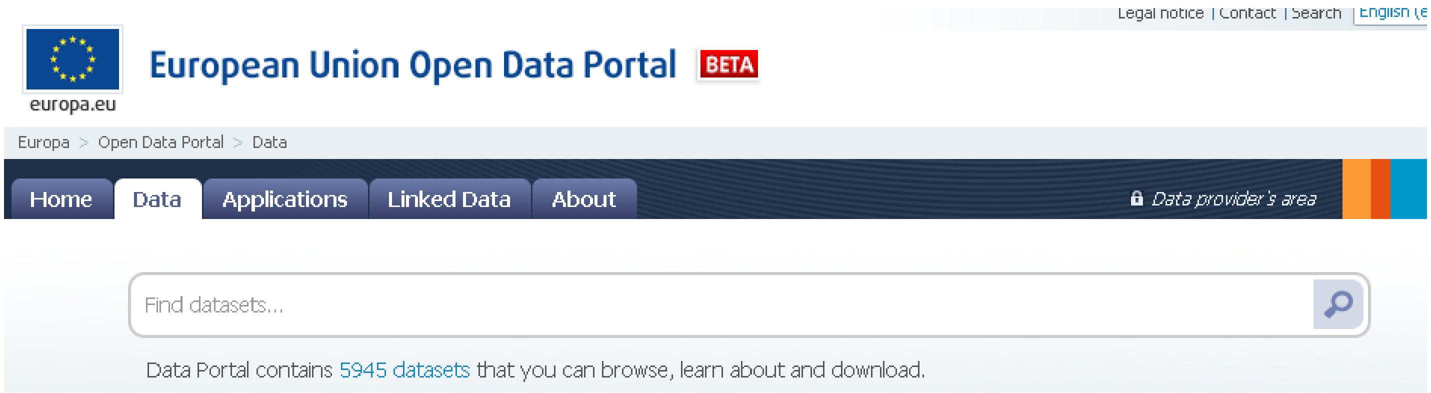
MT 2811 migrazione
BT1 politica migratoria
BT2 migrazione
RT libera circolazione delle persone [1231]

LANGUAGE EQUIVALENTS

BG	миграционна политика на ЕС [4.4]
ES	política migratoria de la UE [4.4]
CS	migrační politika EU [4.4]
DA	EU's migrationspolitik [4.4]
DE	EU-Migrationspolitik [4.4]
ET	ELi migratsioonipoliitika [4.4]
EL	μεταναστευτική πολιτική της ΕΕ [4.4]
EN	EU migration policy [4.4]
FR	politique migratoire de l'UE [4.4]
IT	**politica migratoria dell'UE [4.4]**
LV	ES migrācijas politika [4.4]
LT	ES migracijos politika [4.4]
HU	uniós migrációs politika [4.4]
MT	politika dwar il-migrazzjoni tal-UE [4.4]
NL	migratiebeleid van de EU [4.4]
PL	polityka migracyjna UE [4.4]
PT	política migratória da UE [4.4]
RO	politica UE în domeniul migrației [4.4]
SK	migračná politika EÚ [4.4]
SL	migracijska politika EU [4.4]
FI	EU:n siirtolaispolitiikka [4.4]
SV	EU:s migrationspolitik [4.4]
HR	migracijska politika EU-a [4.4]
SR	EU migration policy *(under translation)*

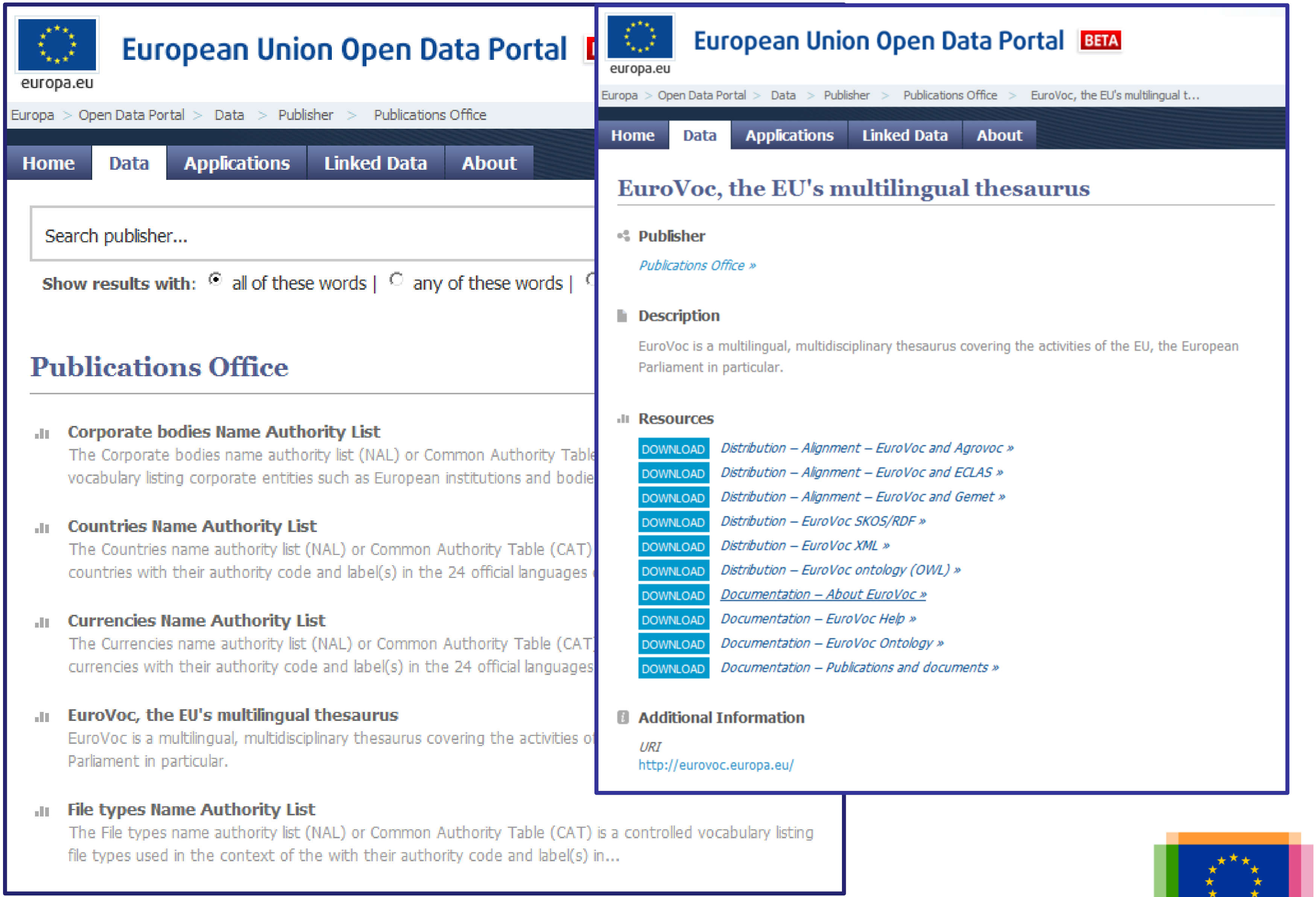
EuroVoc – Distribution

- EU Open Data Portal http://odp/en/
  - Single point of access to metadata produced by the EU institutions and other bodies
  - Free to use, reuse, link and redistribute for commercial or non-commercial purposes.
  - EuroVoc is distributed in RDF or XML


