Analysis of the Quality Metadata in GEOSS Clearinghouse

SEVILLANO Eva¹, DÍAZ Paula², NINYEROLA Miquel¹, MASÓ Joan², ZABALA Alaitz¹, PONS Xavier¹

¹ UAB Universitat Autònoma de Barcelona.
² CREAT Centre for Ecological Research and Forestry Applications.
Objectives

• To get a first analysis of the data quality in the Clearinghouse

• Analyze the quality contained in the metadata (ISO 19115)
  – Quality indicators
  – Lineage
  – Usage

• Start building components for the GEO Portal
  – Quality Broker
  – Quality searcher
  – Quality visualization
• Harvest all XML documents, ISO 19115. (October 2011)
Methodology

- Massive extraction of MD quality elements
  - Quality indicators
  - Lineage
  - Usage
Overall Results

- Total metadata records in the Clearinghouse – 97203
- Total number of quality indicators – 52187
- Metadata records with quality indicators – 19107
- Metadata records with lineage – 10899 (9261 process, 3771 source)
- Metadata with usage – 1226
19.66% Metadata records with quality indicators
   – 2.7 quality indicator per metadata record
1. Quality indicators

- 19.66% Metadata records with quality

Generic Quality Indicators

- Temporal Accuracy: 6.81%
- Completeness: 35.71%
- Consistency: 19.78%
- Positional Accuracy: 37.20%
- Thematic Accuracy: 0.50%
• 19.66% Metadata records with quality
  – 2.7 QI/MD
Quality indicators – Comparison
Clearinghouse - IDEC

Generic Quality Indicators

- Thematic Accuracy: 6.81%
- Completeness: 35.71%
- Temporal Accuracy: 0.50%
- Positional Accuracy: 37.20%
- Consistency: 19.78%

Quality Indicators in IDEC Metadata

- Temporal Accuracy: 0.06%
- Logical Consistency: 0.02%
- Thematic Accuracy: 2.60%
- Completeness: 0.46%
- Positional Accuracy: 95.38%
Quality indicator result

14.18%
(3669 QI)
mainly conformance to INSPIRE

85.8%
(22275 QI)

0.02%
(5 QI)

19115-2 Extension for "per pixel" quality
Quality indicators - Quantitative measures

Quality elements:
- Absolute External Positional Accuracy
- Accuracy Of A Time Measurement
- Completeness Commission
- Conceptual Consistency
- Domain Consistency
- Grided Data Positional Accuracy
- Non Quantitative Attribute Accuracy
- Quantitative Attribute Accuracy
- Relative Internal Positional Accuracy
- Temporal Consistency

Number of quality elements:
- Complete value
- Declare value
- Quantitative type
Quality indicators - Qualitative

Quality elements - Conformance measures

- Number of quality elements
- Absolute External Positional Accuracy
- Accuracy Of A Time Measurement
- Completeness Commission
- Conceptual Consistency
- Grided Data Positional Accuracy
- Non Quantitative Attribute Accuracy
- Quantitative Attribute Accuracy
- Relative Internal Positional Accuracy
- Temporal Consistency

Legend:
- Conformance to specification
- Declare conformance
- Conformance type
• Clearinghouse record ID: 273234, 273232, 273233, 273235, 273236
• Only 5 records use this. Bad news for visualizing data + quality maps

• Title: OMNO2e:OMI Column Amount NO2:ColumnAmountNO2CS30

  <gmd:DQ_QuantitativeAttributeAccuracy>
    <gmd:measureDescription>
      <gco:CharacterString>The 'version 003’ product is the second public release. It is based on improved radiance calibration. For details, please see document: http://disc.sci.gsfc.nasa.gov/Aura/OMI/OMTO3e_v003.shtml</gco:CharacterString>
    </gmd:measureDescription>
    <gmd:result><gmi:QE_CoverageResult>
      <gmi:spatialRepresentationType><gmd:MD_SpatialRepresentationTypeCode codeList="/resources/codeList.xml#MD_SpatialRepresentationTypeCode" codeListValue="grid">grid</gmd:MD_SpatialRepresentationTypeCode></gmi:spatialRepresentationType>
      <gmi:resultFile gco:nilReason="missing" />
      <gmi:resultFormat>
        <gmd:MD_Format>
          <gmd:name><gco:CharacterString>CF-netCDF</gco:CharacterString></gmd:name>
        </gmd:MD_Format>
      </gmi:resultFormat>
    </gmi:QE_CoverageResult></gmd:result>
  </gmd:DQ_QuantitativeAttributeAccuracy>
2. Lineage

If `count(source) + count(processStep) = 0` and 
`DQ_DataQuality.scope.level = 'dataset' or 'series'` then 
statement is mandatory

"source" role is mandatory if 
`LI_Lineage.statement` and "processStep" 
role are not documented

"processStep" role is mandatory if 
`LI_Lineage.statement` and "source" role are 
not documented

---

www.geoviqua.org
2. Lineage
Clearinghouse record ID 131007 (simplified)

- **Compile survey input data from the best and most current survey records.**
  - BLM database of the index to all official (microfilm, CD, other) BLM survey records.
  - USFS survey records.
  - Private land surveyor records.

