eENVplus

State of play of Environmental Thesauri in the Web and their adherence to (Open) Linked Data Best Practices

M. De Martino,R.Albertoni, P.Podestà

CNR- IMATI

INSPIRE2014

Aalborg, June 16-20 2014

Summary

- Overview
  - Objectives
  - Motivation
- SoP Approach
  - Terminological Resources Cataloguing
  - Reusability Criteria Identification
  - Evaluation of the catalogue
- Conclusions
  - Consideration and Recommendation
  - Conclusion and Future Activity

Overview

General Objective

Analysis of the current state of play of the environment thesauri available on the Web and the assessment of their reusability according with a priori defined criteria.

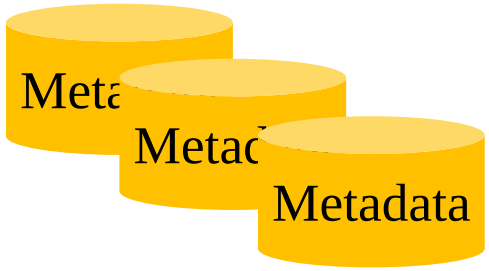
Reusability
«Easiness to access and to exploit Thesaurus content"

Licence Type
- Openness of licence

LD Compliance
- 5 star LD
- Stressing dereferenceable HTTP URIs as identiers for resources

Why Thesauri

Thesauri are employed as solution to the multilingual and multicultural issues in the environmental data sharing



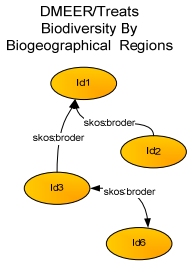
Uniformity in Data description

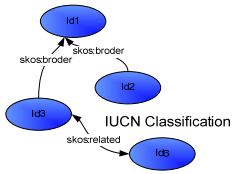
Information discovery across applications and platforms

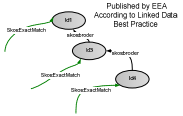
INSPIRE Implementation rules
recommend the adoption of (multilingual) thesauri when compiling metadata for data/services

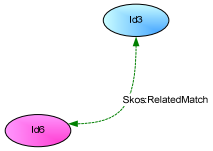
Overview
INSPIRE SDI vs thesauri

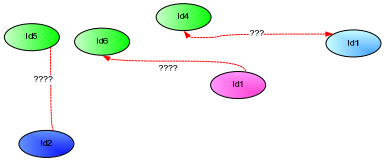
Different thesauri have been developed, and may be deployed for cataloguing the geographical, e.g.,











EARTh *GEMET* *THiST* …

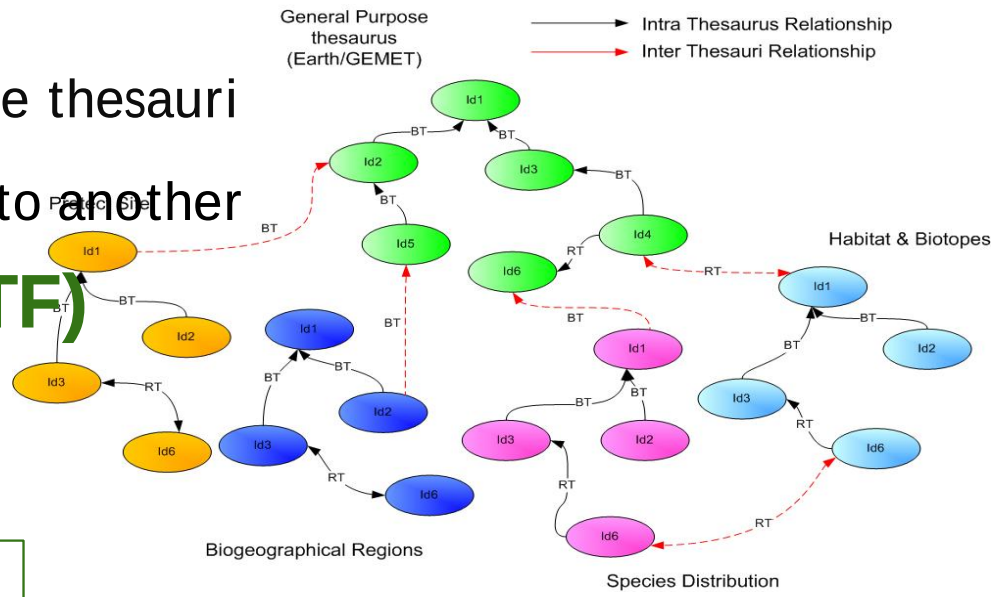
Thesauri heterogeneity wrt thematic coverage, multilingualism, granularities, popularity in certain communities

Heterogeneity is precious!!!

