Arne Rümmler (arne.ruemmler@tu-dresden.de), Christin Henzen (christin.henzen@tu-dresden.de)

Geospatial Data Quality in Research Data Infrastructures

ESIP Information Quality

November 2021

Our Project and Use Cases

UFZ HELMHOLTZ Centre for Environmental Research

Modelling **human-environment relations** with geospatial time-series of global land use data

- Quality (and provenance) information needs of data users
- Best practices on determining fitness for use of geospatial dataset

Our Project and Use Cases

UFZ HELMHOLTZ Centre for Environmental Research

iDiv

Modelling **human-environment relations** with geospatial time-series of global land use data

- Quality (and provenance) information needs of data users
- Best practices on determining fitness for use of geospatial dataset

Developing **consistent data products** on recent global dynamics in multiple land-use variable

- Methods and scripts for validating and quality-improving global land-use data
- Best practices for land-use data quality assurance

Our Project and Use Cases

UFZ HELMHOLTZ Centre for Environmental Research

TECHNISCHE UNIVERSITÄT DRESDEN

iDiv

Modelling **human-environment relations** with geospatial time-series of global land use data

Concepts, guidance, tools

Developing **consistent data products** on recent global dynamics in multiple land-use variable

- Quality (and provenance) information needs of data users
- Best practices on determining fitness for use of geospatial dataset

- Methods and scripts for validating and quality-improving global land-use data
- Best practices for land-use data quality assurance

Geospatial Data Quality

In Research Data Infrastructures

„We do not create data, just maps.“

Geospatial Data Quality

In Research Data Infrastructures

„We do not use a repository. We store our data and code on GitHub.“

„We do not create data, just maps.“

Geospatial Data Quality
In Research Data Infrastructures

„We do not use a repository. We store our data and code on GitHub.“

„We do not create data, just maps.“

The quality is assessed to be good.

