eosc | AquaINFRA

Joint EuroGeographics and EuroSDR Virtual Workshop on Geodata Discovery

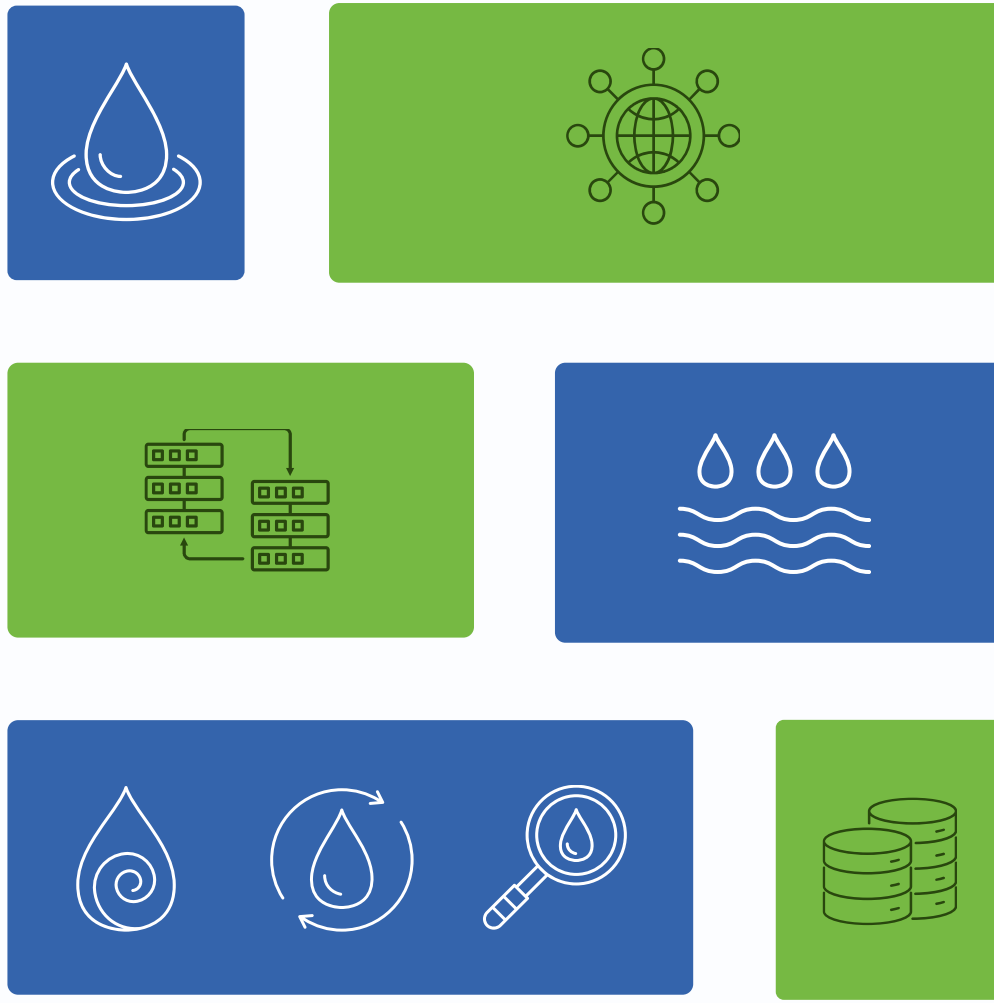
16th January 2024

Enhancing Geospatial Data Discoverability with Ontology and Thesaurus Data in the AquaINFRA Project

Pekka Latvala

Finnish Geospatial Research Institute FGI

pekka.latvala@nls.fi



This project has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No 101094434. The European Commission is not responsible for any use that may be made of the information it contains.

