CÁCERES 2019 JIIDE

Las IDE locales, acercando la información digital a los ciudadanos.

23, 24 y 25 de octubre

Spatial data discovery and indexing tools: an approach based on metadata and fitness for use

Pedro Castro[1,3,4], Joaquim Alonso[2,3,4], Ivone Martins[3,4], João Honrado[3,5], Johannes Peterseil[6]

(1) Arc4DigiT - Applied Research Centre for Digital Transformation
(2) ProMetheus - Research Unit in Materials, Energy and Environment for Sustainability
(3) CIBIO-InBIO - Research Centre in Biodiversity and Genetic Resources
(4) Instituto Politécnico de Viana do Castelo
(5) Faculdade de Ciências, Universidade do Porto
(6) Umweltbundesamt Gmbh · Ecosystem Research and Environmental Information Management

This work has been carried out within the H2020 project 'ECOPOTENTIAL: Improving Future Ecosystem Benefits Through Earth Observations' (http://www.ecopotential-project.eu). The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 641762.

Overview

- Context
- Objectives
- Framework Overview
- Workflow

Spatial Data Quality

Spatial Data Quality issues are important:

- the **increasing amount of spatial data production, handling and sharing** with different sources, different frequency of acquisition, different spatio-temporal scales, different levels of accuracy, different processing methods or techniques leads to many challenges in SDQ assessment
- Spatial Data is used **in very different application contexts** (data is often used with purposes other than producer' intended ones)
- It is necessary to **consider data quality** to identify datasets that satisfy the requirements of a particular application for specific user

Metadata

- Efforts have been made in metadata development and in meta-evaluation of external and (in)direct quality by the end-user(s), taking advantage of **metadata documentation possibilities and quality communication**

- Standard metadata profiles can contain a description of attributes about **dataset/database content, access and use conditions**, thus **allowing the assessment of data quality components and elements** (ISO 19157) as well as data quality management (ISO 19158).

- In this context, metadata catalogues offer opportunities for the implementation and improvement of spatial data quality evaluation/assessment tools related to knowledge discovery, searching and indexing

ThemisE

THEmatic Metadata-based and fItness-for-use Spatial data quality Evaluation platform – **ThemisE** platform:

- implemented as an autonomous and modular **Web application** to perform **quality evaluation BASED ON METADATA** considering that
  - Metadata can contain information about the content, quality, condition and other characteristics of the data (ISO 2005) that can be used for (meta)quality evaluation
  - Frequent limitations to data access and use
  - Increasing availability of metadata catalogues allowing a (simple) integration with an evaluation platform
- With the aim to support the **quality-driven discovery and selection of relevant datasets** (or the identification of data gaps) necessary for environmental/ecological modelling based on (well documented) datasets' metadata

Framework overview

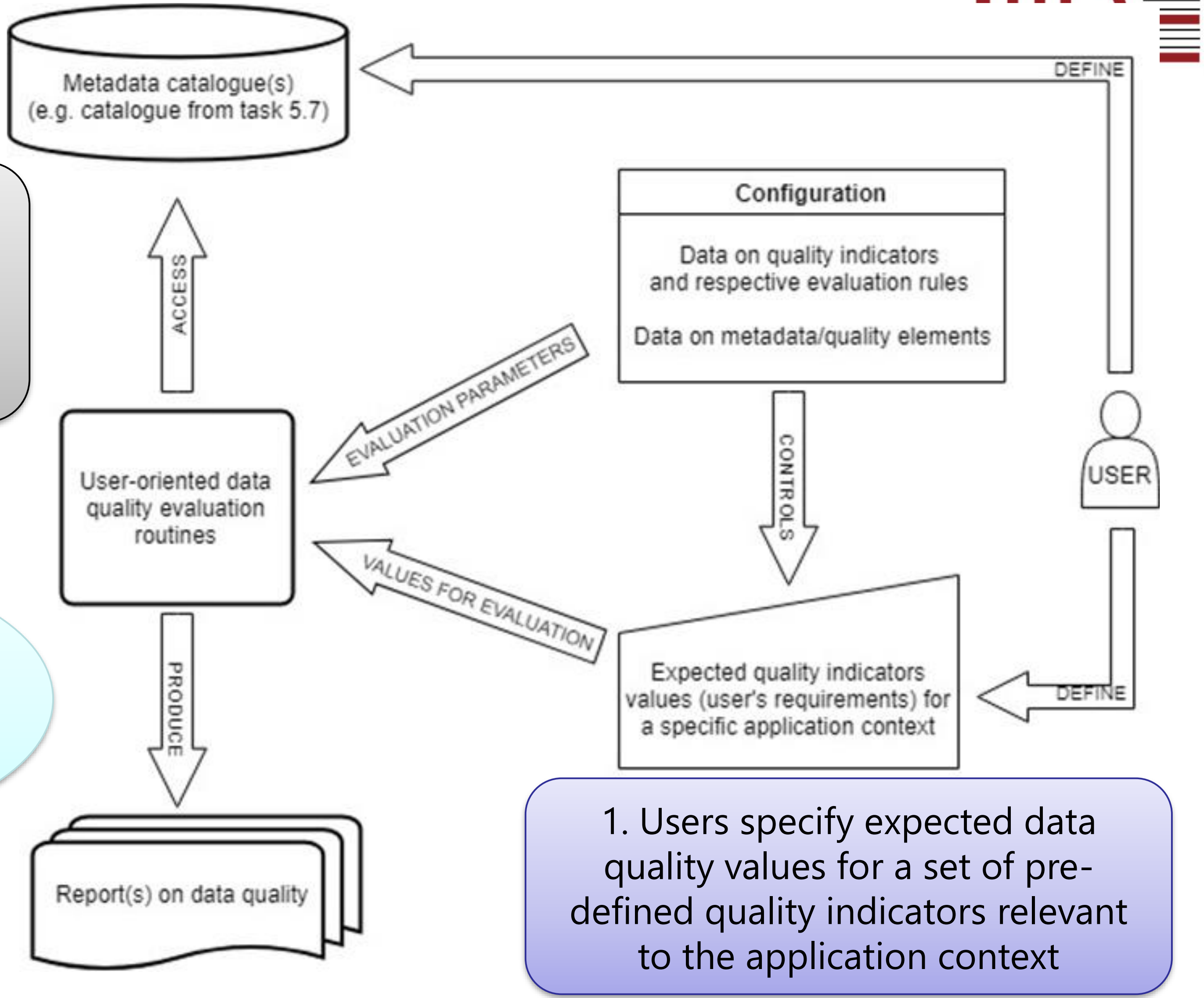
ThemisE platform allows two types of evaluation:

- an internal evaluation centered on the comparison of the characteristics of the dataset, as detailed in metadata by the producer, with the required elements according to a predefined standard profile;

- an external quality evaluation that is based on determining the **matching level** (fitness-for-use) between the **characteristics of the dataset** (detailed by its **metadata**) and the **characteristics of the data required by the user** that describe the user's requirements for a given application context (and defined through expected values for predefined quality indicators)

focused on evaluating how data will fit the users' needs, to bring data sets closer to users' applications

General process

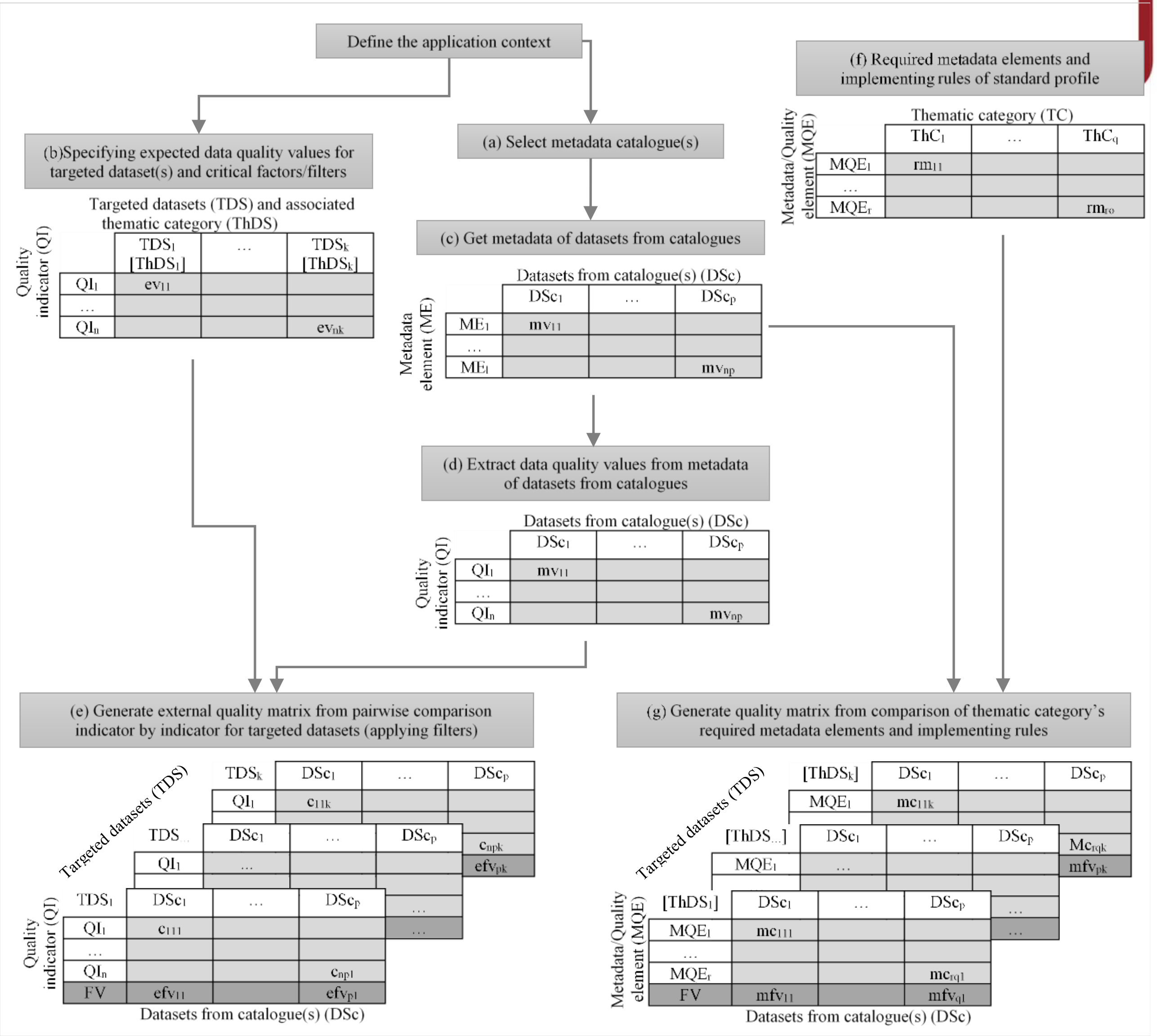

Metadata catalogue(s)
(e.g. catalogue from task 5.7)
DEFINE
ACCESS
Configuration
Data on quality indicators
and respective evaluation rules
Data on metadata/quality elements
EVALUATION PARAMETERS
User-oriented data
quality evaluation
routines
CONTROLS
USER
VALUES FOR EVALUATION
Expected quality indicators
values (user's requirements) for
a specific application context
DEFINE
PRODUCE
Report(s) on data quality