„... but we need spatially explicit quality (and provenance) information.“

Geospatial Data Quality
In Research Data Infrastructures

„We do not use a repository. We store our data and code on GitHub.“

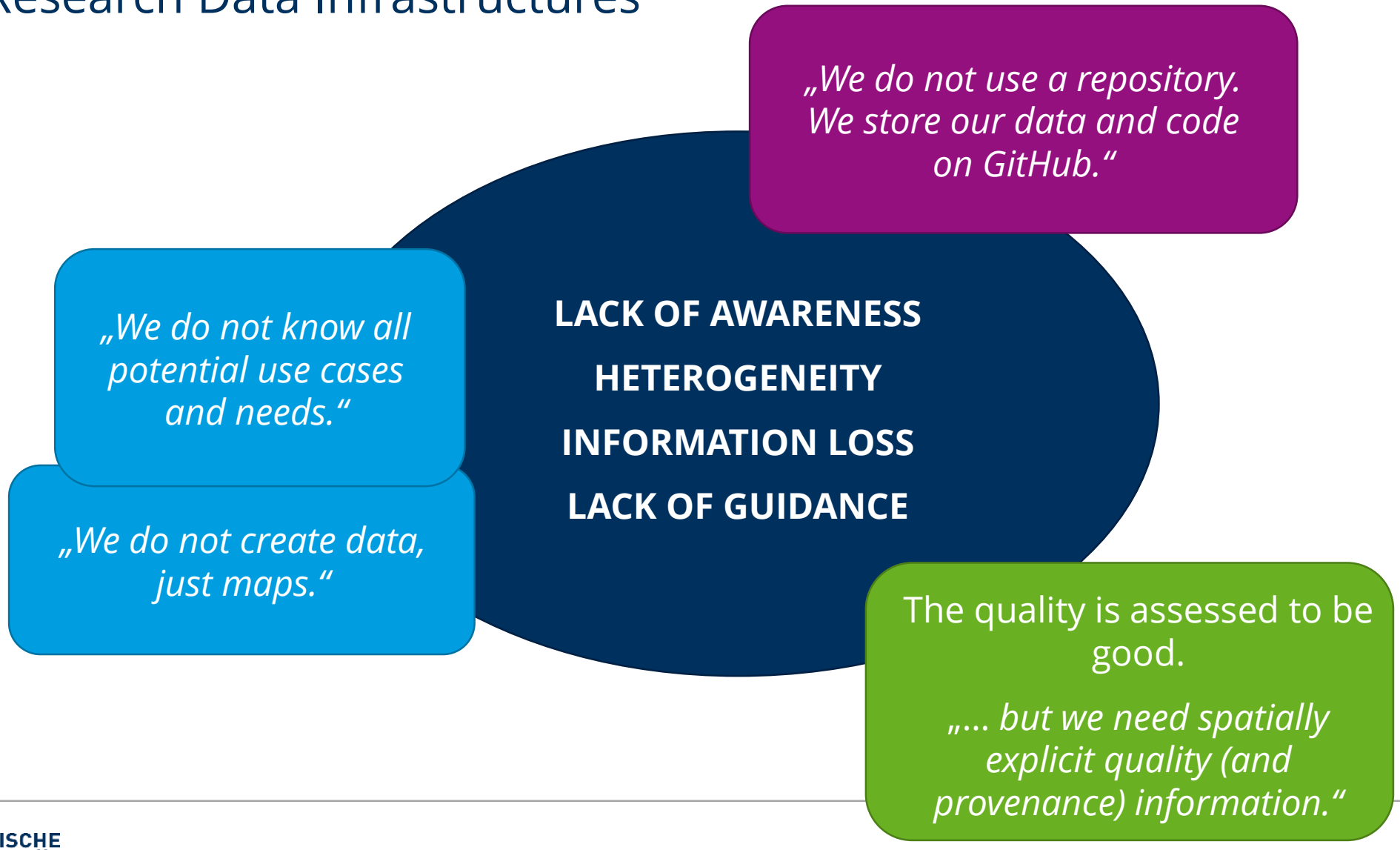
„We do not know all potential use cases and needs.“

„We do not create data, just maps.“

The quality is assessed to be good.