/aquainfraeu /aquainfraeu /aquainfra

AquaINFRA project

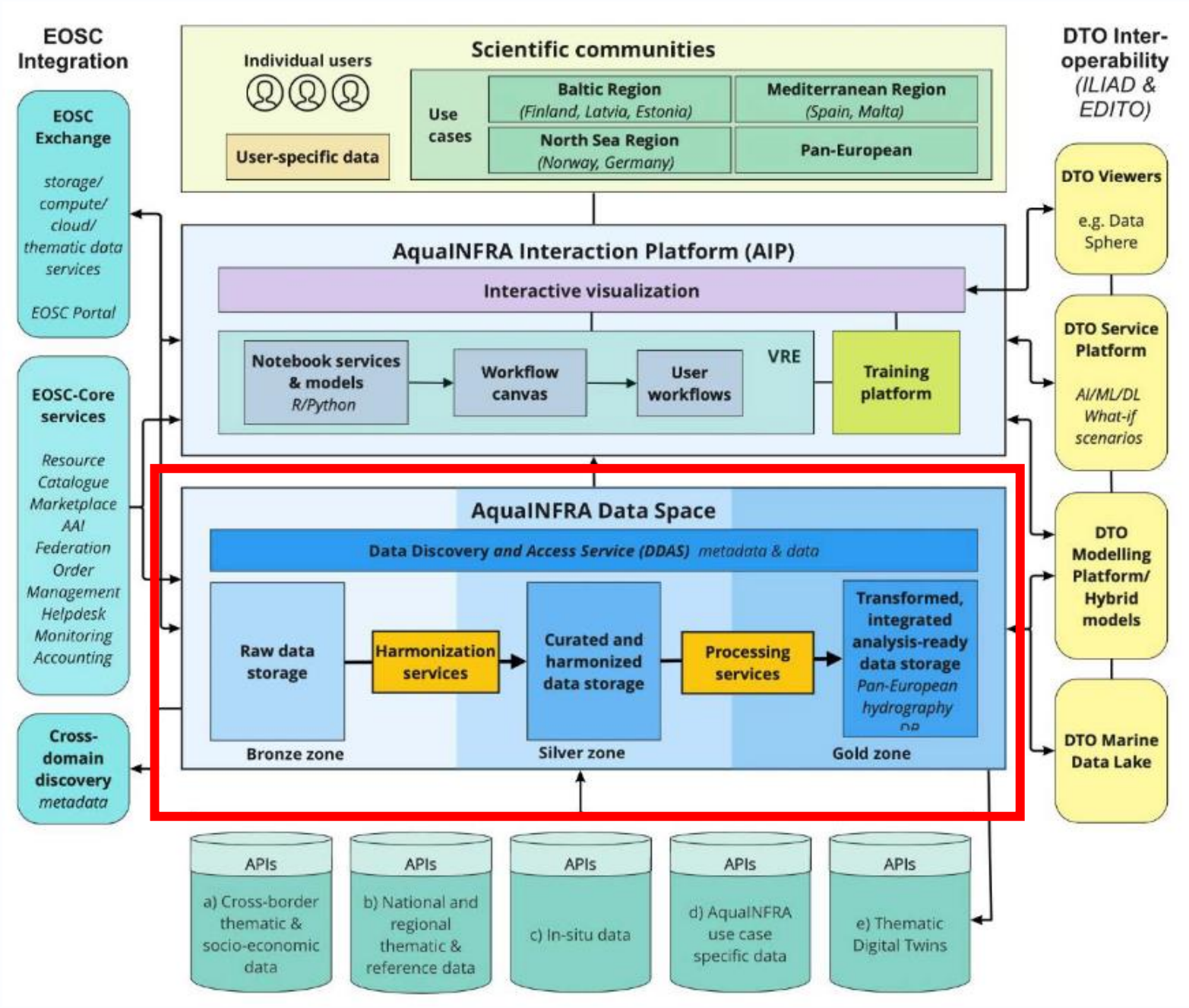
- AquaINFRA is an EOSC (European Open Science Cloud) project
  - Running for 4 years (2023 – 2026)
- 21 partners from 10 countries:
  - Denmark, Finland, Norway, Estonia, Latvia, Germany, Spain, Malta, Austria, United Kingdom
- Objective: To develop a virtual environment that contains FAIR data and services that support the research activities in marine and freshwater domains.