1. Users specify expected data quality values for a set of pre-defined quality indicators relevant to the application context

2. Extract internal data quality values from metadata catalogues for quality indicators selected by the user

3. Run comparison of information provided by users and metadata

Functional workflow



Define the application context

(a) Select metadata catalogue(s)

(b)Specifying expected data quality values for targeted dataset(s) and critical factors/filters

Targeted datasets (TDS) and associated thematic category (ThDS)

Quality indicator (QI)

	TDS$_1$ [ThDS$_1$]	…	TDS$_k$ [ThDS$_k$]
QI$_1$	ev$_{11}$		
…			
QI$_n$			ev$_{nk}$

(c) Get metadata of datasets from catalogues

Datasets from catalogue(s) (DSc)

Metadata element (ME)

	DSc$_1$	…	DSc$_p$
ME$_1$	mv$_{11}$		
…			
ME$_l$			mv$_{np}$

(d) Extract data quality values from metadata of datasets from catalogues

Datasets from catalogue(s) (DSc)

Quality indicator (QI)

	DSc$_1$	…	DSc$_p$
QI$_1$	mv$_{11}$		
…			
QI$_n$			mv$_{np}$

(f) Required metadata elements and implementing rules of standard profile

Thematic category (TC)

Metadata/Quality element (MQE)

	ThC$_1$	…	ThC$_q$
MQE$_1$	rm$_{11}$		
…			
MQE$_r$			rm$_{ro}$

(e) Generate external quality matrix from pairwise comparison indicator by indicator for targeted datasets (applying filters)

Targeted datasets (TDS)

TDS$_k$	DSc$_1$	…	DSc$_p$
QI$_1$	c$_{11k}$		
			c$_{npk}$
			efv$_{pk}$

TDS$_{…}$	DSc$_1$	…	DSc$_p$
QI$_1$	…		
			…
			…

Quality indicator (QI)

TDS$_1$	DSc$_1$	…	DSc$_p$
QI$_1$	c$_{111}$		
…			
QI$_n$			c$_{np1}$
FV	efv$_{11}$		efv$_{p1}$

Datasets from catalogue(s) (DSc)

(g) Generate quality matrix from comparison of thematic category's required metadata elements and implementing rules

Targeted datasets (TDS)

[ThDS$_k$]	DSc$_1$	…	DSc$_p$
MQE$_1$	mc$_{11k}$		
			Mc$_{rqk}$
			mfv$_{pk}$

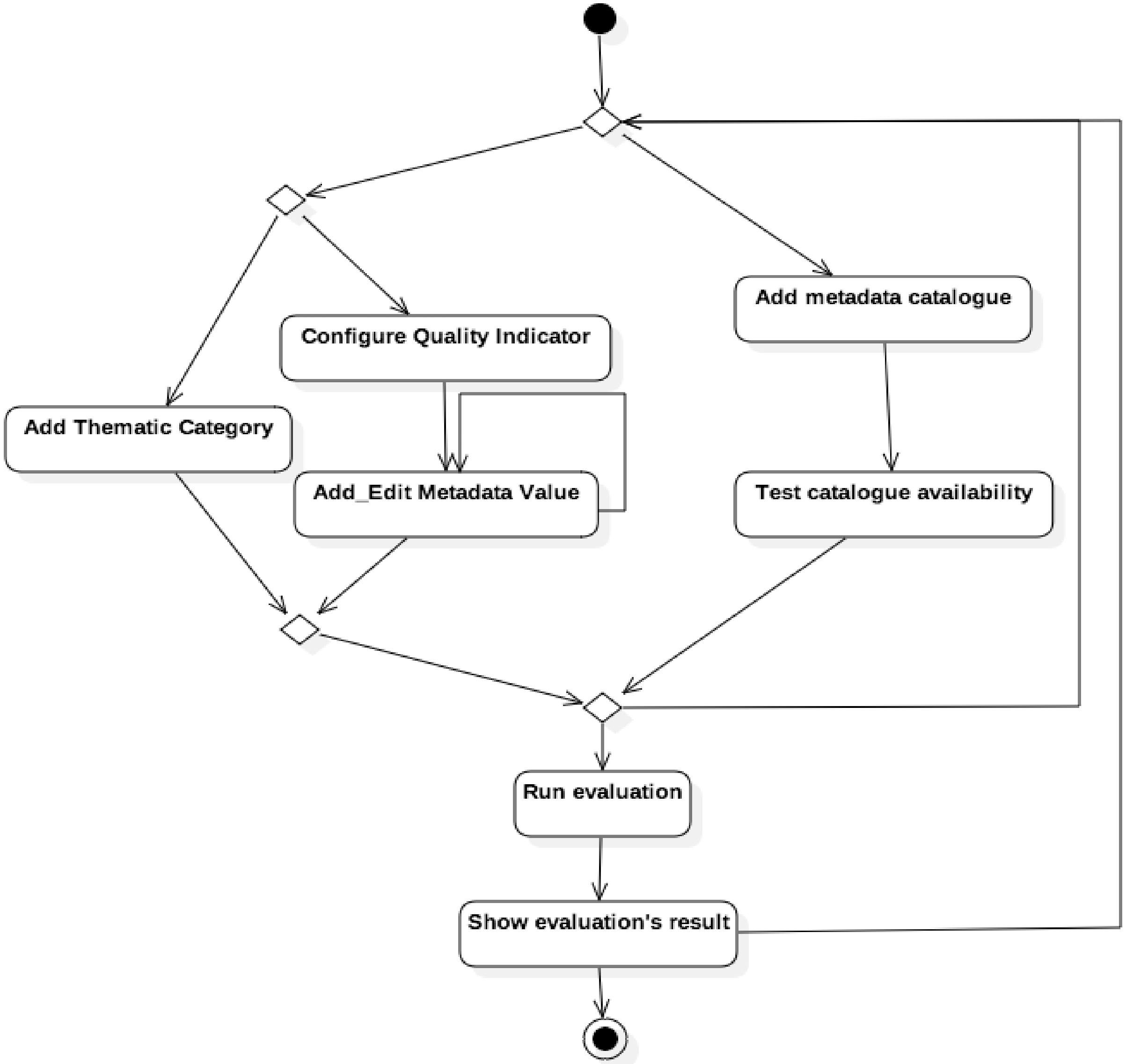
[ThDS$_{…}$]	DSc$_1$	…	DSc$_p$
MQE$_1$	…		
			…
			…