„... but we need spatially explicit quality (and provenance) information.“

Geospatial Data Quality
In Research Data Infrastructures



Geospatial Data Quality
In Research Data Infrastructures

Software + Repositories

Geoinformation systems

Catalogues

(Spatial) Web services

Tracking & Extraction Concepts and Tools; Catalogue & Registry

Spatial

Temporal

Thematic

LACK OF AWARENESS

HETEROGENEITY

INFORMATION LOSS

LACK OF GUIDANCE

DMPs & Guidance Patterns

Metadata Profile, Management concepts,

Quality

Provenance

Content + Analysis

Metadata

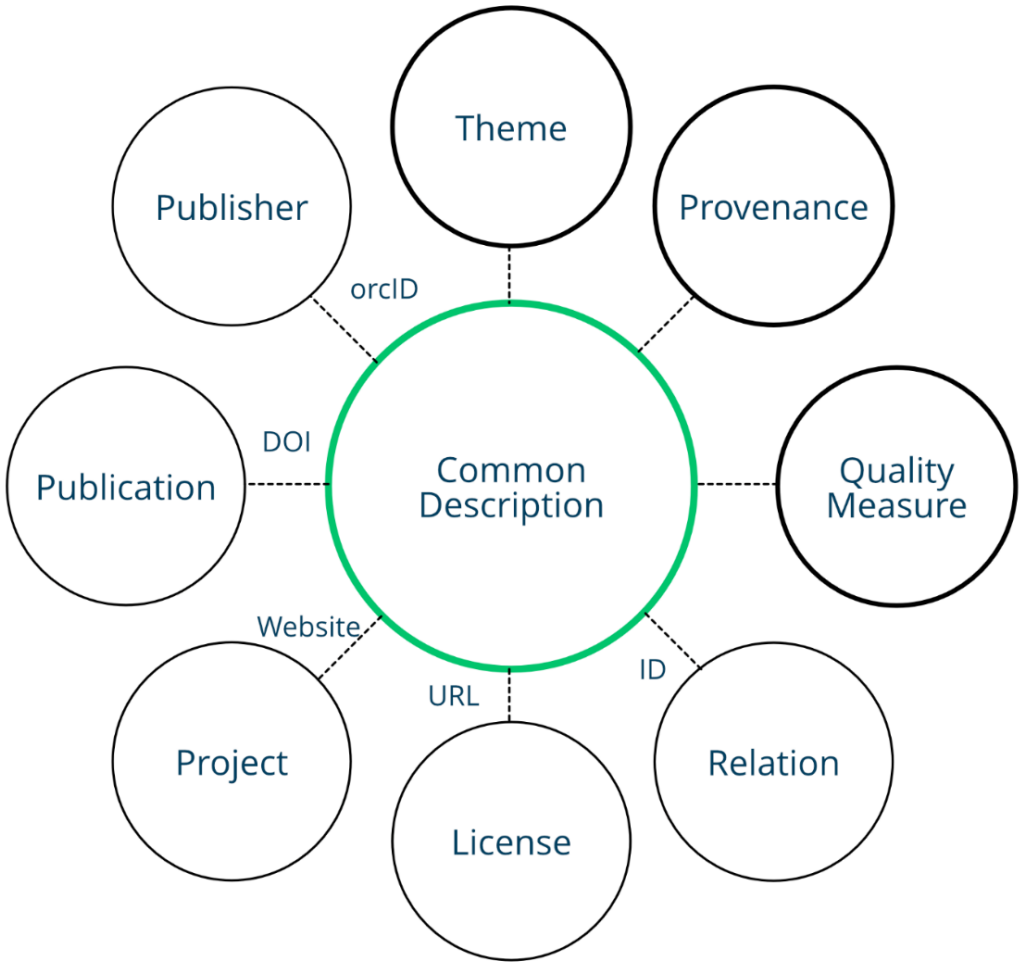
Guidance Patterns for DMPs

Examples for Earth System Science

Table 1: Data management in a Geo-Catalogue

Name	Example
Guidance name	Data Management in a Geo-Catalogue