Legal notice | Contact | Search | English (e
europa.eu
European Union Open Data Portal BETA
Europa > Open Data Portal > Data
Home
Data
Applications
Linked Data
About
Data provider's area
Find datasets...

Data Portal contains 5945 datasets that you can browse, learn about and download.

EuroVoc – Distribution



europa.eu
European Union Open Data Portal
Europa > Open Data Portal > Data > Publisher > Publications Office
Home
Data
Applications
Linked Data
About
Search publisher...
Show results with: ● all of these words | ○ any of these words | ○

Publications Office

Corporate bodies Name Authority List
The Corporate bodies name authority list (NAL) or Common Authority Table vocabulary listing corporate entities such as European institutions and bodie

Countries Name Authority List
The Countries name authority list (NAL) or Common Authority Table (CAT) countries with their authority code and label(s) in the 24 official languages

Currencies Name Authority List
The Currencies name authority list (NAL) or Common Authority Table (CAT) currencies with their authority code and label(s) in the 24 official languages

EuroVoc, the EU's multilingual thesaurus
EuroVoc is a multilingual, multidisciplinary thesaurus covering the activities of Parliament in particular.

File types Name Authority List
The File types name authority list (NAL) or Common Authority Table (CAT) is a controlled vocabulary listing file types used in the context of the with their authority code and label(s) in...

europa.eu
European Union Open Data Portal BETA
Europa > Open Data Portal > Data > Publisher > Publications Office > EuroVoc, the EU's multilingual t...
Home
Data
Applications
Linked Data
About

EuroVoc, the EU's multilingual thesaurus

Publisher

Publications Office »

Description

EuroVoc is a multilingual, multidisciplinary thesaurus covering the activities of the EU, the European Parliament in particular.

Resources

- DOWNLOAD *Distribution – Alignment – EuroVoc and Agrovoc »*
- DOWNLOAD *Distribution – Alignment – EuroVoc and ECLAS »*
- DOWNLOAD *Distribution – Alignment – EuroVoc and Gemet »*
- DOWNLOAD *Distribution – EuroVoc SKOS/RDF »*
- DOWNLOAD *Distribution – EuroVoc XML »*
- DOWNLOAD *Distribution – EuroVoc ontology (OWL) »*
- DOWNLOAD *Documentation – About EuroVoc »*
- DOWNLOAD *Documentation – EuroVoc Help »*
- DOWNLOAD *Documentation – EuroVoc Ontology »*
- DOWNLOAD *Documentation – Publications and documents »*

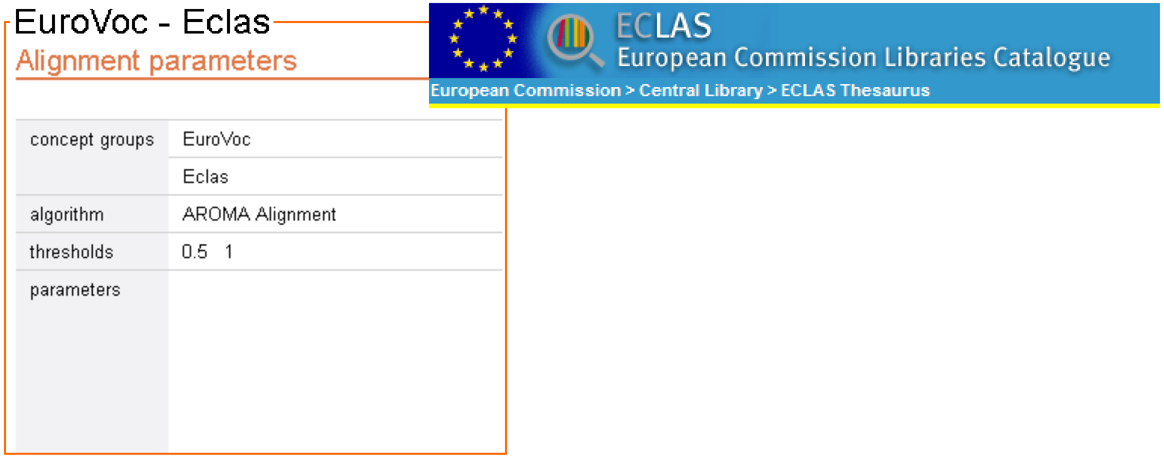
Additional Information

URI
http://eurovoc.europa.eu/

Thesaurus interoperability initiatives

- 2008: Study for a proposed methodology for thesaurus interoperability
  - Feasibility study
    - ITTIG - Institute of Legal Information Theory and Techniques (IT)
  - EuroVoc ← → Eclas, Gemet, Unesco, ETT

- 2010 : Development of a Thesaurus Alignment Tool
  - TAE - Thesaurus Alignment Environment
  - Developed under a contract of the Publications Office
  - Interface for automated mapping and manual validation