Metadata/Quality element (MQE)

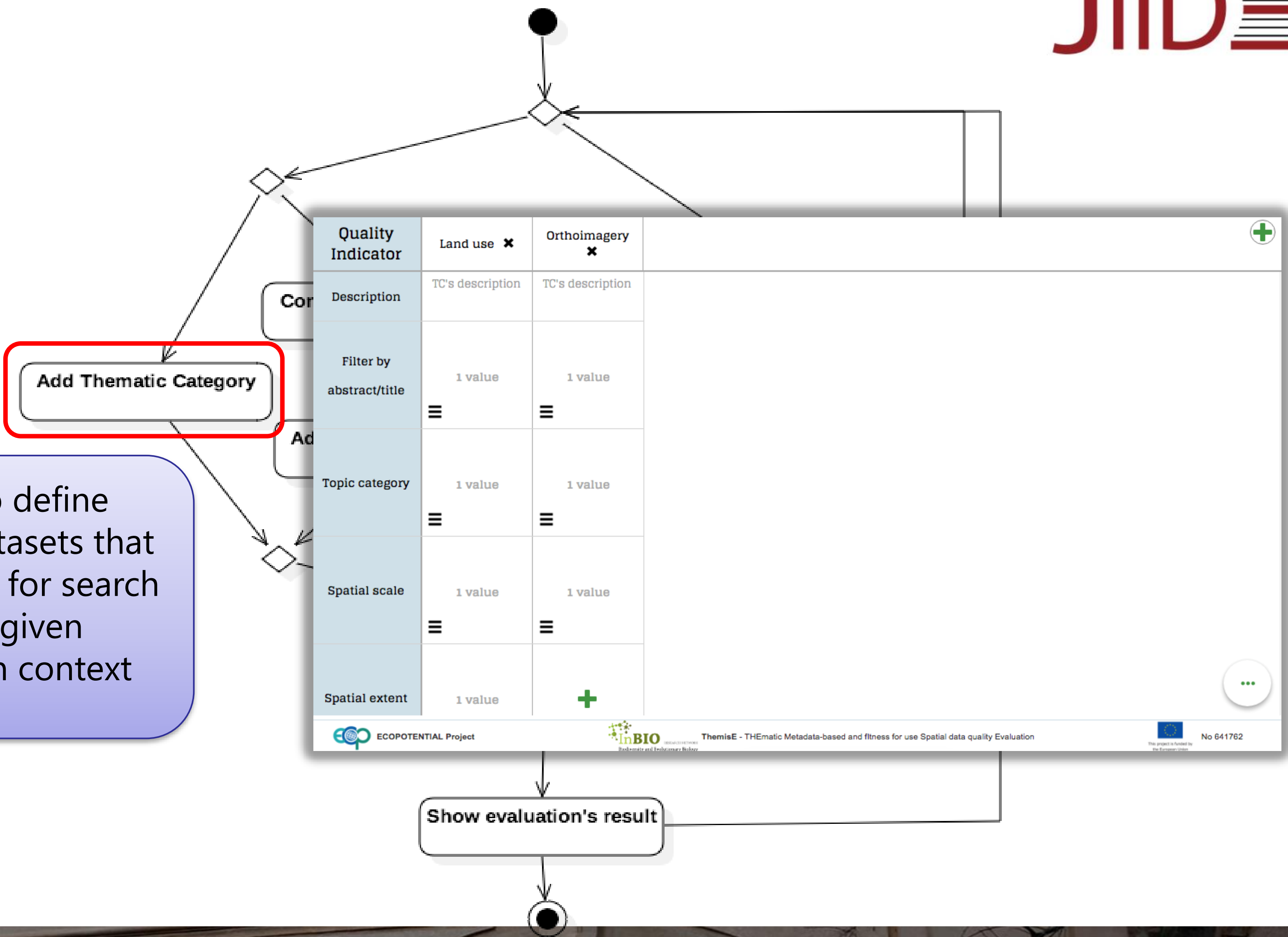
[ThDS$_1$]	DSc$_1$	…	DSc$_p$
MQE$_1$	mc$_{111}$		
…			
MQE$_r$			mc$_{rq1}$
FV	mfv$_{11}$		mfv$_{q1}$