Motivation / Intent / Aim of guidance	In Earth System Sciences, Geo-(metadata) catalogues are used t manage geospatial data and related metadata by providing discipline specific user interfaces, e.g. spatial filter and search menus, and APIs e.g. for spatial requests.
Recommended activities	Manage geospatial data in a Geo-catalogue directly from the projec beginning. Whenever possible, use an existing catalogue, e.g. a institutional. When not having the option to use an existing Geo-catalogue, you ca choose from a list of various existing (open-source) catalogues.
Example / Use	The BMBF project GeoKur aims to support the curation and qualit

Metadata Profile with Provenance & Quality



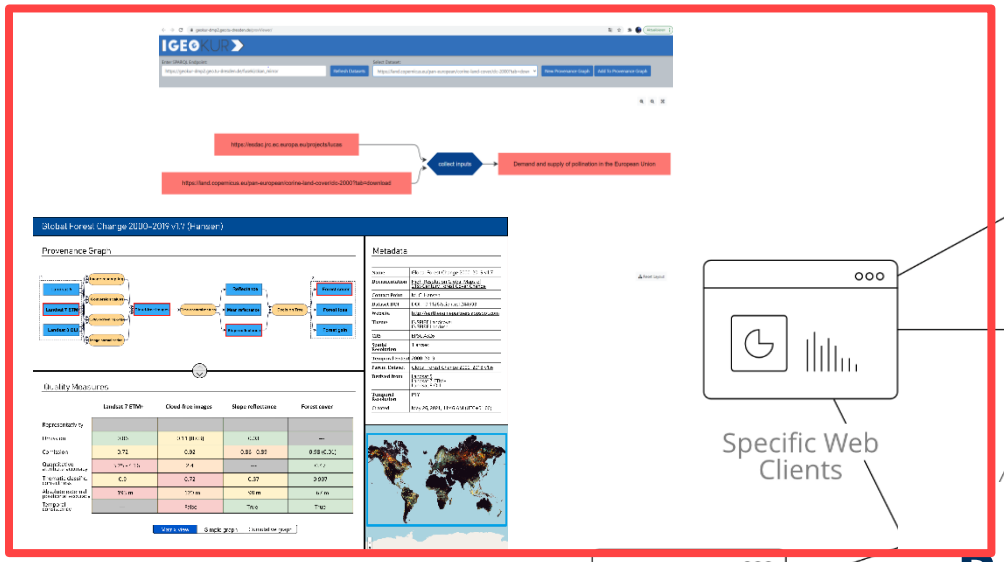
Recommendations for Developing a Metadata Profile for Earth System Science Data