EuroVoc - Eclas

Alignment parameters

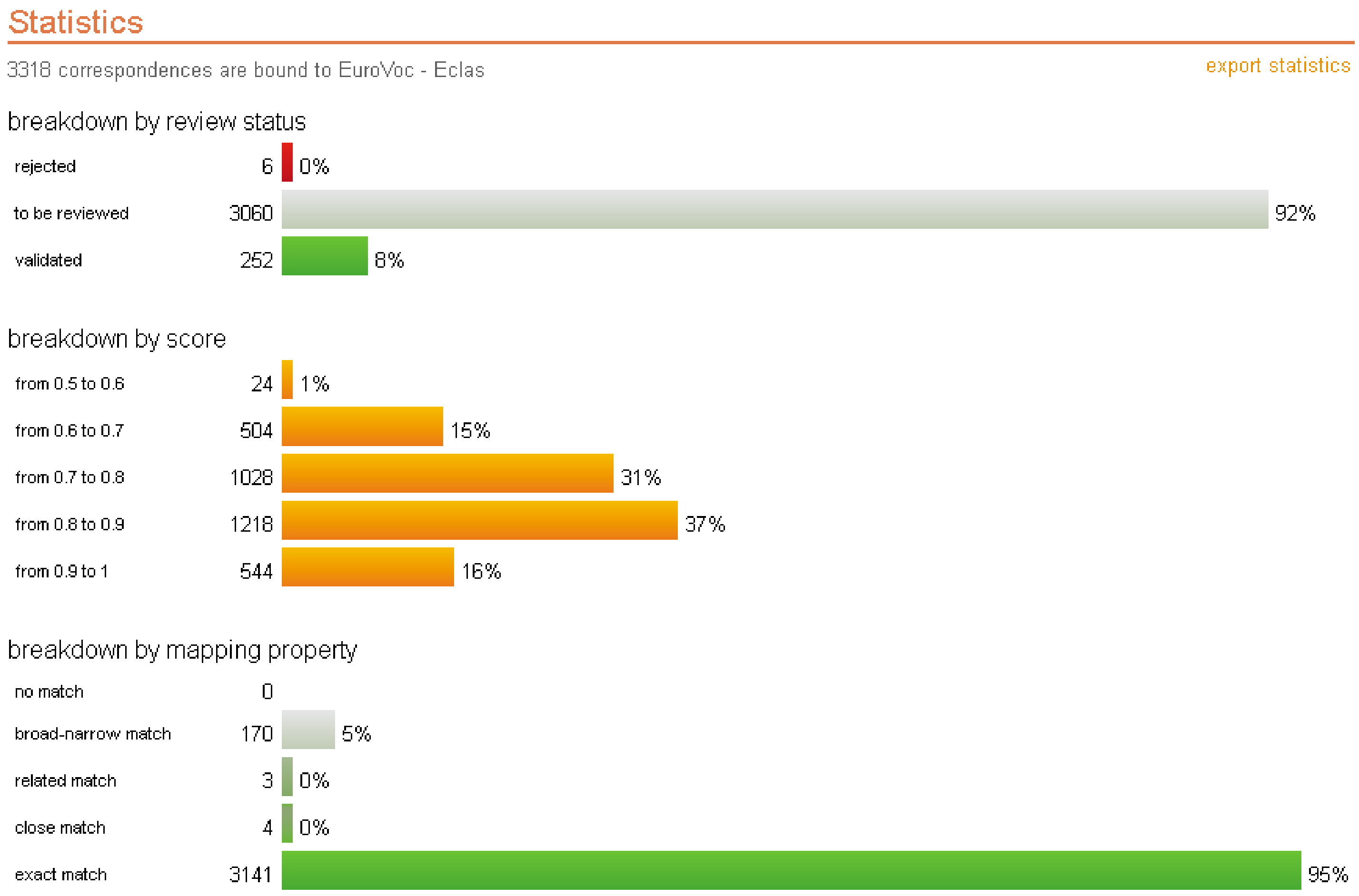
concept groups	EuroVoc
	Eclas
algorithm	AROMA Alignment
thresholds	0.5 1
parameters	

ECLAS
European Commission Libraries Catalogue
European Commission > Central Library > ECLAS Thesaurus

TAE - Thesaurus Alignment Environment (1/4)

ALIGNMENT

Number of correspondences between two thesauri



Statistics

3318 correspondences are bound to EuroVoc - Eclas

export statistics

breakdown by review status

rejected	6	0%
to be reviewed	3060	92%
validated	252	8%

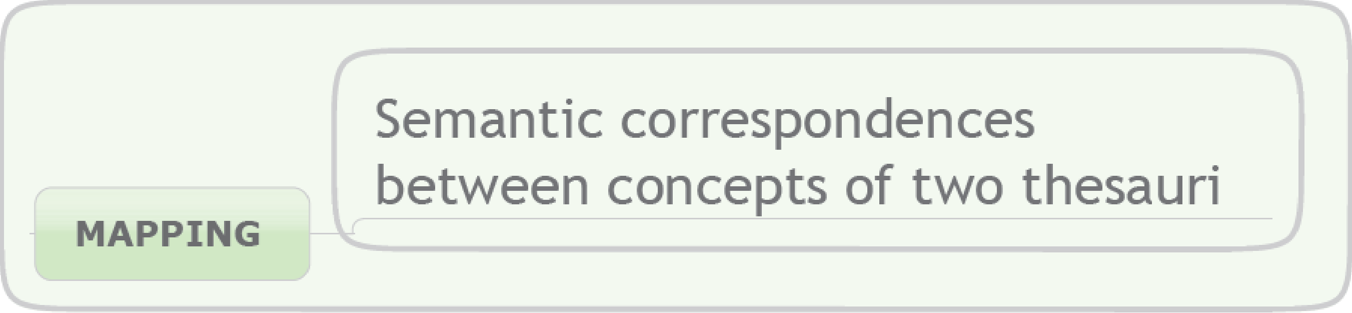
breakdown by score

from 0.5 to 0.6	24	1%
from 0.6 to 0.7	504	15%
from 0.7 to 0.8	1028	31%
from 0.8 to 0.9	1218	37%
from 0.9 to 1	544	16%