Datasets from catalogue(s) (DSc)

Actions workflow



Thematic category



Quality Indicator	Land use ✖	Orthoimagery ✖
Description	TC's description	TC's description
Filter by abstract/title	1 value ≡	1 value ≡
Topic category	1 value ≡	1 value ≡
Spatial scale	1 value ≡	1 value ≡
Spatial extent	1 value	+

ECOPOTENTIAL Project

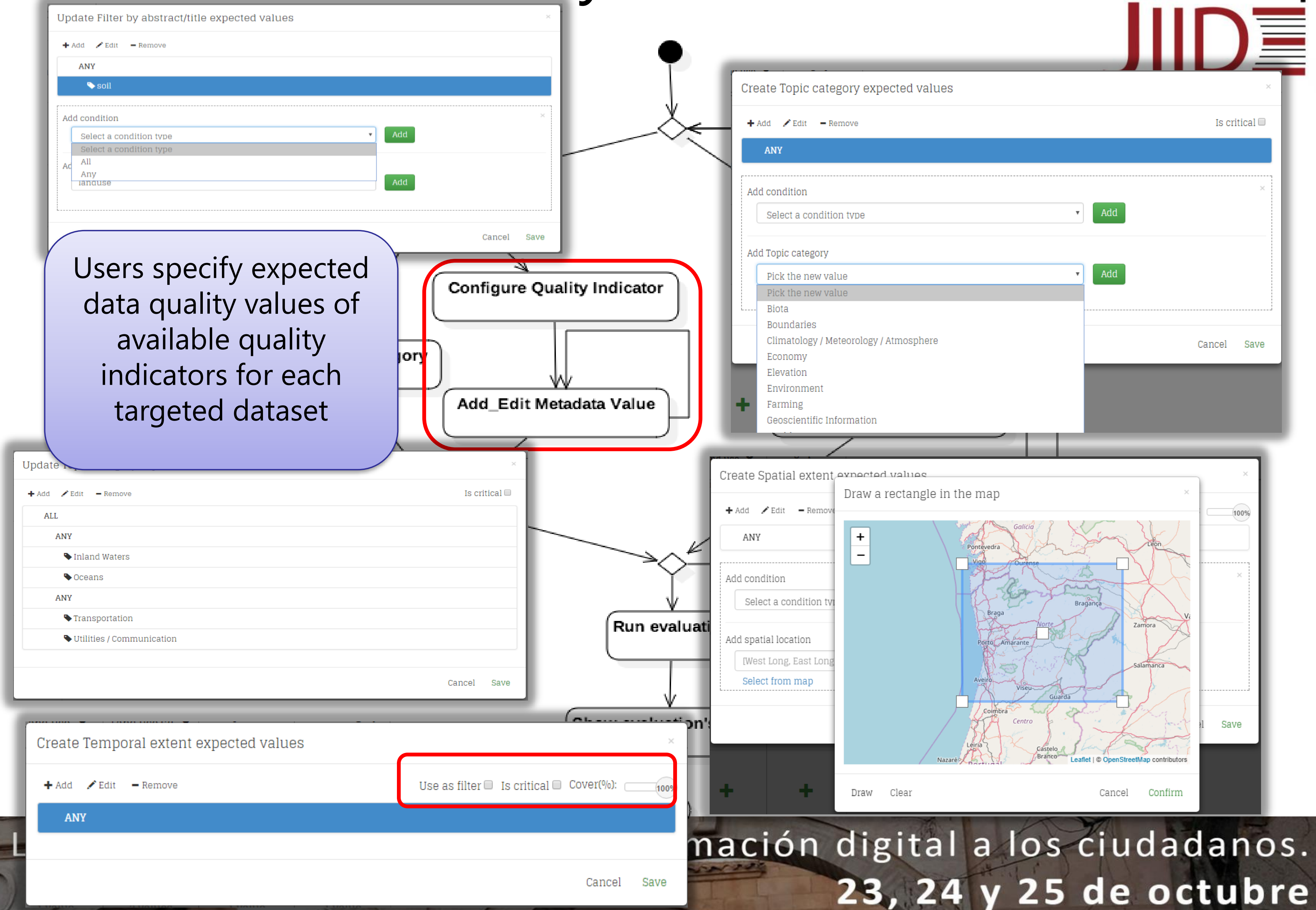
InBIO

ThemisE - THEmatic Metadata-based and fitness for use Spatial data quality Evaluation

No 641762

Allows to define different datasets that are targeted for search for the given application context

Quality indicators



Users specify expected data quality values of available quality indicators for each targeted dataset

Update Filter by abstract/title expected values

✚ Add ✎ Edit ━ Remove

ANY

soil

Add condition

Select a condition type ▾ Add

Select a condition type
All
Any

landuse Add

Cancel Save

Create Topic category expected values

✚ Add ✎ Edit ━ Remove Is critical □

ANY

Add condition

Select a condition type ▾ Add

Add Topic category

Pick the new value ▾ Add

Pick the new value
Biota
Boundaries
Climatology / Meteorology / Atmosphere
Economy
Elevation
Environment
Farming
Geoscientific Information