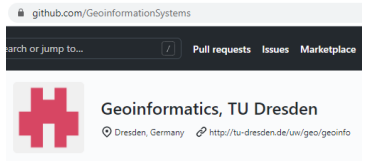
Authors: Christin Henzen, Arne Rümmler, Michael Wagner
Affiliation: Geoinformatics, Technische Universität Dresden
Publication date: June, 2021

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,

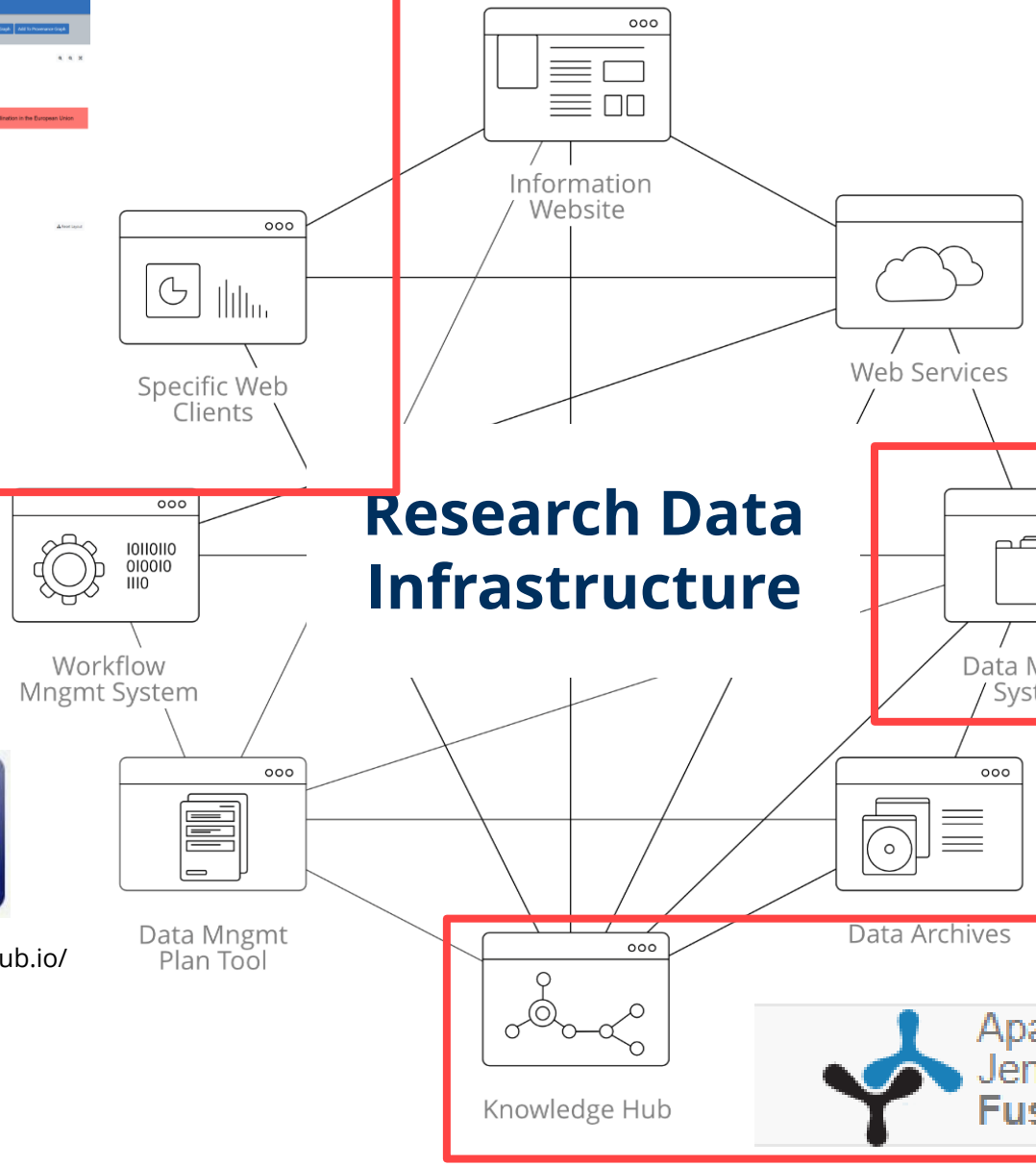
GeoKur Research Data Infrastructure





RDMO

https://rdmorganiser.github.io/



Research Data Infrastructure

OGC®
Making location count.

www.opengeospatial.org

ckan

ckan.org

Apache Jena Fuseki

jena.apache.org

Interaction between Components

- Implement metadata schema
- User interface for reviewing / managing metadata entries
- Links to Web clients
- User interface for reviewing / managing quality metrics

- Provide (CKAN) metadata as linked open data (LOD)
- Provide interface for Web clients
- Manage definitions of quality metrics

- Visualize LOD that adheres to standards
- W3C Provenance ontology (PROV-O)
- W3C Data quality vocabulary (DQV)