breakdown by mapping property

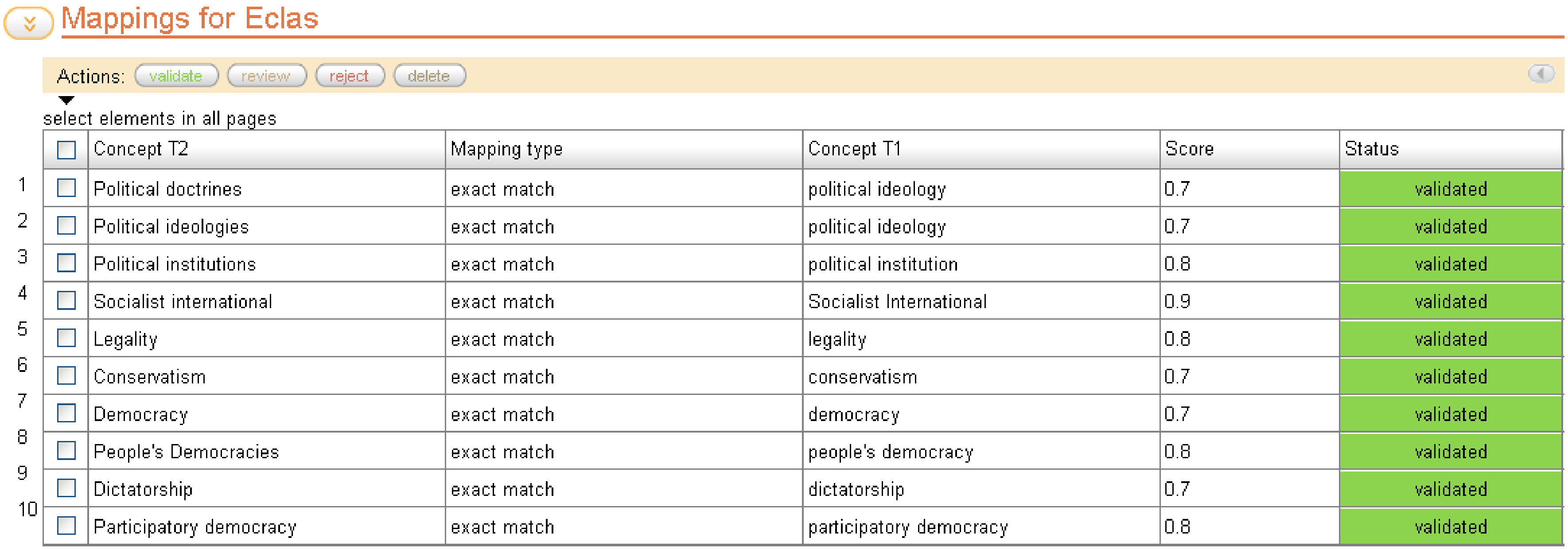
no match	0	
broad-narrow match	170	5%
related match	3	0%
close match	4	0%
exact match	3141	95%

TAE - Thesaurus Alignment Environment (2/4)



MAPPING

Semantic correspondences between concepts of two thesauri



Mappings for Eclas

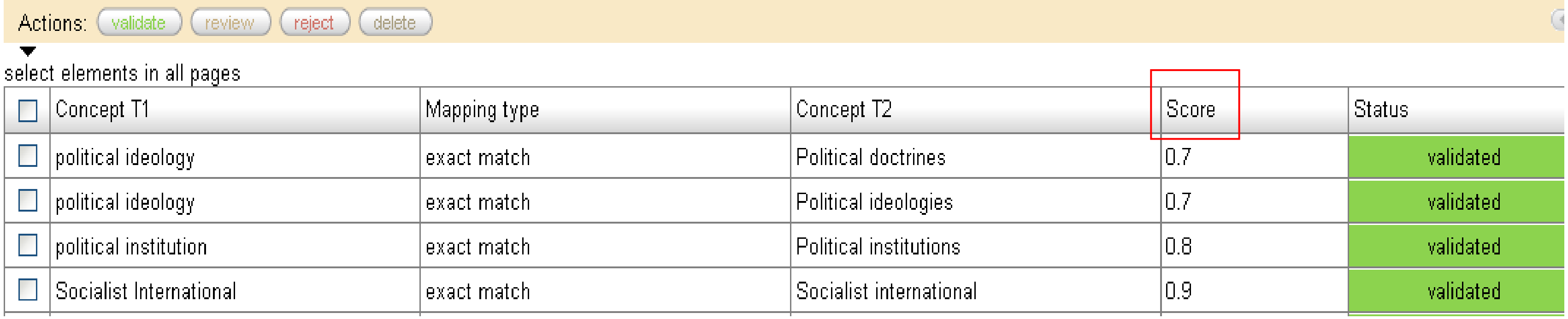
select elements in all pages

	□	Concept T2	Mapping type	Concept T1	Score	Status
1	□	Political doctrines	exact match	political ideology	0.7	validated
2	□	Political ideologies	exact match	political ideology	0.7	validated
3	□	Political institutions	exact match	political institution	0.8	validated
4	□	Socialist international	exact match	Socialist International	0.9	validated
5	□	Legality	exact match	legality	0.8	validated
6	□	Conservatism	exact match	conservatism	0.7	validated
7	□	Democracy	exact match	democracy	0.7	validated
8	□	People's Democracies	exact match	people's democracy	0.8	validated
9	□	Dictatorship	exact match	dictatorship	0.7	validated
10	□	Participatory democracy	exact match	participatory democracy	0.8	validated

TAE - Thesaurus Alignment Environment (3/4)

Automated alignment

- Algorithm: Aroma
- No pivot language
  - Alignment of concepts (URIs) instead of terms
- Score
  - Assigned according to the number of matching lexical values
  - At least 3 matching values between any languages