EOSC | AquaINFRA

Architecture

- Data Discovery and Access Service (DDAS)
- Implementation: pygeoapi, CKAN
- Provides federated metadata search and data access
- The ontology search component is a web service that is part of the DDAS

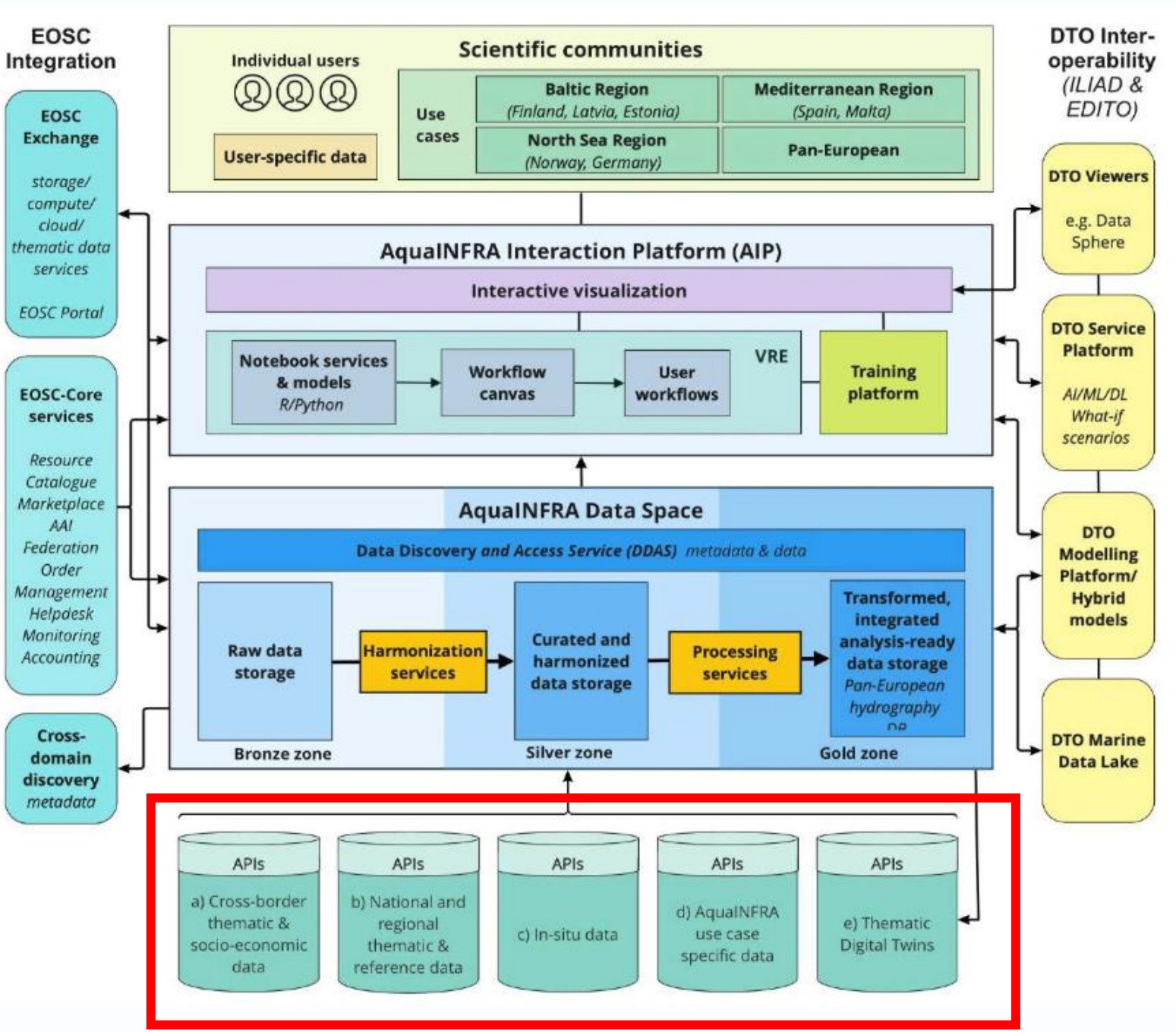
- Data zones:
  - Bronze – Raw data
  - Silver – Harmonized data
  - Gold – Transformed and integrated, analysis-ready data



