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

**Pedro Castro**¹,³,⁴, Joaquim Alonso²,³,⁴, Ivone Martins³,⁴, João Honrado³,⁵, Johannes Peterseil⁶

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

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

Las IDE locales, acercando la información digital a los ciudadanos.
23, 24 y 25 de octubre
Functional workflow

(a) Select metadata catalogue(s)

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

(c) Get metadata of datasets from catalogues

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

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

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

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

<table>
<thead>
<tr>
<th>TDS</th>
<th>DSc1</th>
<th>...</th>
<th>DScn</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>eVQ1</td>
<td></td>
<td>eVQn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Metadata element (MF):

| ME1 | MV11 | ...
|-----|------|-----|
|     |      | ...

Thematic category (TC):

<table>
<thead>
<tr>
<th>ThC1</th>
<th>...</th>
<th>ThCn</th>
</tr>
</thead>
</table>
| ME1  | MV11| ...
|      |     |      |

Dataset from catalogue(s) (DSc):

<table>
<thead>
<tr>
<th>DSc1</th>
<th>...</th>
<th>DScn</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>eVQ1</td>
<td>eVQn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quality indicator (QI):

<table>
<thead>
<tr>
<th>QL</th>
<th>eVQ1</th>
<th>eVQn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Metadata Quality element (MQE):

| MOE1 | mV11 | ...
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Datasets from catalogue(s) (DSc):

<table>
<thead>
<tr>
<th>DSc1</th>
<th>...</th>
<th>DScn</th>
</tr>
</thead>
<tbody>
<tr>
<td>QL</td>
<td>eVQ1</td>
<td>eVQn</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

23, 24 y 25 de octubre
Actions workflow

Add Thematic Category

Configure Quality Indicator

Add/Edit Metadata Value

Add metadata catalogue

Test catalogue availability

Run evaluation

Show evaluation's result
Thematic category

Allows to define different datasets that are targeted for search for the given application context
Users specify expected data quality values of available quality indicators for each targeted dataset.
Metadata catalogues

Add metadata catalogue

Test catalogue availability

Run evaluation

Show evaluation's result

Las IDE locales, acercando la información digital a los ciudadanos.
23, 24 y 25 de octubre
Evaluation

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

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

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

Fitness for use values

Run evaluation

Show evaluation's result

Las IDE locales, acercando la información digital a los ciudadanos. 23, 24 y 25 de octubre
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.