Need of common thesaurus framework to exploit thesauri heterogeneity

Motivation: NatureSDI and eENVplus

- Not only one thesaurus … But
- integration of different available thesauri
- cross walking from a thesaurus to another



Thesaurus Framework(TF)

Modularity

To pl T

Openness

To kee

Interlinking

Li to or th

Exploitability

To encode in a standard and flexible format in order to encourage the adoption and its enrichment from third party system

Design Principle

Simple Knowledge Organization System (SKOS) to encode the thesaurus content

Linked Data best practices to publish the thesaurus in machine understandable format

LusTRE: Linked Thesaurus fRamework for Environment

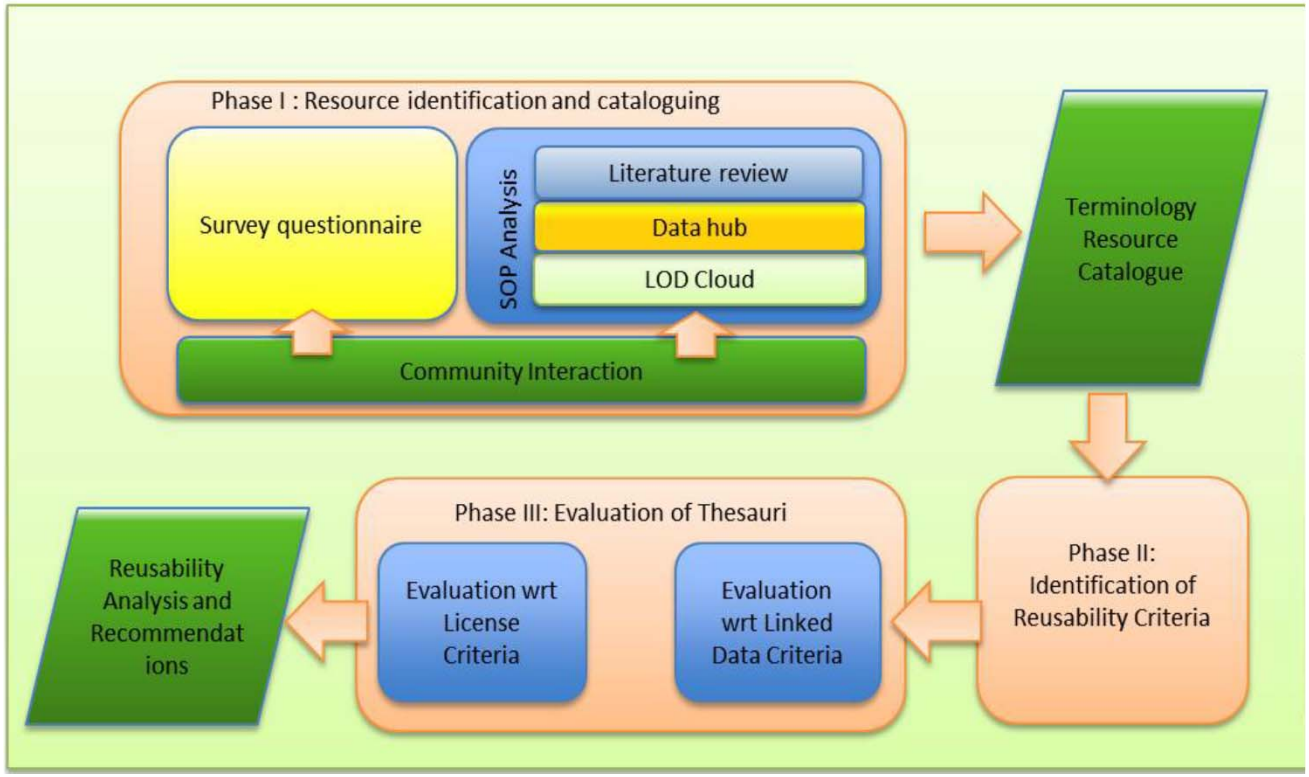
http://linkeddata.ge.imati.cnr.it:2020/