EOSC | AquaINFRA

Architecture

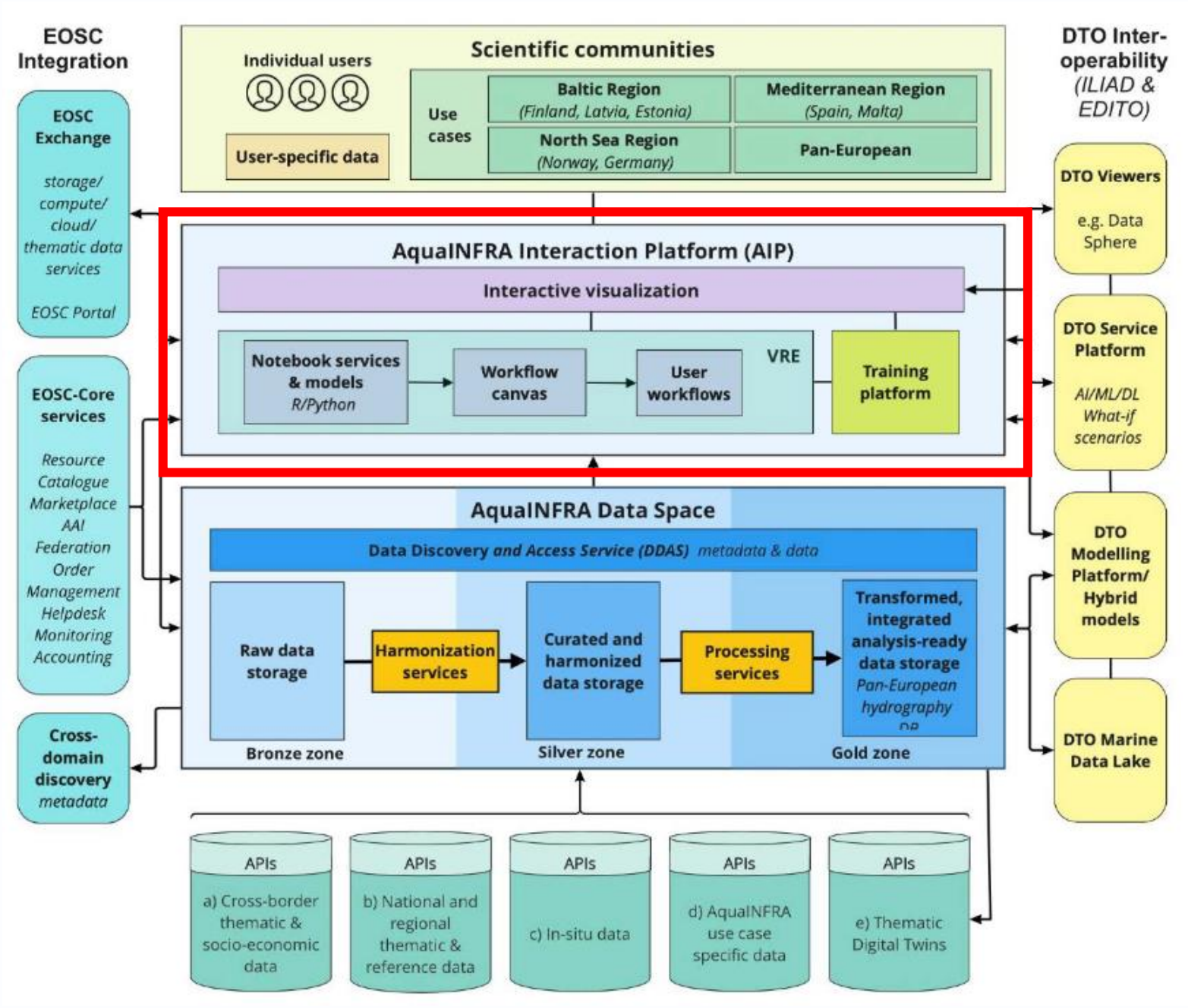
- Supported DDAS background services:
  - Metadata Search
    - OGC's Catalogue Service for Web (CSW)
    - OGC API – Records
    - Custom APIs
      - Data provider plugins
  - Data Access
    - OGC API – Features
    - OGC API – Coverages