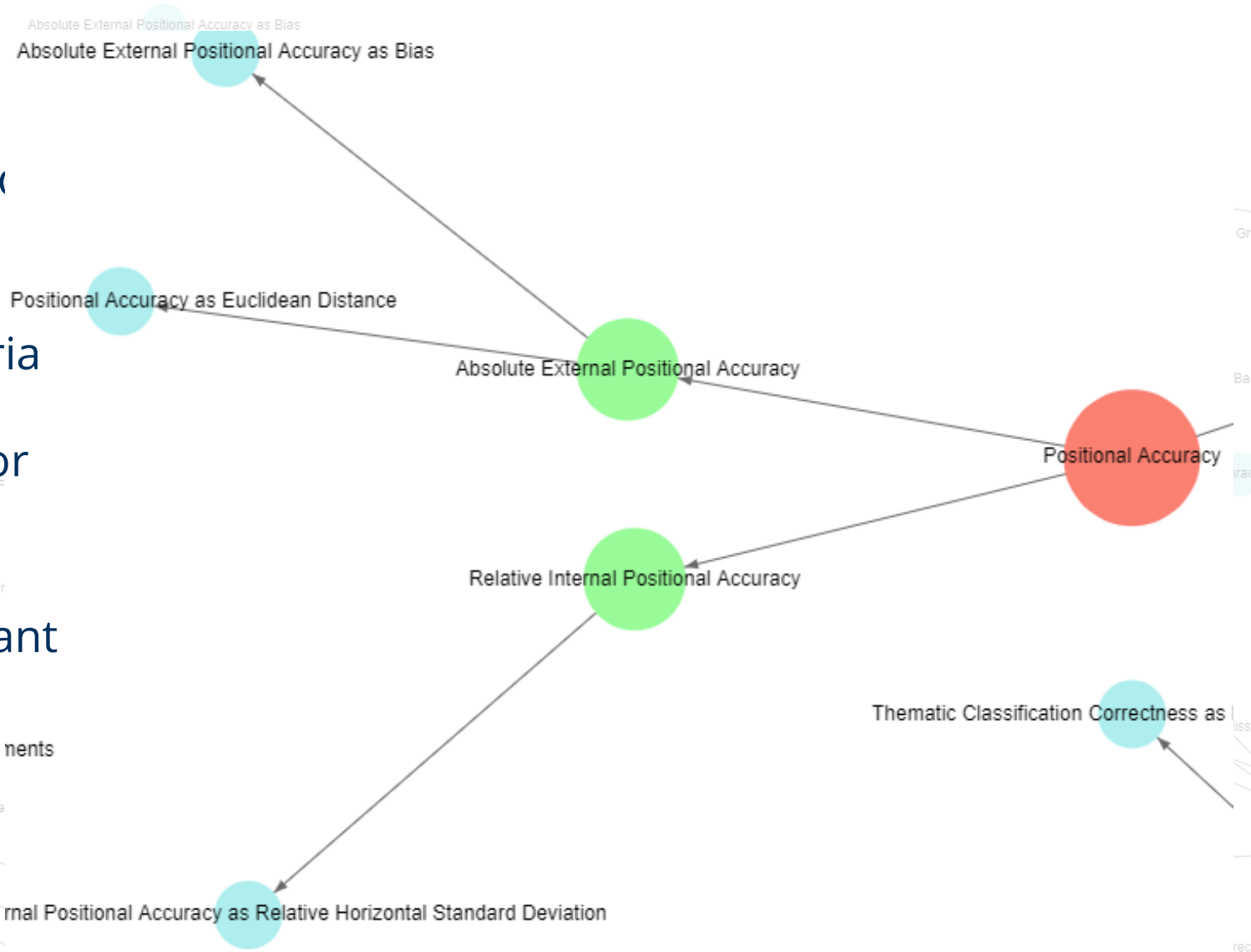
Quality Terms

DQV Metric: “Represents a standard to measure a quality dimension.”

DQV Dimension: “Represents criteria relevant for assessing