Cancel Save

Update

✚ Add ✎ Edit ━ Remove Is critical □

ALL
ANY
Inland Waters
Oceans
ANY
Transportation
Utilities / Communication

Cancel Save

Create Spatial extent expected values

✚ Add ✎ Edit ━ Remove

ANY

Add condition

Select a condition ty

Add spatial location

[West Long, East Long

Select from map

Save

Draw a rectangle in the map

Draw Clear Cancel Confirm

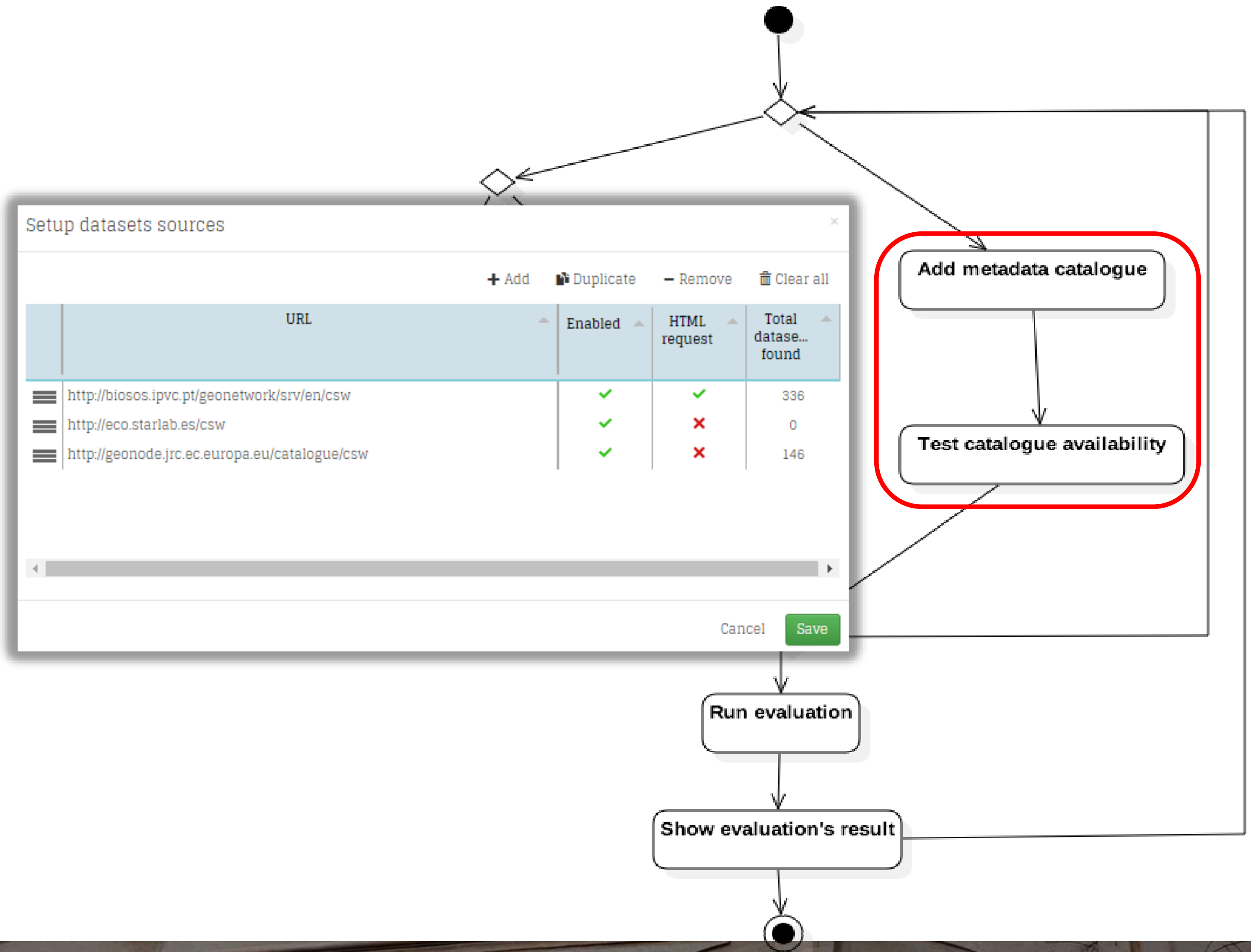
Create Temporal extent expected values