eosc | AquaINFRA

Architecture

- AquaINFRA Interaction Platform
  - Utilizes the DDAS service
  - Search
  - Access
  - Data processing
  - Visualization
  - Virtual Research Environment
    - Data analysis workflows
  - Training platform



Ontology search in AquaINFRA DDAS

- Purpose: To enhance DDAS metadata searches by providing related terms search to the metadata search keyword.
- Use of ontology / thesaurus data for searching the related keywords.

Inventory of Potential Ontologies and Thesauri

Ontologies:

- EIFFEL ontology
    - 3 domains: Essential Climate Variables, Sustainable Development Goals and Earth Observation.
    - Contains only some relevant classes.
- Ocean Data ontology
    - 76 classes.
    - Not very relevant.
- InWaterSense ontology (Intelligent Wireless Sensor Networks for Monitoring Surface Water Quality)
    - ~250 classes.
    - Focused on sensors and sensor data.
- Surface Water Ontology
    - 94 classes.
    - Contains various feature types and waterbody types.
- AFO - Natural resource and environment ontology
    - Available in Finnish and English (2/3 of concepts).
    - 6000 concepts, and also ~25000 concepts from general finnish ontology.
    - Contains ~60 relevant hydrographic concepts.

Inventory of Ontologies and Thesauri

Thesauri:

- USGS thesaurus
  - ~1000 concepts, mainly geology-related terminology.
  - Contains broader and narrower terms, alternative labels and related terms.
- UNESCO thesaurus
  - A controlled and structured list of concepts in the fields of education, culture, natural sciences, social and human sciences, communication and information.
  - Contains narrower, and related concepts.
  - 4581 concepts, ~140 related to hydrography.
- UNBIS thesaurus
  - A multilingual database of the controlled vocabulary used to describe UN documents and other materials in the Library's collection.
  - ~80 concepts related to hydrography.
- AGROVOC Multilingual thesaurus
  - Multilingual, 42 languges, number of concepts varies between languages.
  - Narrower broader and related concepts.
  - ~600 hydrographic concepts, including many lake, river etc. names.

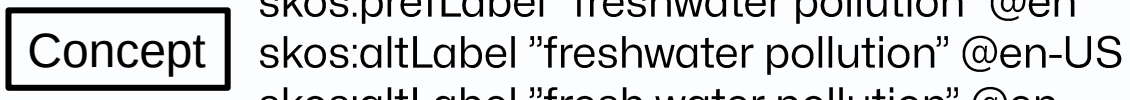
Inventory of Ontologies and Thesauri

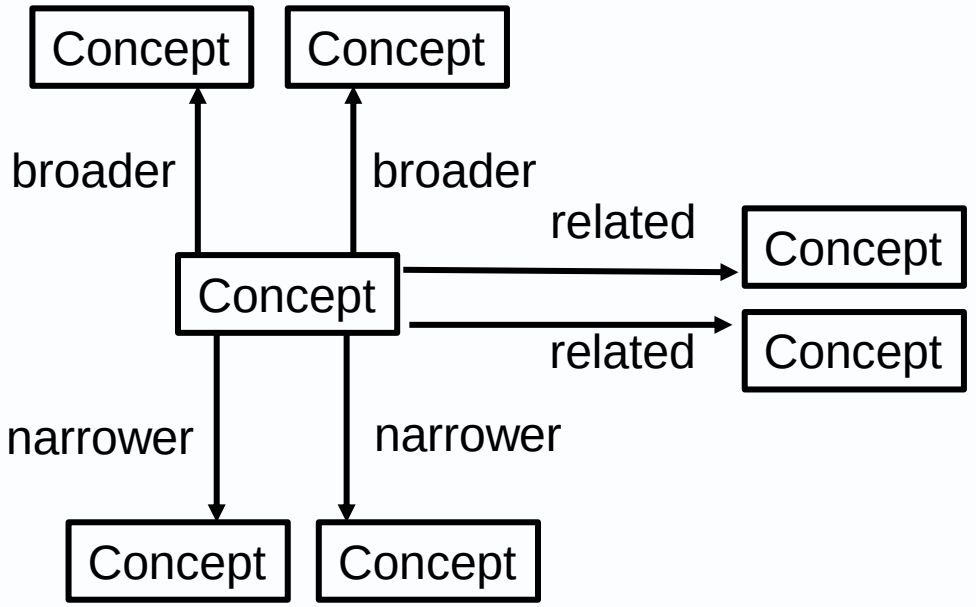
Thesauri:

- GEMET (General Multilingual Environmental Thesaurus)
  - Contains relevant themes. hydrosphere, water.
  - ~100 relevant concepts.
  - Multilingual data.

GEMET Data

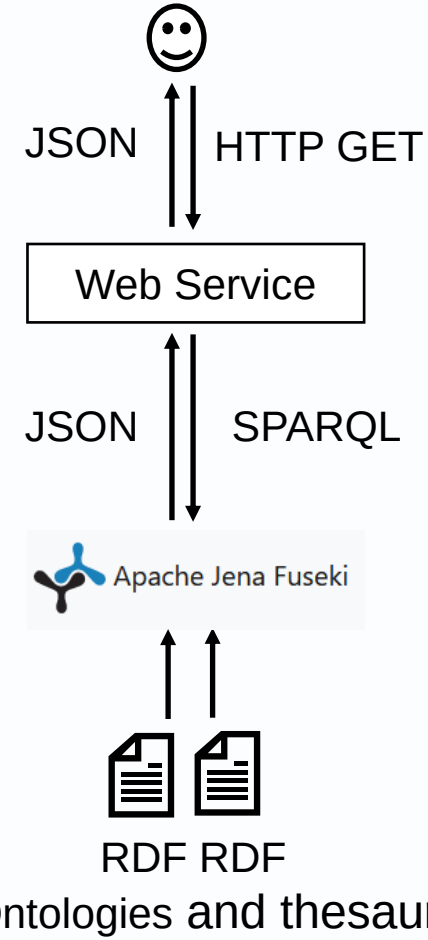
- GEMET (General Multilingual Environmental Thesaurus) was selected as a starting point for the implementation work.
- Multilingual terms.
  - Terms written with up to 37 languages.
  - Work is focused only on the English language.
    - English and American English terms (en, en-US).
- Concepts may have preferred labels and alternative labels.
- Concepts may have links to other concepts with broader, narrower and related relations.





Web Service Implementation

- Data store:
  - Apache Jena Fuseki SPARQL server
  - Ontology / thesaurus data are loaded in the RDF format
  - SPARQL queries can be made to the RDF data



- Web Service
  - Python implementation.
    - django, django REST Framework, SPARQLWrapper libraries.
  - Output available in JSON format.
  - Parameters:
    - keyword [string] - search word
    - broader [boolean] - includes broader concept labels to output [not their children]
    - related [boolean] - includes related concept labels to output [including all children if narrower = true
    - narrower [boolean] - includes narrower concepts labels to output [including all children]

Example 1: Search only for concepts that match with the original metadata search keyword.

keyword=lake&broader=false&related=false&narrower=false

Keywords are searched from the skos:prefLabel and skos:altLabel elements

Results

Concept 579	Label ”artificial lake” @en Label ”artificial lake” @en-US
Concept 4590	Label ”lake basin” @en Label ”lake basin” @en-US
Concept 4593	Label ”lake pollution” @en Label ”lake pollution” @en-US
Concept 4594	Label ”lake” @en Label ”lake” @en-US

Duplicate label names are removed

Output:
1. artificial lake
2. lake
3. lake basin
4. lake pollution

Example 2: Search for concepts that match with the original metadata search keyword including broader concepts.

keyword=lake&broader=true&related=false&narrower=false

Only the first immediate parent of found concepts are included to the results.
Related and narrower parameters don't have any influence on broader concepts.