State of Play Approach

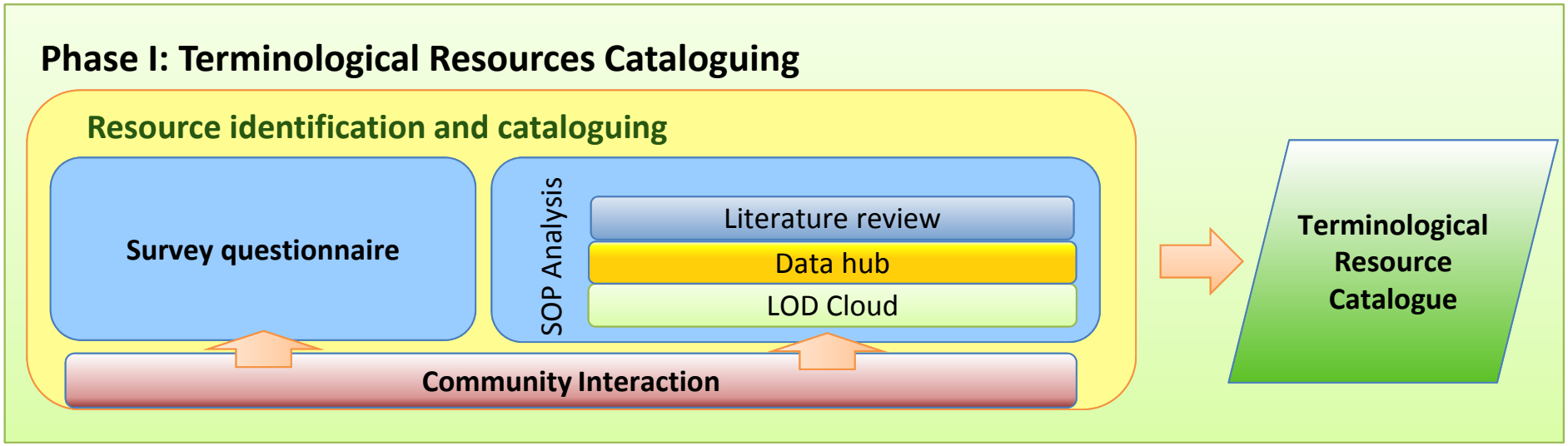
Approach overview

❑ Approach and Outcomes

- ❑ Phase I: **Terminological Resources Cataloguing**: **live Catalogue**
- ❑ Phase II: **Identification reusability criteria**
- ❑ Phase II: **Evaluation of the catalogue** : **Reusability analysis**



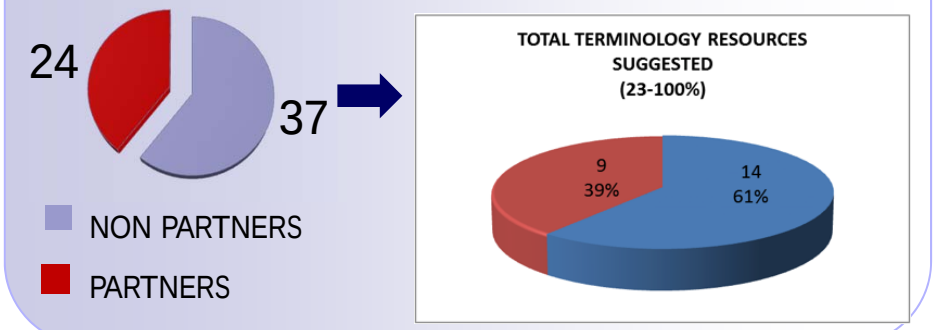
Approach
Teminological Resources Cataloguing



Questionnaire

N Answers (61 -100%)

N suggested Terminology (23 -100%)



SoP

Literature review

- Scientific international journals (i.e. SWJ)

Data hub

- Resource associated to the keywords "thesaurus skos".
- Thesauri for Environment, Geology, GI

LOD Cloud

- resources in the data hub and included in the LOD Cloud datasets published by the LOD from 2007-2011