✚ Add ✎ Edit ━ Remove Use as filter □ Is critical □ Cover(%): 100%

ANY

Cancel Save

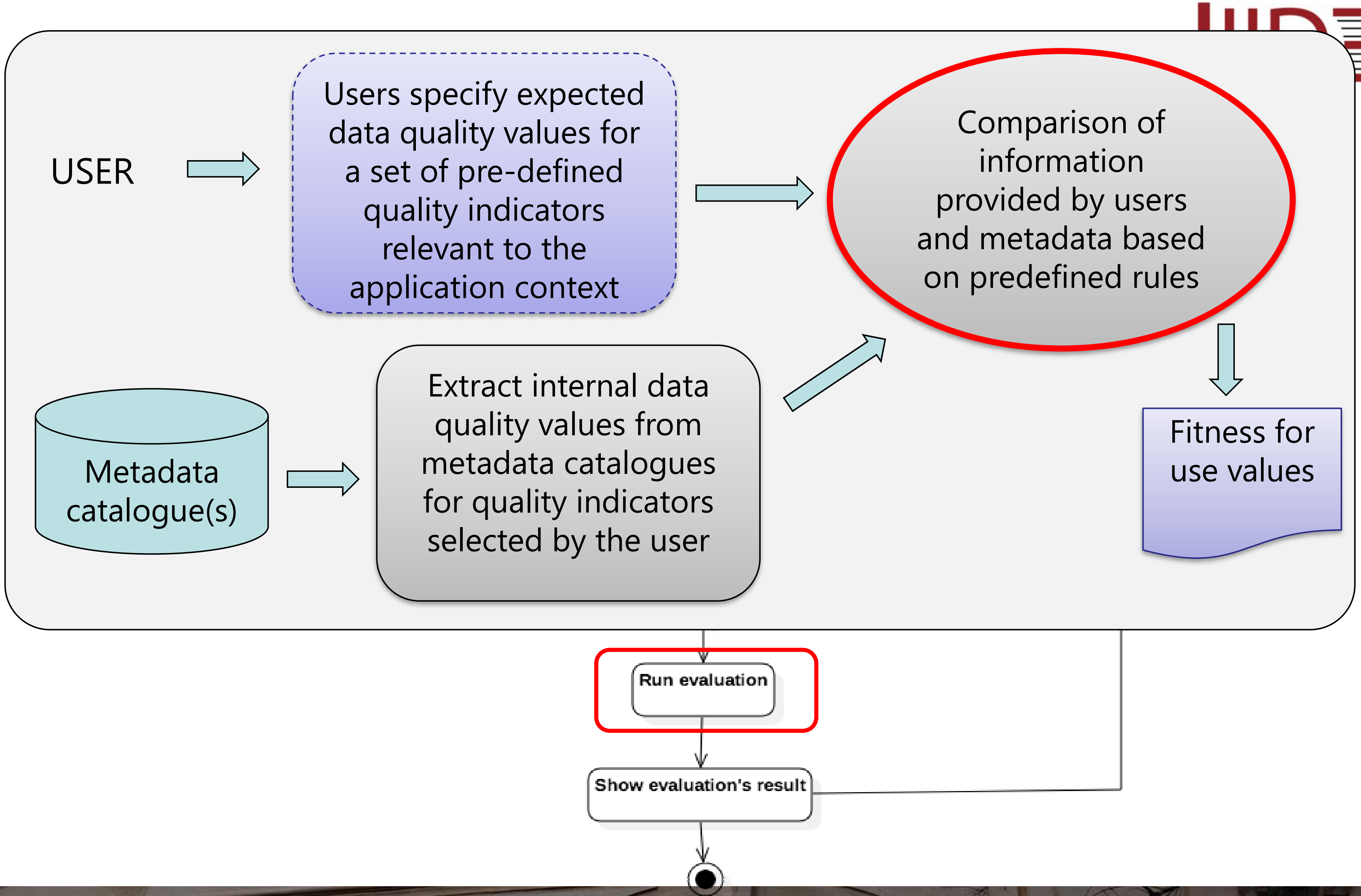
Metadata catalogues



Setup datasets sources

URL	Enabled	HTML request	Total datase... found
http://biosos.ipvc.pt/geonetwork/srv/en/csw	✓	✓	336
http://eco.starlab.es/csw	✓	✗	0
http://geonode.jrc.ec.europa.eu/catalogue/csw	✓	✗	146

Evaluation



USER

Users specify expected data quality values for a set of pre-defined quality indicators relevant to the application context