Results:

Concept 579	Label "artificial lake" @en Label "artificial lake" @en-US	skos:broader	Concept 7138	Label "reservoir" @en Label "reservoir" @en-US
Concept 4590	Label "lake basin" @en Label "lake basin" @en-US	skos:broader	Concept 8386	Label "terrestrial area" @en Label "terrestrial area" @en-US
Concept 4593	Label "lake pollution" @en Label "lake pollution" @en-US	skos:broader	Concept 3493	Label "freshwater pollution" @en Label "freshwater pollution" @en-US AltLabel "fresh water pollution" @en
Concept 4594	Label "lake" @en Label "lake" @en-US	skos:broader	Concept 4124	Label "hydrosphere" @en Label "hydrosphere" @en-US

Duplicate label names are removed

Output:
1. artificial lake
2. fresh water pollution
3. freshwater pollution
4. hydrosphere
5. lake
6. lake basin
7. lake pollution
8. reservoir
9. terrestrial area

Broader concepts may or may not be relevant to the keyword

Example 3: Search for concepts that match with the original metadata search keyword including related concepts.

keyword=lake&broader=false&related=true&narrower=false

All immediate related concepts are included to the results.
Broader parameter doesn't have any influence on related concepts.
Narrower parameter includes also narrower concepts of found related concepts.

Results:

Concept 579	Label "artificial lake" @en Label "artificial lake" @en-US			
Concept 4590	Label "lake basin" @en Label "lake basin" @en-US	skosrelated	Concept 12248	Label "water reservoir" @en Label "water reservoir" @en-US
		skosrelated	Concept 15281	Label "river basin" @en
		skosrelated	Concept 15282	Label "river basin management" @en
Concept 4593	Label "lake pollution" @en Label "lake pollution" @en-US			
Concept 4594	Label "lake" @en Label "lake" @en-US	skos:related	Concept 5967	Label "overturn (limnology)" @en Label "overturn (limnology)" @en-US

Duplicate label names are removed

Output:
1. artificial lake
2. lake
3. lake basin
4. lake pollution
5. overturn (limnology)
6. river basin
7. river basin management
8. water reservoir

Related concepts may or may not be relevant to the keyword

Example 4: : Search for concepts that match with the original metadata search keyword including related and narrower concepts.

keyword=lake&broader=false&related=true&narrower=true

All immediate related and narrower concepts are included to the results
Broader parameter doesn't have any influence on related concepts
Narrower parameter includes also narrower concepts of related concepts

Results:

Concept 579
Label "artificial lake" @en
Label "artificial lake" @en-US

Concept 4590
Label "lake basin" @en
Label "lake basin" @en-US

skosrelated — Concept 12248
Label "water reservoir" @en
Label "water reservoir" @en-US

skosrelated — Concept 15281
Label "river basin" @en

skos:narrower — Concept 4440
Label "international river basin" @en
Label "international river basin" @en-US

skos:narrower — Concept 7251
Label "river basin development" @en
Label "river basin development" @en-US

skos:narrower — Concept 15282
Label "river basin management" @en

skosrelated — Concept 15282
Label "river basin management" @en

skos:narrower — Concept 4440
Label "international river basin" @en
Label "international river basin" @en-US

skos:narrower — Concept 7251
Label "river basin development" @en
Label "river basin development" @en-US

… Example 4: Example 4: Search for concepts that match with the original metadata search keyword including narrower concepts.

keyword=lake&broader=false&related=true&narrower=true

… Results:

Concept 4593	Label ”lake pollution” @en Label ”lake pollution” @en-US			
Concept 4594	Label ”lake” @en Label ”lake” @en-US	skos:related	Concept 5967	Label ”overturn (limnology)” @en Label ”overturn (limnology)” @en-US