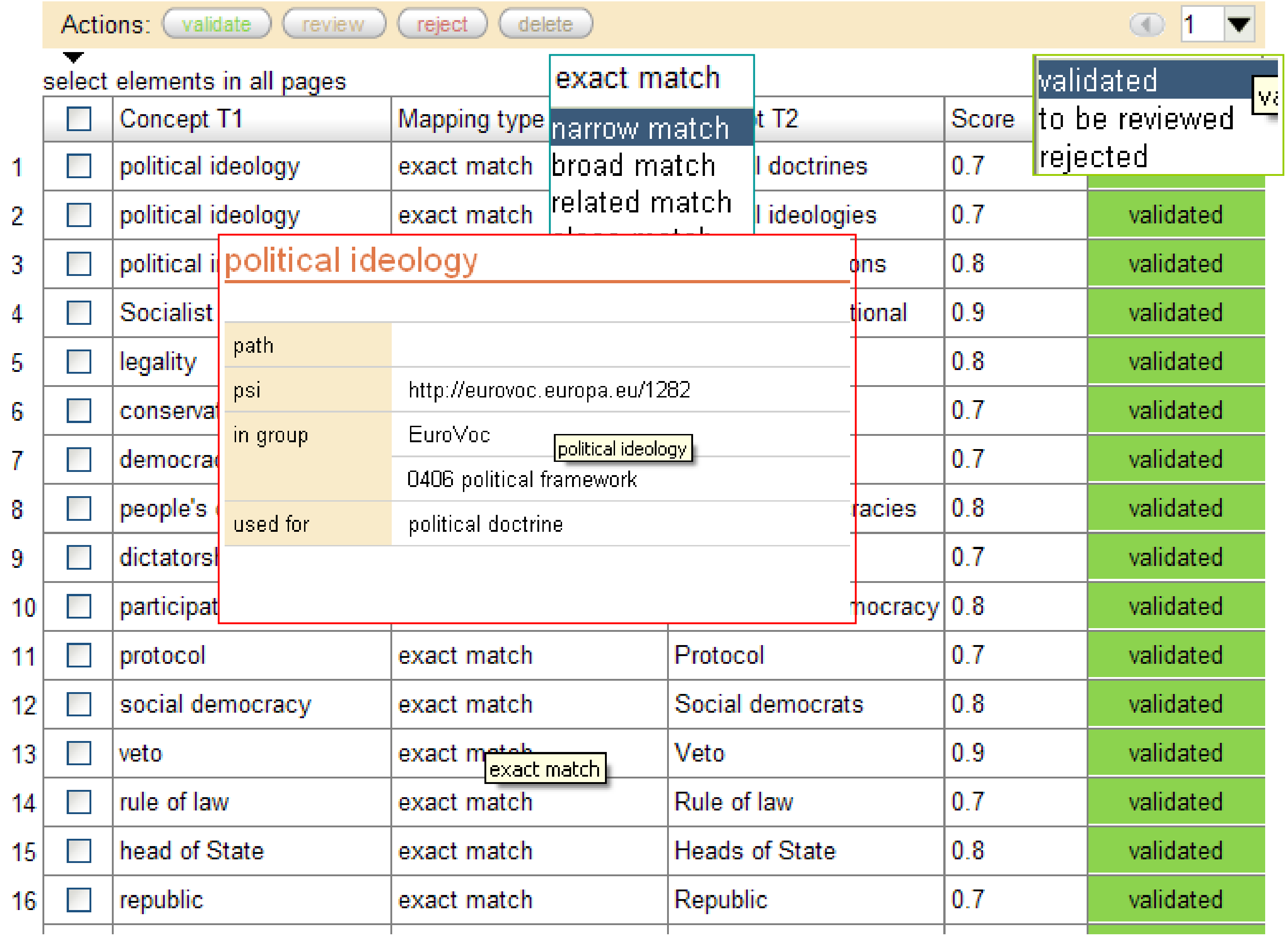
□	Concept T1	Mapping type	Concept T2	Score	Status
□	political ideology	exact match	Political doctrines	0.7	validated
□	political ideology	exact match	Political ideologies	0.7	validated
□	political institution	exact match	Political institutions	0.8	validated
□	Socialist International	exact match	Socialist international	0.9	validated

- Export the final Results
  - SKOS-Core (RDF)
  - EDDOAL

TAE - Thesaurus Alignment Environment (4/4)

Human validation

- Status: validated, to be reviewed, rejected



Actions: validate review reject delete
1
select elements in all pages

	□	Concept T1	Mapping type	t T2	Score	
1	□	political ideology	exact match	l doctrines	0.7	
2	□	political ideology	exact match	l ideologies	0.7	validated
3	□	political i		ons	0.8	validated
4	□	Socialist		tional	0.9	validated
5	□	legality			0.8	validated
6	□	conserva			0.7	validated
7	□	democra			0.7	validated
8	□	people's		racies	0.8	validated
9	□	dictators			0.7	validated
10	□	participat		nocracy	0.8	validated
11	□	protocol	exact match	Protocol	0.7	validated
12	□	social democracy	exact match	Social democrats	0.8	validated
13	□	veto	exact match	Veto	0.9	validated
14	□	rule of law	exact match	Rule of law	0.7	validated
15	□	head of State	exact match	Heads of State	0.8	validated
16	□	republic	exact match	Republic	0.7	validated

political ideology

path	
psi	http://eurovoc.europa.eu/1282
in group	EuroVoc
	0406 political framework
used for	political doctrine

Thesaurus Alignment (Onagui)

- Ontology Alignment GUI
  - Open-source software

- Pros
  - Compilation of alignments achieved by
    - By language
    - Methods: I-subdistance, Levenstein, Exact
  - Export in SKOS, RDF, CSV

- Cons - Human validation
  - Not user-friendly
  - Import in TAE for validation

Thesaurus alignment – Constraints (1/2)

- Datasets must be uploaded as SKOS-Core
    - Conversion in SKOS-Core (from SKOS-XL)
- Vocabularies must be uploaded at each new releases
- TAE = a dedicated installation for alignment
    - No gateway to the Thesaurus Maintenance System
        - SKOS-XL

Thesaurus alignment – Constraints (2/2)