Approach
Synthesis of Resources Catalogue

Thesauri 30

Thesaurus	QUESTIONNAIRE	Datahub	LITERATURE review	LOD Cloud	Community suggestions	TF from NATURESDIPlus
ADL FTT					X	
AGROVOC	X	X	X	X	X	
EcoLexicon	X					
EEA - GEMET	X	X		X		X
EnvThes	X					
EOSterm	X					
EuroVoc		X			X	
Geological Survey of Austria (GBA)- thesaurus		X		X		
ICAN Dem. Thes.					X	
InterWATER					X	
IUGS-CGI Thes. of Geosciences					X	
NALT	X	X		X		
NERC NVS2.0	X	X				
SEMIDE					X	
SnowTerm	X					
SoilThes	X					
STW Thesaurus for Economics		X		X		
Tesauro Multilingüe de Medio Ambiente	X					
TF - DMEER						X
TF - EARTh		X	X	X		X
TF - EEA biogeographic region						X
TF - EUNIS HABITAT						X
TF - EUNIS SPECIES						X
TF - IUCN Protected Areas						X
TheSoz		X	X			
ThIST	X					
UMTHES		X				
UNESCO		X			X	
U.S. Geological Survey					X	
WQPB					X	

Other KOS 32

	QUESTIONNAIRE	Datahub	LITERATURE review	LOD Cloud	Community suggestions	TF from NATURESDIPlus
Codelists for metadata/data modelling						
BODC	X					
EEA - EIONET AQ Pollutants	X				X	
EEA - EIONET DATA DICTIONARY	X				X	
IUGS-CGI vocabularies	X				X	
INSPIRE Glossary	X				X	
INSPIRE IFCD	X				X	
OneGeology	X				X	
Ontology						
SWEET	X					
Taxonomic Datasets						
EEA-EUNIS HABITATS		X		X		
EEA-EUNIS SPECIES		X		X		X
PESI	X					
TAXREF	X					
Datasets						
EEA-EUNIS SITES		X		X		
EEA- E-PRTR	X	X				
NTUS	X	X				
Gazetteer						
FAO GeopoliticalOntology	X	X				
GeoNames Semantic Web		X		X	X	
GeoNames	X	X				
Metacarta					X	
TGN					X	
British Place Names					X	
SCHEMA/RDF vocabularies						
DCAT					X	
Data Cube vocabulary					X	
GEOVOCAB	X					
ORG					X	
WaterML2.0					X	
Glossary						
AwwaRF glossary					X	
Monterey Bay Aquarium Glossary					X	
WQA Glossary					X	
Vocabulary						
HYDROGRAPHIC DICTIONARY					X	
Other						
BIO_SOS					X	
OCLC Terminology Services					X	

- Not only thesauri, but different kinds of artefact
- The presence of the same terminological resources in LOD Cloud, SWJ dataset section, or data hub provides a thumb rule for reusability and for dataset popularity in the Linked Data community

Approach

Phase II: reusability criteria

- Language coverage
- % of concepts translated in each supported language
- Licence of language translation

Multi-lingualism

Licence

- Basic criteria "Openess of licence"

LD Compliance

5 stars classification

Tim Berners-Lee

- Basic criteria for LD compliance: "Dereferenceable URI "

Approach
Reusability: LD Criteria definition

- 5 Start classification of LD by Tim Berners-Lee
- HTTP dereferenceability of the URI mandatory LD prerequisite
  - to check authoritativeness of information associated to thesaurus concepts
  - to exploit mappings among thesauri concepts in order to discover further information in a follow-your-nose fashion

1 star	resources available on the web (whatever format)
2 stars	resources available as machine-readable structured data (e.g., Excel)
3 stars	as 2 stars plus non-proprietary format (e.g., CSV instead of Excel)
3,5 stars	resources available as RDF dump without dereferenceable HTTP URI
3,9 stars	resources provided as RDFa (RDF embedded in XHTML) or SPARQL end point which are very close to be LD ready but without dereferenceable HTTP URI
4 stars	all the above plus, use open standards from W3C (RDF and SPARQL) and HTTP dereferenceable URI to identify things, so that people can point at published resources
5 stars	all the above, plus links to other data to provide context