quality. Each quality dimension must have one or more metric to measure it.”

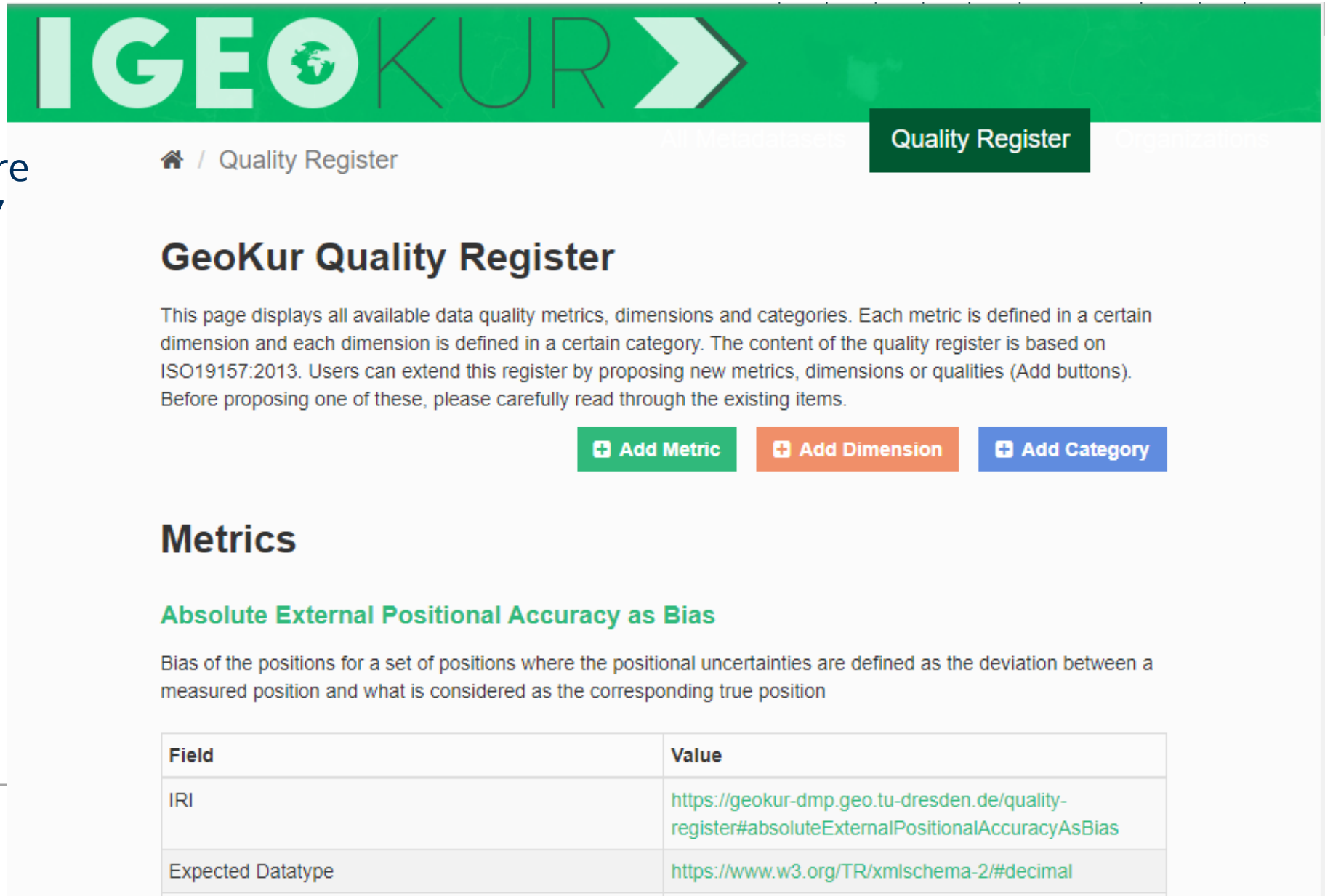
DQV Category: “Categories are meant to systematically organize dimensions.”