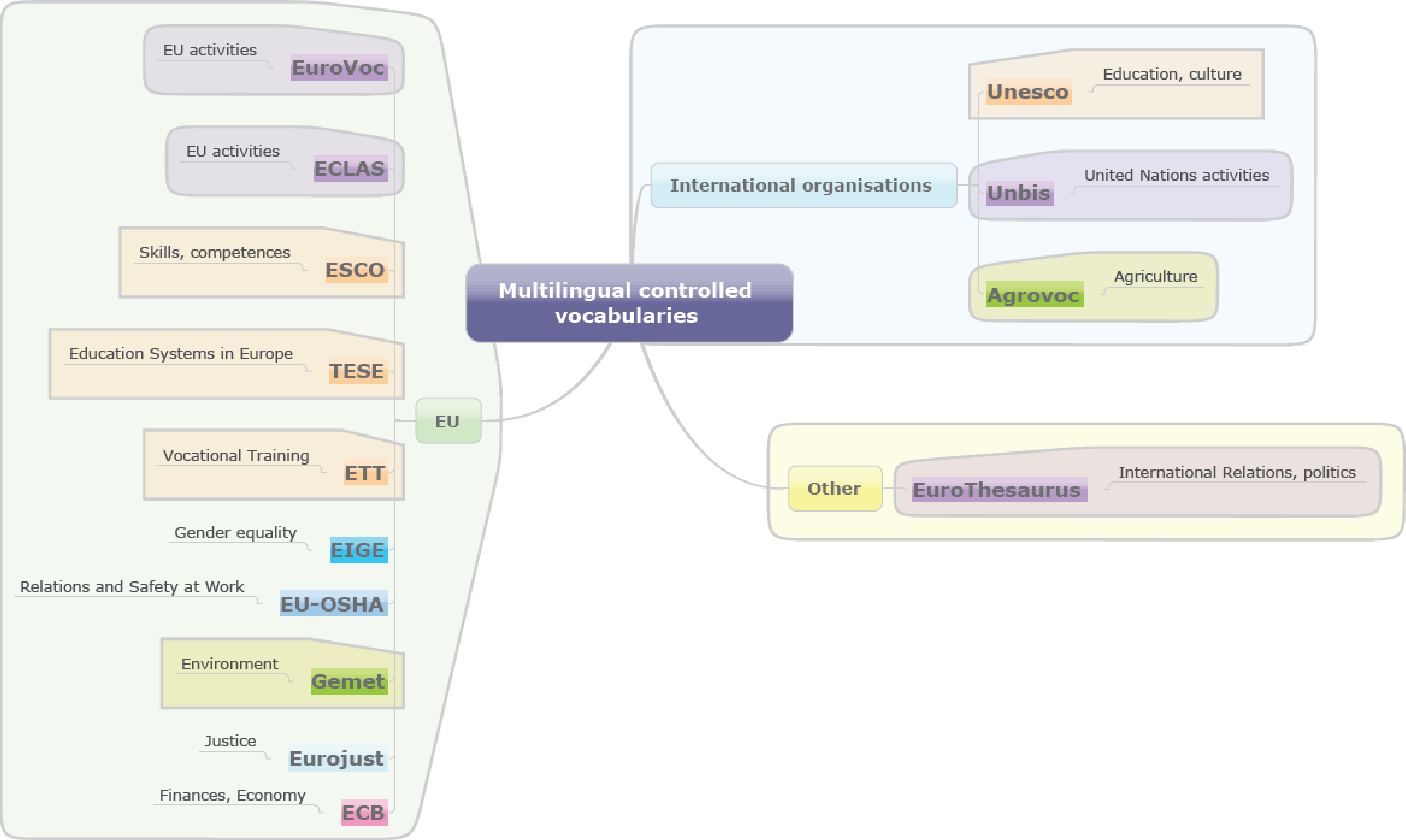
- Published URI’s
  - Resources available as LOD
    - Agrovoc (FAO - UN)
    - Gemet (General Multilingual Environmental Thesaurus - EEA)
- Published URL's – Not dereferencable
    - ECLAS (European Commission’s Library)
    - UNBIS (UN Thesaurus)
    - EuroVoc
- Published (PDF, HTML, Excel)
  - No URI's
    - ETT - European Training Thesaurus (Cedefop)
    - TESE – Thesaurus of education systems in Europe
- Not Published
    - EIGE (Agency for gender equality)
    - Eurojust (EU judicial cooperation unit)
    - OSHA – European agency for health at work

Thesaurus Alignment - Results

Thesauri	Source	Total number of Concepts in TAE	ExactMatches with EuroVoc	% shared concepts with EuroVoc
TESE	European Education Executive Agency	1.387	146	11
OSHA	Agency for Health and Security at Work	1.730	299	18
Gemet	European Environment Agency	1.723	1685	98
ECLAS	EC Central Library	6.400	2751	43
ETT	Cedefop - Agency for Vocational Training	1.566	939	60
Agrovoc	FAO Thesaurus	19.702	1318	7
UNBIS	United Nations Thesaurus	938	214	23

11 % of TESE concepts are matching with EuroVoc

A network of multilingual thesauri (1/2)



Multilingual controlled vocabularies

- EU
  - EuroVoc: EU activities
  - ECLAS: EU activities
  - ESCO: Skills, competences
  - TESE: Education Systems in Europe
  - ETT: Vocational Training
  - EIGE: Gender equality
  - EU-OSHA: Relations and Safety at Work
  - Gemet: Environment
  - Eurojust: Justice
  - ECB: Finances, Economy
- International organisations
  - Unesco: Education, culture
  - Unbis: United Nations activities
  - Agrovoc: Agriculture
- Other
  - EuroThesaurus: International Relations, politics

A network of multilingual Thesauri (2/2)

- Maintained by
  - EU institutions and bodies
  - International organisations
  - National or governmental bodies
- Coverage
  - Different levels of granularity
    - Generic ←→ specialised
  - Overlapping
    - Domains, concepts
    - Language coverage
- Dissemination
  - Individual publication
  - Different formats (SKOS/RDF, HTML, PDF, Excel)
  - Different presentations
    - Hierarchy
    - Alphabetical list of terms
  - Online vs. Not published (PDF, XLS)

Thesaurus Alignment – Lessons learnt

- We maintain
  - Similar or close concepts
  - Similar translations

- We use
  - Different maintenance systems
    - Maintenance, hosting or licensing fees
  - Different formats and standards (data model)
  - Different means of dissemination

- Our concepts and metadata are not interlinked (LOD)

- Duplicated efforts and resources in terms of thesaurus maintenance, infrastructure and dissemination

EU Thesaurus Working Group

- Discussion forum between multilingual thesaurus managers
  - EU multilingual thesauri
    - EU institutions (ECLAS, EuroVoc)
    - EU agencies (Cedefop, Eurydice Tese, Gemet, ESCO, EU-OSHA)
  - International multilingual thesauri
    - UNBIS, FAO, UNESCO
  - National multilingual thesaurus
    - Eurothesaurus - European Thesaurus on International Relations and Area Studies (Stiftung Wissenschaft und Politik)

- First meeting in June 2010
  - Publications Office, Luxembourg
  - Informal bilateral discussions

- Raise the need of a future thesaurus collaboration
  - Optimise our resources

EuroVoc granularity

- DCAT Application Profile for European Data Portals
  - Description of EU public sector datasets
  - Enables cross-data portal search for data sets available at national level

- EuroVoc selected for the categorization of datasets
  - EuroVoc is a generic thesaurus, not enough specialized

- Requirements
  - Expand the coverage of EuroVoc with specialized concepts
  - Create synergies with EU multilingual vocabularies
    - Integration of specialised concepts in EuroVoc
    - Integration of new languages
  - Development of new specialised microthesauri

Collaborative Thesaurus Environment (1/3)

Role of the Publications Office

- Offer a Thesaurus Management Platform
  - Selected platform: Vocbench 2.0
    - Opened to EU institutions and bodies
    - Collaborative maintenance for similar concepts and their translations
- Offer a dissemination platform
  - EuroVoc Front-Office, OP Common Portal
- Responsible for hosting, technical maintenance and developments

Why Vocbench?