Reusability: Licence definition

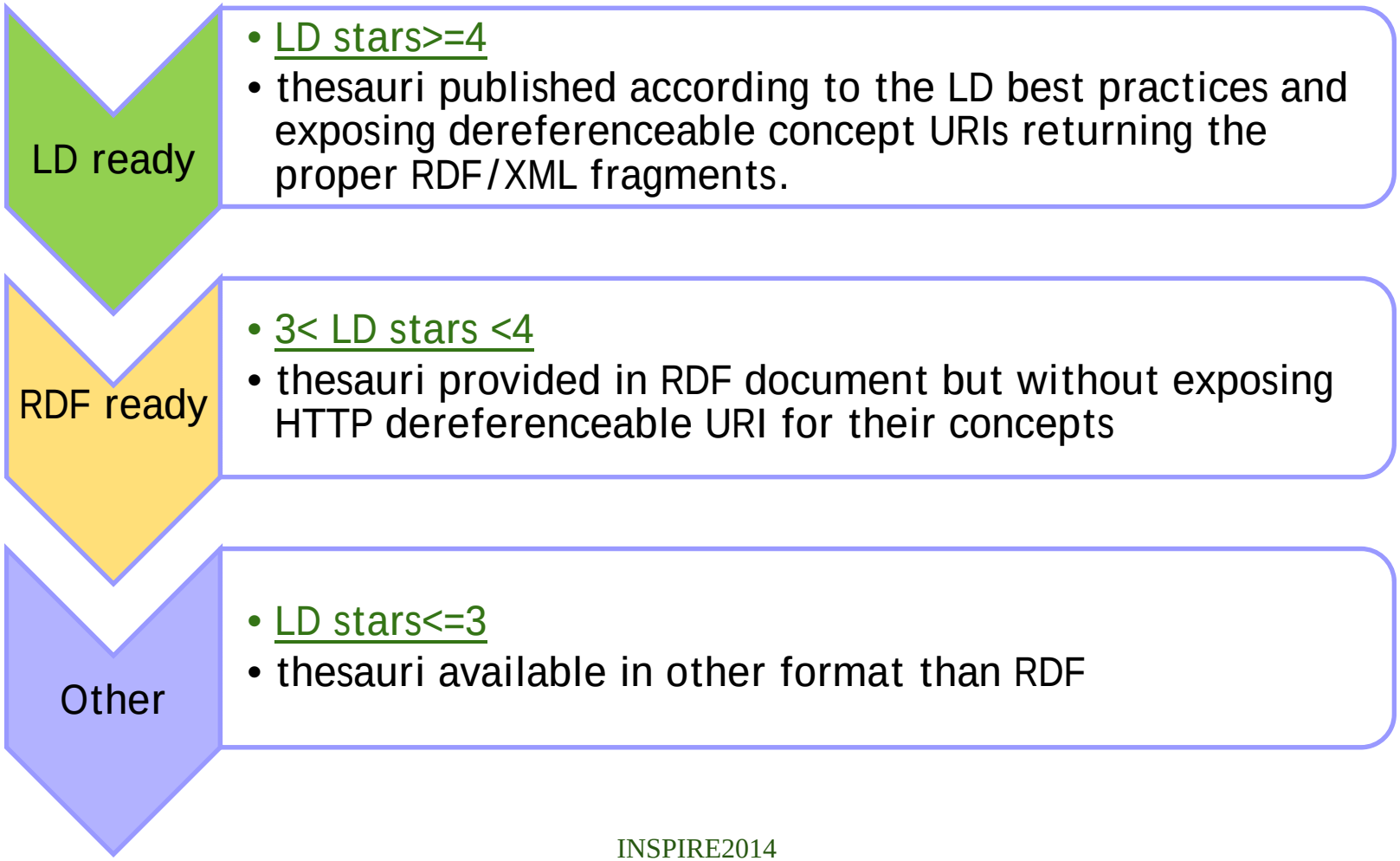
- Categories based on some existing and well-known type of licences (e.g. Creative Commons)
  - presented in "Rodrguez-Doncel, V., Gomez-Perez, A., Mihindukulasooriya, N.: Rights declara-tion in linked data. In: 4th Int. Work. on Consuming Linked Data (2013)"
- Level of reusability: 1=low reusability … 5= high reusability

Licence (acronym)	Characteristics	Licence reusability evaluation
Public Domain (CC0)	All the rights have been waived	5
Attribution (CC-BY)	Attribution is required	4.5
Share alike (CC-SA)	Copyleft licence	4
With restrictions (CC-NC , CC-ND, CC-NC-ND)	More severe restrictions	3.5
Closed (CR)	Closed licence	3
In progress (Pr)	Licence is going to be defined soon	2
Not found (NF)	No licence has been found in the website	1

Open licences , without severe restrictions:
complete reuse, transformation and publication of a resource

PhaseIII: LD Thesauri Evaluation

- LD analysis of thesauri in the reference catalogue
- Identification of three Macro Categories of LD Thesauri



PhaseIII: Licence Thesauri Evaluation

- Licence analysis of thesauri in the reference catalogue
- Identification of three Licence Macro Categories

Open Licenced Thesauri

- licence evaluation>=4
- highly reusable thesauri released under public domain, attribution or share-alike licences. They can be modified and extended and deployed in commercial and non-commercial context

Partially Open Licenced

- Licence evaluation =3.5
- thesauri licenced with some further restrictions in reusability.