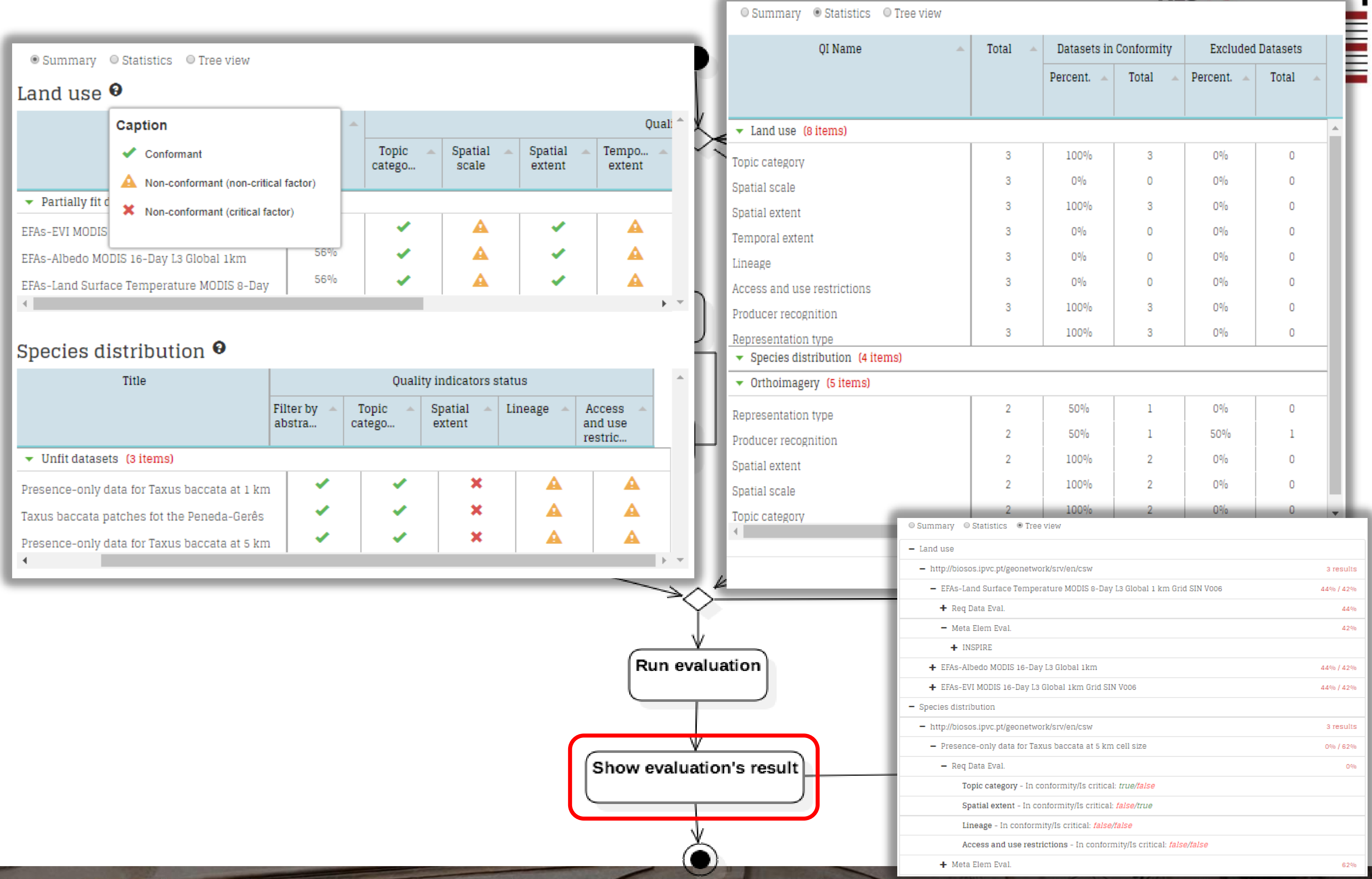
Comparison of information provided by users and metadata based on predefined rules

Metadata catalogue(s)

Extract internal data quality values from metadata catalogues for quality indicators selected by the user

Fitness for use values

Results



○ Summary ○ Statistics ○ Tree view

Land use

Caption

- ✓ Conformant
- ⚠ Non-conformant (non-critical factor)
- ✖ Non-conformant (critical factor)

		Quali			
		Topic catego...	Spatial scale	Spatial extent	Tempo... extent
Partially fit d					
EFAs-EVI MODIS		✓	⚠	✓	⚠
EFAs-Albedo MODIS 16-Day L3 Global 1km	56%	✓	⚠	✓	⚠
EFAs-Land Surface Temperature MODIS 8-Day	56%	✓	⚠	✓	⚠

Species distribution

Title	Quality indicators status				
	Filter by abstra...	Topic catego...	Spatial extent	Lineage	Access and use restric...
Unfit datasets (3 items)					
Presence-only data for Taxus baccata at 1 km	✓	✓	✖	⚠	⚠
Taxus baccata patches fot the Peneda-Gerês	✓	✓	✖	⚠	⚠
Presence-only data for Taxus baccata at 5 km	✓	✓	✖	⚠	⚠

○ Summary ● Statistics ○ Tree view

QI Name	Total	Datasets in Conformity		Excluded Datasets	
		Percent.	Total	Percent.	Total
Land use (8 items)					
Topic category	3	100%	3	0%	0
Spatial scale	3	0%	0	0%	0
Spatial extent	3	100%	3	0%	0
Temporal extent	3	0%	0	0%	0
Lineage	3	0%	0	0%	0
Access and use restrictions	3	0%	0	0%	0
Producer recognition	3	100%	3	0%	0
Representation type	3	100%	3	0%	0
Species distribution (4 items)					
Orthoimagery (5 items)					
Representation type	2	50%	1	0%	0
Producer recognition	2	50%	1	50%	1
Spatial extent	2	100%	2	0%	0
Spatial scale	2	100%	2	0%	0
Topic category	2	100%	2	0%	0

○ Summary ○ Statistics ● Tree view

- − Land use
  - − http://biosos.ipvc.pt/geonetwork/srv/en/csw — 3 results
    - − EFAs-Land Surface Temperature MODIS 8-Day L3 Global 1 km Grid SIN V006 — 44% / 42%
      - + Req Data Eval. — 44%
      - − Meta Elem Eval. — 42%
        - + INSPIRE
    - + EFAs-Albedo MODIS 16-Day L3 Global 1km — 44% / 42%
    - + EFAs-EVI MODIS 16-Day L3 Global 1km Grid SIN V006 — 44% / 42%
- − Species distribution
  - − http://biosos.ipvc.pt/geonetwork/srv/en/csw — 3 results
    - − Presence-only data for Taxus baccata at 5 km cell size — 0% / 62%
      - − Req Data Eval. — 0%
        - **Topic category** - In conformity/Is critical: *true/false*
        - **Spatial extent** - In conformity/Is critical: *false/true*
        - **Lineage** - In conformity/Is critical: *false/false*
        - **Access and use restrictions** - In conformity/Is critical: *false/false*
      - + Meta Elem Eval. — 62%

Thank you for your attention

This work has been carried out within the H2020 project 'ECOPOTENTIAL: Improving Future Ecosystem Benefits Through Earth Observations' (http://www.ecopotential-project.eu). The project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 641762.