- Open-Source software developed by FAO
- SKOS is the native storage format
- No licensing fees
- Modular architecture and flexibility
  - Develop extensions without modifying the core system
- Vocbench community
  - Contribution to open source development

Collaborative Thesaurus Environment (2/3)

- Generic and EU concepts maintained at the EuroVoc level
  - EU treaties, EU institutions, EU policies, EU activities
  - Type of documents, Countries, organisations

- Specific concepts maintained in EuroVoc, at the level of the specialised thesauri
  - Expand the coverage below EuroVoc Top Terms
  - Create new Microthesauri
    - Example: Vocational training, gender equality, occupational safety

- Translations achieved by
  - DGT (General Direction for Translations, EC) for EuroVoc
  - CDT (Translation Centre of the EU) for the EU agencies
  - Potential collaboration between the stakeholders
    - Translation network

Collaborative Thesaurus Environment (3/3)

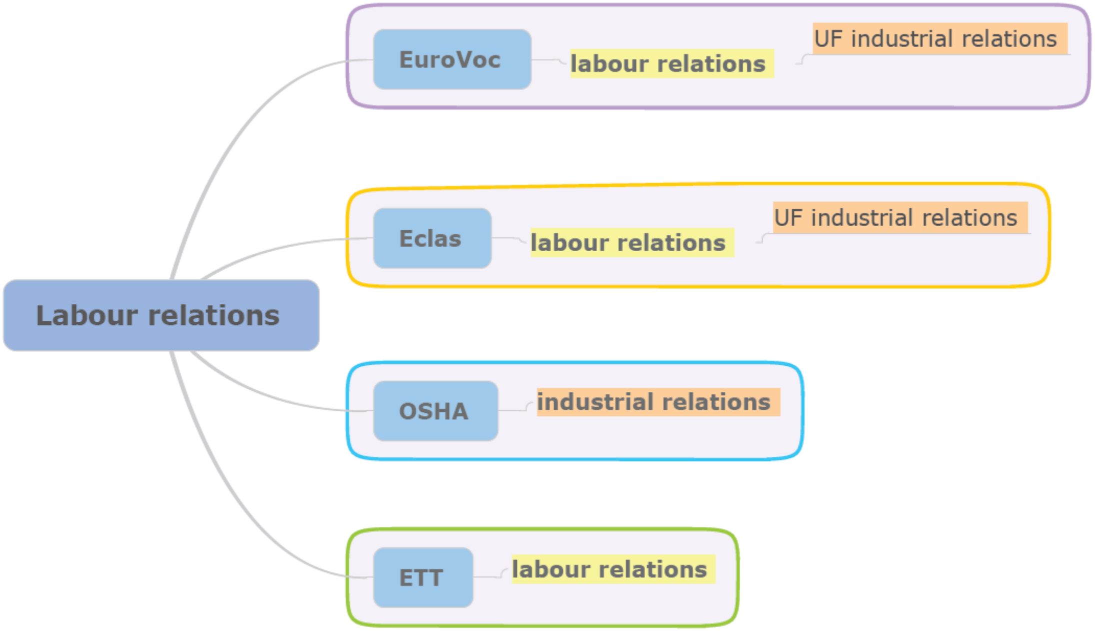
- Several thesauri uploaded in Vocbench
  - One concept might belong to different conceptSchemes (1,*)

- Two categories of controlled vocabularies
  - Vocabulary with no namespace and no concept's URI
  - Vocabulary with a namespace and concept's URI

ConceptSchemes with no namespace (1/2)

- ConceptSchemes with no namespace/URI
  - New URI
  - For example http://data.europa.eu/

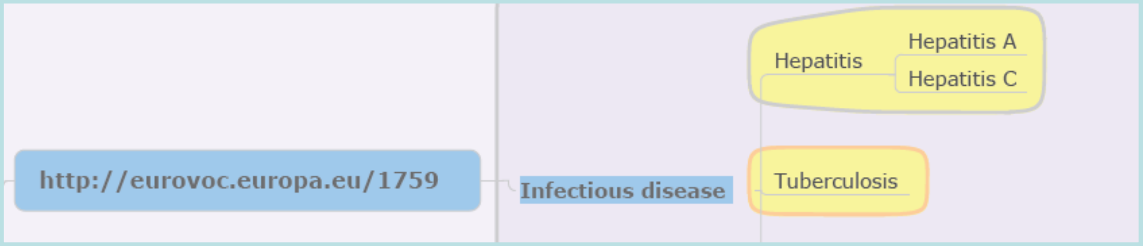
- Concepts identified as *exactMatch* in TAE
  - Taken over by Eurovoc



ConceptSchemes with no namespace (2/2)

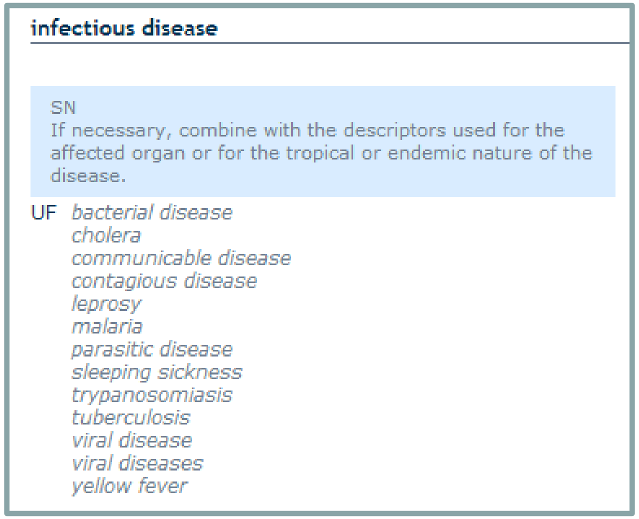
Concepts identified as *narrowMatch* in TAE

- Integrated in EuroVoc
- Assigned to an EuroVoc Top Term or in a new Microthesaurus



Terminology review

- altLabel promotes as a new specific concept



infectious disease

SN
If necessary, combine with the descriptors used for the affected organ or for the tropical or endemic nature of the disease.

