Data Quality in Earth System Sciences - The GeoKur Approach

Christin Henzen | christin.henzen@tu-dresden.de
Geoinformatics, Technische Universität Dresden

OGC Data Quality Workshop, June 2021
Data Quality in Earth System Sciences - The GeoKur Approach

**Research in Environmental Sciences**

- Digitalisation is established
- Culture to publish/share data & tools partly exist
- Common standards for metadata, data schema, formats etc.
- Various repositories available

**But, in short:**

- Too many special solutions
- Lack of management plans, software and guidance for research data
Data Quality in Earth System Sciences - The GeoKur Approach

Project GeoKur

Data consumer’s perspective

2 Use cases on Human-nature interactions
Use Cases from Different Perspectives

Data consumer’s perspective

Not available
[MapSPAM IFPRI]

"The accuracy for the final data delivery is assessed to be good."
[Crop production (apro_cp) EUROSTAT]

Forest loss 2000-2012: 87.0 (2.8)
[Global Forest Change 2000–2019 University of Maryland]
### Survey

**Availability, relevance and needs related to data quality information**

Based on your experience, how often is information about these quality elements available (consumer) or provided (producer)?

<table>
<thead>
<tr>
<th>Positional accuracy</th>
<th>Temporal quality</th>
<th>Thematic accuracy</th>
<th>Completeness</th>
<th>Logical consistency</th>
<th>Usability element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very often</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Often</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sometimes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Never</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Don’t know</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Are the provided information structured using a standard?

- Choose one of the following answers

Yes. Please specify: [ ]

If information for a quality element is available, where do you obtain (consumer) or provide (producer) information from?

<table>
<thead>
<tr>
<th>Positional accuracy</th>
<th>Temporal quality</th>
<th>Thematic accuracy</th>
<th>Completeness</th>
<th>Logical consistency</th>
<th>Usability element</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Associated publication</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Associated report</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Directly from/within the data (e.g. grid based)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Geo-Kur Approach

- GeoDashboards
  - Dynamic DMPs as living documents
  - Collaboratively used
  - Managed in proper tools
  - Still lacks clear guidance and
  - Community-specific and
  - Adapted DMP templates

How can we improve future DMPs for earth system sciences?

Geo-Dashboard Concept

https://doi.org/10.5194/agile-giss-2-25-2021
Data Management Plans

Data provider's perspective

Recommendations for Future Data Management Plans in Earth System Sciences
https://doi.org/10.5194/agile-giss-2-31-2021
http://doi.org/10.5281/zenodo.4916856
Generic Linked Data Quality Concept

1. Create an initial set of project-specific quality measures
2. Add dataset-specific quality measures
3. Upload metadata and data

TRIPLE STORE

Data Quality Measures

RESEARCH DATA MANAGEMENT PLATFORM

DATA CATALOGUE

Metadata including Quality

Quality-related Data

Geodata
Metadata Profile

Developer’s perspective

Recommendations for Developing a Metadata Profile for Earth System Science Data

Authors: Christin Henzen, Arne Rümmler, Michael Wagner
Affiliation: Geoinformatics, Technische Universität Dresden
Publication date: June, 2021

Executive Summary
Most Earth System Science (ESS) research projects are data driven and/or produce data sets as main results. Metadata management is core to support discovery and reuse of such results, and ultimately to allow for reproducibility of the research findings. Thus, ensuring acquisition and provision of meaningful and quality assured metadata should become an integral part of such projects. Here, choosing a suitable metadata schema and/or developing a proper metadata profile is a relevant task at the beginning of each project. Building on available, well-known and well-used, often standardized formats and schemas is strongly recommended. This document serves as guideline for researchers, who need to manage metadata with certain project-specific requirements. It guides metadata managers to create a suitable metadata schema, which meets the project-specific requirements. Metadata consumers will get an overview about relevant aspects, basic principles and standards to understand available metadata and to aggregate or summarize metadata for their purposes.

http://doi.org/10.5281/zenodo.4916698
### Generic Linked Data Quality Concept

**Completeness Omission as Number of Missing Items**

- **value of quality metric:** 14
- **ground truth dataset:**
- **confidence term:**
- **confidence value:**
- **thematic representativity:** supp_info
- **spatial representativity:** Global
- **temporal representativity:** 1819-2021
- **name of quality source:** MetadataFromGeodata Extraction Tool
- **type of quality source:** software
- **link to quality source:** https://github.com/GeoinformationSystems/MetadataFromGeodata
### Generic Linked Data Quality Concept

**Completeness Ommission as Number of Missing Items**

- value of quality metric: 14
- ground truth dataset:
- confidence term:
- confidence value:
- thematic representativity: supp_info
- spatial representativity: Global
- temporal representativity: 1819-2021
- name of quality source: MetadataFromGeo
- type of quality source: software
- link to quality source: [https://github.com/](https://github.com/)

**Quantitative Attribute Accuracy as Attribute Value Uncertainty at 99% Significance Level**

- value of quality metric: \( R^2 = 0.71-0.91; \; \text{RMSE} = 231-307 \)
- ground truth dataset: CDL2010 dataset
- confidence term:
- confidence value:
- thematic representativity:
- spatial representativity: Global
- temporal representativity:
- name of quality source: A cultivated planet in 2010 – Part 2: The global gridded ag
duction maps
- type of quality source: Publication
- link to quality source: [https://doi.org/10.5194/essd-12-3545-2020](https://doi.org/10.5194/essd-12-3545-2020)
**Generic Linked Data Quality Concept**

**Completeness Omission as Number of Missing Items**

<table>
<thead>
<tr>
<th>value of quality metric:</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>ground truth dataset:</td>
<td></td>
</tr>
<tr>
<td>confidence term:</td>
<td></td>
</tr>
<tr>
<td>confidence value:</td>
<td></td>
</tr>
<tr>
<td>thematic representativity:</td>
<td>supp_info</td>
</tr>
<tr>
<td>spatial representativity:</td>
<td></td>
</tr>
<tr>
<td>temporal representativity:</td>
<td></td>
</tr>
<tr>
<td>name of quality source:</td>
<td></td>
</tr>
<tr>
<td>type of quality source:</td>
<td></td>
</tr>
<tr>
<td>link to quality source:</td>
<td></td>
</tr>
</tbody>
</table>

**Quantitative Attribute Accuracy as Attribute Value Uncertainty at 99% Significance Level**

<table>
<thead>
<tr>
<th>value of quality metric:</th>
<th>$R^2 = 0.71-0.91$; $\text{RMSE} = 231-307$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ground truth dataset:</td>
<td>CDL2010 dataset</td>
</tr>
<tr>
<td>confidence term:</td>
<td></td>
</tr>
<tr>
<td>confidence value:</td>
<td></td>
</tr>
</tbody>
</table>

**Absolute External Positional Accuracy as Bias**

- The data in the WDPA and OECM database come from a wide range of sources, often using different scales and techniques to generate their data. This results in variation in accuracy and resolution.
Management and Extraction Tools

Java Tool: MetadataFromGeodata
https://doi.org/10.5194/agile-giss-2-41-2021
**Why do we need guidelines for quality in GeoKur?**

<table>
<thead>
<tr>
<th>Guidance</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discipline-specific DMP Template</td>
</tr>
<tr>
<td></td>
<td>Metadata Profile</td>
</tr>
<tr>
<td></td>
<td>Generic Linked Data Quality Concept</td>
</tr>
<tr>
<td></td>
<td>Extraction Tool</td>
</tr>
<tr>
<td></td>
<td>Geo-dashboard Concept</td>
</tr>
</tbody>
</table>