- **Compile listings of known locations of PLSS corners.**
  - USGS topographic quadrangles and other sources.
  - USC&GS published coordinate data.
  - NGS published coordinate data.
  - BLM global positioning Data.
  - USFS global positioning data.

- **Coordinates of control stations are entered into a control data base with associated reliabilities.**

- **Topologically correct GIS coverages are modified to use FGDC compliant naming conventions and then loaded into the LSI database. These layers can then be downloaded as shapefiles through the LSI website.**

- **GCDB Data was downloaded for Kiowa and Cheyenne Counties, Colorado.**
  - C:\gis_data\sand\zipped\kiowa\township.shp.xml

- **Metadata imported and data was exported from regions format to shapefile format**

- **Dataset copied.**
  - C:\gis_data\sand\data\basedata\plss\ck_gcdb_region_township

- **Source Contribution: Survey data in the form of official (microfilm, CD, other) survey and BLM, abstracted into a vector digital format online.**

- **Source Contribution: Survey and control data from the Cartographic Feature File (CFF) data set.**

- **Source Contribution: Digitized control data from standard topological quadrangle sheets.**
6.02% metadata records (5851) contain **direct list of the data sources**.  
- 1.85% (1798) with temporal extent

- Gives credit (attribution, and eventually some trust on them)

- If quality indicators are not provided for the dataset, the quality indicators from sources can be a clue.
• 8.26% metadata records (8035) contain the **direct list of the processes** without sources
  – 292 (0.30%) contain date
• With the order of these processes.
• If quality indicators are not provided for the dataset, it’s difficult to infer resource quality with only a process list
1.26% metadata records (1226) with more complete provenance process.

How and when the data sources where used

If quality indicators are not provided for the dataset, we can deduce which sources have more influence in the quality of the final result.
• LI_ProcessStep includes a LE_Processing that has a runTimeParameters attribute that allows us describing the exact list of parameters used in the execution.

• There is a citation of the algorithm used (LI_Algorithm).

• All these extensions were done for the benefit of the EO gridded data, but there are not in the Clearinghouse.

• We can completely evaluate the quality of the resulting product if we know the uncertainties that sources have in their metadata (sourceMetadata citation in LI_Source).

From ISO 19115-2:2009 shown for informative purposes only
3. Usage - User feedback

- There is one small entry for user feedback in the current ISO-19115:
- MD_Usage
  - Brief description of ways in which the resource is currently or has been used
• There are **1.2% (1133)** entries
  – SpecificUsage and
  – UserContactInfo, only

• All made by the **same institution!**:
  – Landesvermessung und Geobasisinformation Brandenburg (LGB)
  – Tel +49-331-8844-123, Fax. +49-331-8844-16123
  – Heinrich-Mann-Allee 103, Potsdam, Brandenburg 14473, Deutschland
  – kundenservice@geobasis-bb.de
Conclusions

- There are many different kinds of quality indicators
  - There is a lack of a complete description of values provided (no units, missing measure name, missing evaluation method)
- Quality coverage results (by pixel) are almost inexistent and the link is not there
- Lineage information is rich in many records, some with more than 100 entries in source or ProcessSteps
- We have usage examples -> Feedback
- Current data is enough to demonstrate search and visualization with some limitations. Good for GeoViQua.

- Next steps:
  - Assess the Quality of Quality Metadata?
  - Extend this analysis to other capacity catalogues integrated in the EuroGEOSS Broker
Thank you!  
Danke!  
Grazie!  
Merci!  
Ευχαριστίες!  
Vő'aka!  
Tak!  
Díky!  
Tänan!  
Kiitos!

Diolch!  
Bedankt!  
Köszönöm!  
Αčίů!  
Благодаря!  
Dzięki!  
Mulţumiri!  
Хваля!  
Tack!  
Teşekkürler!

Спасибі!  
Спасибо!  
Obrigado!  
Takk!  
Gràcies!  
Gracias!