UF *bacterial disease*
cholera
communicable disease
contagious disease
leprosy
malaria
parasitic disease
sleeping sickness
trypanosomiasis
tuberculosis
viral disease
viral diseases
yellow fever

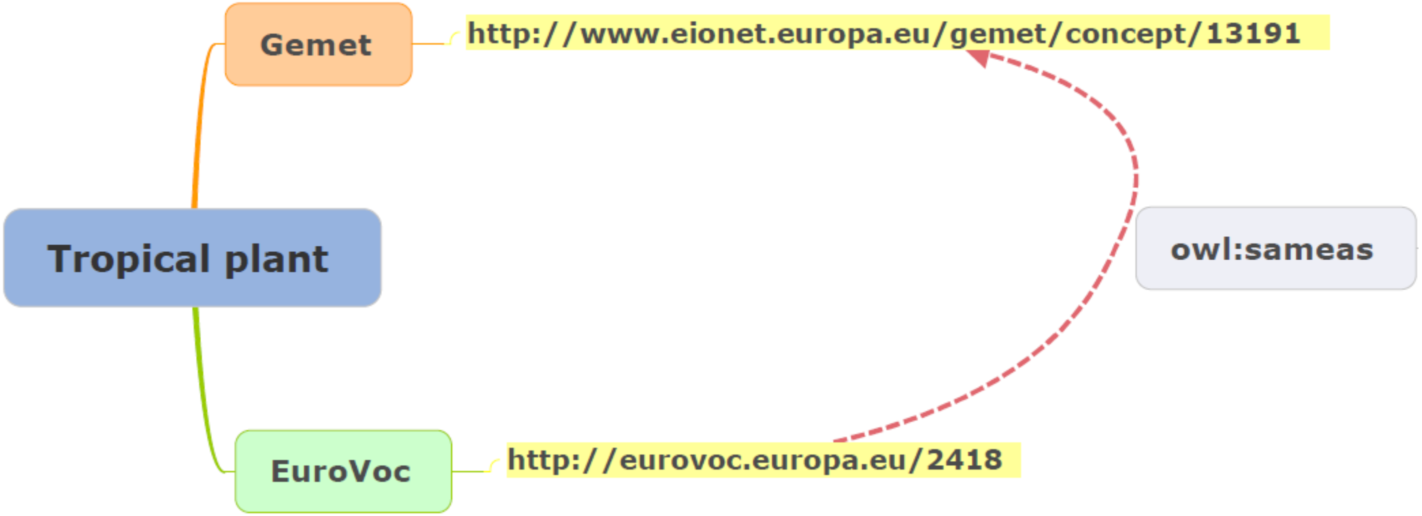
Collaborative management and terminology review

- prefLabel and altLabel
- Concepts identified as *closeMatch* in TAE
  - political doctrine – political ideology
  - Nationality – citizenship
- Plural/singular forms
- Notes (scope notes, definition, history notes)
- Concepts identified as *broadMatch* in TAE
  - Lack of granularity in the specialised thesauri
  - EuroVoc has a BroadMatch in a specialised thesaurus
  - Forest ranger "hasBroadMatch" forestry worker
- Long-term process

ConceptSchemes with a namespace (1/2)

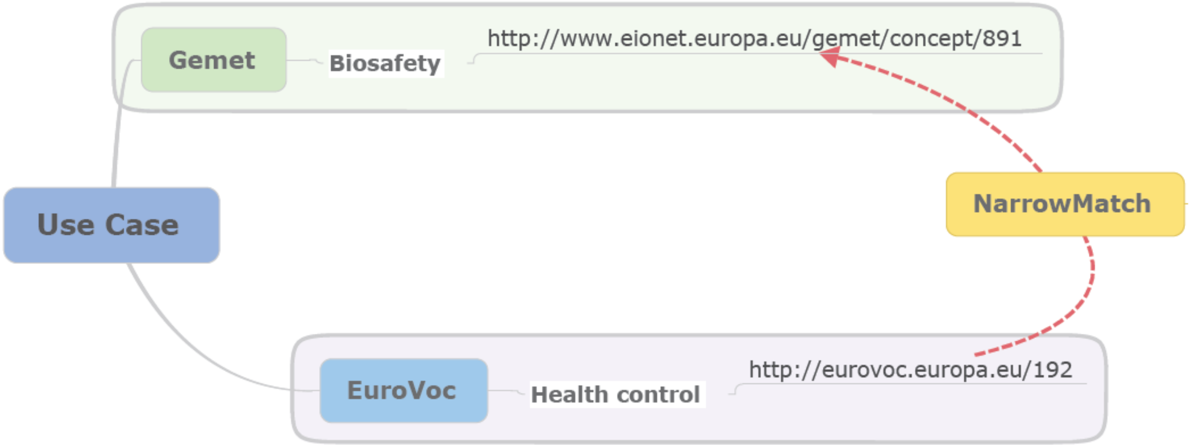
- ConceptSchemes with current namespace/URI
  - Upload in Vocbench for maintenance purpose

- Concepts identified as *exactMatch* in TAE
  - One physical concept (maintenance purpose and translation)
  - Discussion about a EU unique URI http://data.europa.eu/


Gemet
http://www.eionet.europa.eu/gemet/concept/13191
Tropical plant
owl:sameas
EuroVoc
http://eurovoc.europa.eu/2418

ConceptSchemes with a namespace (2/2)

- Concepts identified as *narrowMatch* in TAE
  - Linked by a *narrowMatch* association


Gemet
Biosafety
http://www.eionet.europa.eu/gemet/concept/891
Use Case
NarrowMatch
EuroVoc
Health control
http://eurovoc.europa.eu/192

- Dissemination
  - Individual export by conceptScheme and relevant namespace
    - Automated generation of SKOS mapping between ConceptScheme at export
  - Publish in the PO Common Portal as SKOS Mapping
    - New search option: LOD Search engine

Collaborative Thesaurus Environment - Phases

- Phase 1: Install Vocbench 2.0 in the PO (End 2013)
  - Upload EuroVoc and the EU multilingual thesauri in Vocbench
- Phase 2: Vocbench developments
  - Collaborative maintenance module
  - Customized exports
- Phase 3: Collaborative maintenance
  - Integration of specific terms from the specialised thesauri
  - Generate Linked Data (*narrowMatch*, *exactMatch*)
  - Maintenance of the candidates (new proposed concepts)
- Phase 4: Merge the *exactMatches*
  - Terminology review

Thesaurus collaboration – Ongoing work

- Alignment EuroVoc – ETT (Cedefop), Eclas and OSHA
  - Identify the exactMatch to be shared
  - Identify the narrowMatch to integrate
  - Development of a new “vocational training” microthesaurus

- Alignment EuroVoc - Unbis (United Nations Thesaurus)
  - Practical use
    - Cross-retrieval in EU and UN documentary collections
    - Expand the language coverage of EuroVoc
      - Arabian, Russian, Chinese

- Other practical applications
  - In-depth indexing of EU documentary collection
  - Double indexing and cross-information retrieval
    - EU-Bookshop, EUR-Lex, Eclas

Thanks for your attention

Questions ?

For more information, please contact :

eurovoc@publications.europa.eu