Duplicate label names are removed

Output:
1. artificial lake
2. International river basin
3. lake
4. lake basin
5. lake pollution
6. overturn (limnology)
7. river basin
8. river basin development
9. river basin management
10. water reservoir

Related concepts and their children may or may not be relevant to the keyword

In general, narrower concepts are always relevant to the parent concept

Future Work

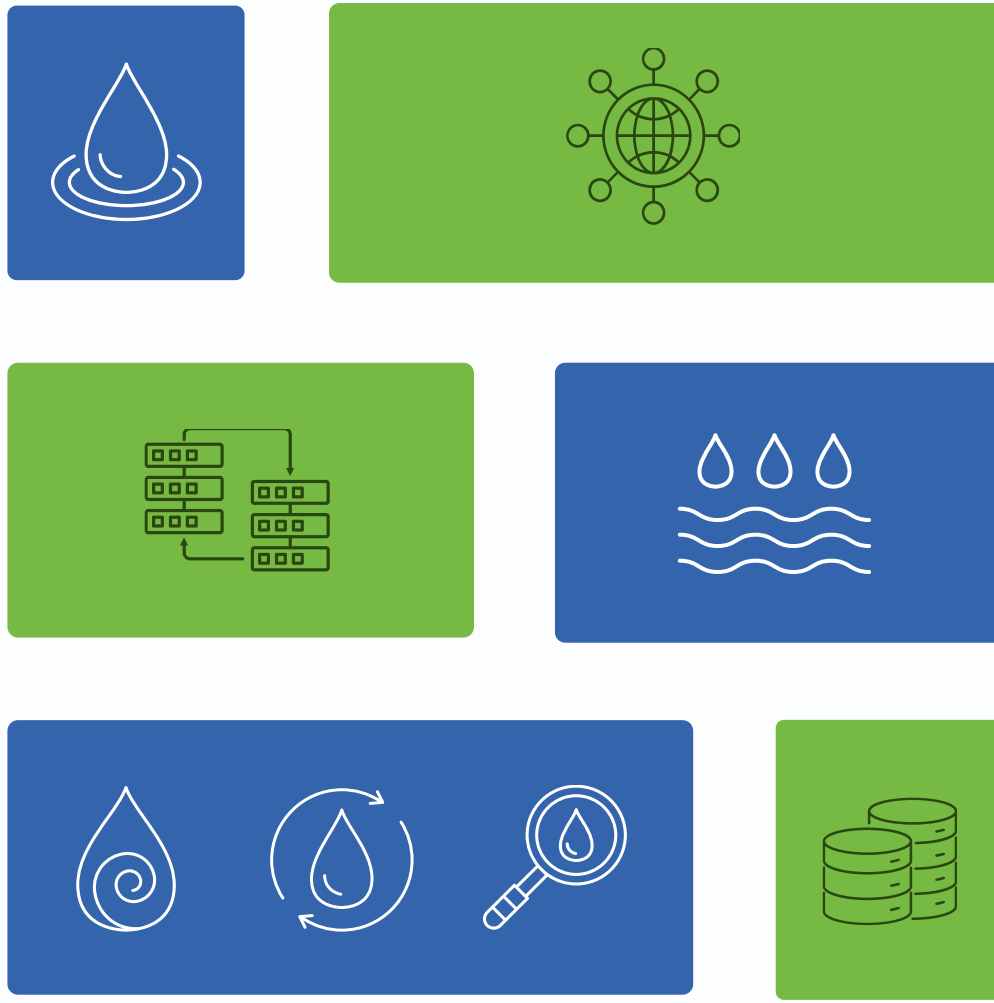
- Handling of plural words.
    - Morphological analysis.
- Expanding the related terms search into other areas.
    - Socio-economical data.
- Handling the results in the user interface.
    - Manually filtering out the results that are unrelated to the original metadata search keyword.

Thank You

Pekka Latvala

Finnish Geospatial Research Institute FGI

pekka.latvala@nls.fi



This project has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No 101094434. The European Commission is not responsible for any use that may be made of the information it contains.

/aquainfraeu /aquainfraeu /aquainfra