Closed Licenced Thesauri

- licence evaluation<3.5
- It considers thesauri in which licence forbids the free reuse or for which a licence is not provided yet

PhaseIII: Overall Thesauri Evaluation

- Analysis of the thesauri respect to the macro-categories identified for LD stars and licence

	LD ready	RDF ready	Other
Open Licenced	SoilThes, GEMET, AGROVOC, NERC NVS2.0, GBA, NALT	EuroVoc, UMTHES	
Partially Open Licenced	TheSoz, EARTh, UNESCO	STW, SnowTerm, EOSterm	SEMIDE, InterWATER
Closed Licenced	EnvThes, ICAN	ThIST, U.S.G.S., ADL FTT	IUGS-CGI , EcoLexicon, WQPB

- Results
  - 12 (45%) Thesauri are LD ready (6 are interlinked with third party thesauri)
  - 8 (33%) have the SKOS deployed in RDF ready
  - Thesauri are equally distributed among Licence categories,
    => only the 33% of thesauri are truly open licence

Consideration and recommendation

- Considerations
  - The Thesaurus Catalogue provides good level of reusability
    - (58% Thesauri are both LD or RDF ready and Open or Partial Open Licence
- Recommendations to improve reusability
  - More attention to HTTP dereferenceability of Concept URIs
    - 54% are not complete in HTTP dereferenceable
  - Licence should be more carefully stated
    - Thesauri are available in more then one sources but rarely licence is stated in all the sources (e.g. thesaurus's portal, datahub)
    - Sometimes it is missing an explicit web link to the licence

Conclusions & Future Work

Outcomes

- Reference catalogue of thesauri on the web and their evaluation with respect to licence and LD compliance.
- Investigation approach and stress of reusability criteria domain independent and recommendation for thesaurus user and publisher

Future work

- Analysis refinement
  - Evaluation of multilingualism
  - SKOS quality (e.g. QSKOS)
  - Quality of interlinking:
    - How enabling are interlinkings in a joint exploitation of the thesauri?
- A web portal to expose the whole catalogue / the reusability evaluation.
- LusTRE … A new release end of year

Thank you !

Contact persons

CNR-IMATI
monica.demartino@ge.imati.cnr.it

riccardo.albertoni@ge.imati.cnr.it

References

eEnvplus project (http://www.eenvplus.eu/),
Deliverable D4.1: Thesaurus Survey
LusTRE Thesaurus Framework http://linkeddata.ge.imati.cnr.it:2020/

Publication:
Albertoni R.,De Martino M.,Podestà P.,
Environmental thesauri under the lens of Reusability,
EGOVIS 2014, (to appear)

Multilingualism: how to deepen the analysis for KOS in LOD

- The percentage of concepts translated in different languages (prefLabel)

	languages																																		
Thesaurus name	ar	bg	ca	cs	de	el	en	en-US	es	eu	fa	fr	ga	hi	hr	hu	it	ja	ko	lo	lt	lv	ms	mt	no	pl	pt	ro	ru	sk	sv	te	th	tr	uk	zh-CN
gemet	100	100	100	100	100	100	100	100	100	100	0	100	100	0	91	0	100	0	0	0	100	100	0	100	100	100	100	100	100	99	92	0	0	100	100	100
earth	0	0	0	0	0		100	0	0	0	0	0	0	0	0	0	96	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
agrovoc	76	0	0	99	64	0	100	0	96	0	61	96	0	62	0	61	63	95	39	53	0	0	1	0	0	61	97	0	61	59	0	10	61	96	1	96

- The percentage of concepts translated in different languages (altLabel)

	languages																																		
Thesaurus name	ar	cs	de	el	en	es	et	eu	fa	fi	fr	hi	hr	hu	id	it	ja	ko	lo	mr	ms	pl	pt	ru	sk	sv	te	th	tr	uk	zh
GEMET	0	0	0	23	0,2	0	0,1	33	0	5,4	0	0	25	0	0	0	0	0	0	0	0	0	0	0	1,8	16	0	0	0	0	0,1
EARTH	0	0	0		8,3	0	0	0	0	0	0	0	0	0	0	5,9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AGROVOC	3,3	26	31	0	27	34	0	0	28	0	24	23	0	22	0	23	23	7	7,7	0	2,1	27	20	24	21	0	3,3	18	31	0,5	17