Currently, our Triplestore manages 62 metric definitions.



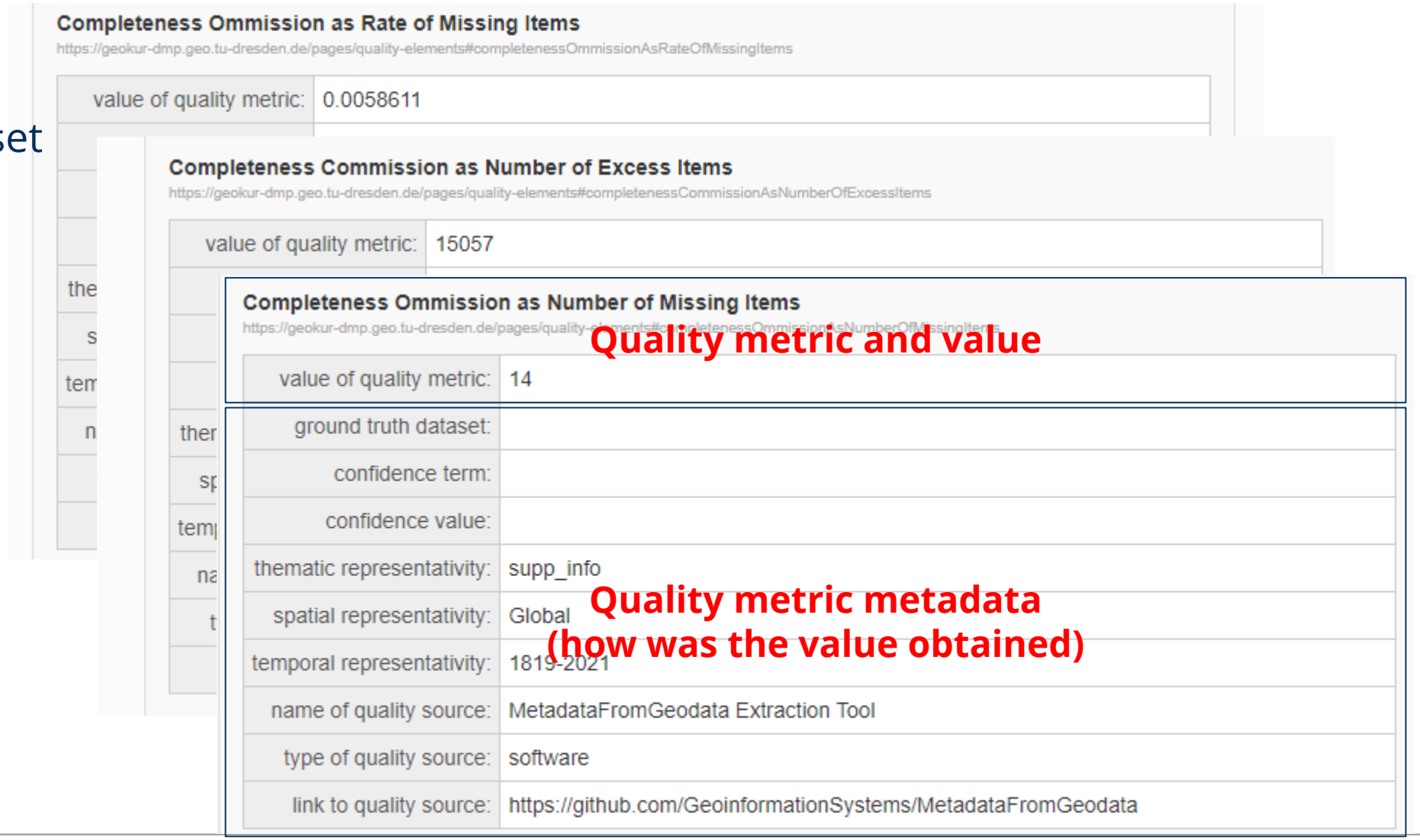
CKAN - DMS and Quality Register

- Available metrics are based on ISO19157
- Users can propose new metrics, dimensions and categories



CKAN - DMS and Quality Register

Each Meta dataset can reference multiple quality metrics



CKAN - DMS and MD Schema

Link to Web client

Profile description on Zenodo: [Recommendations for Developing a Metadata Profile for Earth System Science Data | Zenodo]

The dataset's metadata (subset)

[dataset] WDPA World Database on Protected Areas v1.6

The World Database on Protected Areas (WDPA) is the most comprehensive global database of marine and terrestrial protected areas. Protected areas exist under the authority of diverse governance actors, including indigenous peoples, local communities, private actors, governments, and combinations of these. The WDPA is updated on a monthly basis.

Dataset Provenance | Metadata as JSON | Metadata as RDF-Turtle

Data and Resources

wdpa_oct2021_public_csv.zip — Explore

Areas of Biodiversi... | OECMs | Protected Areas | other effective are...

Additional Info

Field	Value
Identifier	protected-areas
Documentation	https://wdpa.s3-eu-west-1.amazonaws.com/WDPA_Manual/English/WDPA_WDOECM_Manual_1_6.pdf
Contact Point	protectedareas@unep-wcmc.org
Contact Point - ORCID iD or e-mail adress	
Dataset DOI	
Information Website	https://www.protectedplanet.net/en/thematic-areas/wdpa?tab=WDPA
Theme / Vocabulary / Ontology	https://inspire.ec.europa.eu/theme/ps, https://agrovoc.fao.org/browse/agrovoc/en/page/c_37952
Coordinate Reference System	http://www.opengis.net/def/datum/EPSG/0/6326
Spatial Resolution	
Spatial Resolution Measured	
Temporal Coverage	1819-01-01 to 2021-04-30
Temporal Resolution	P1M
Data Quality Metric	
Format Consistency as Physical Structure Conflict Rate https://geokur-dmp.geo.tu-dresden.de/pages/quality-elements#formatConsistencyAsPhysicalStructureConflictRate	

TECHNISCHE UNIVERSITÄT DRESDEN

DRESDEN concept

Enter SPARQL Endpoint:

https://geokur-dmp2.geo.tu-dresden.de/fuseki/ckan_mirror

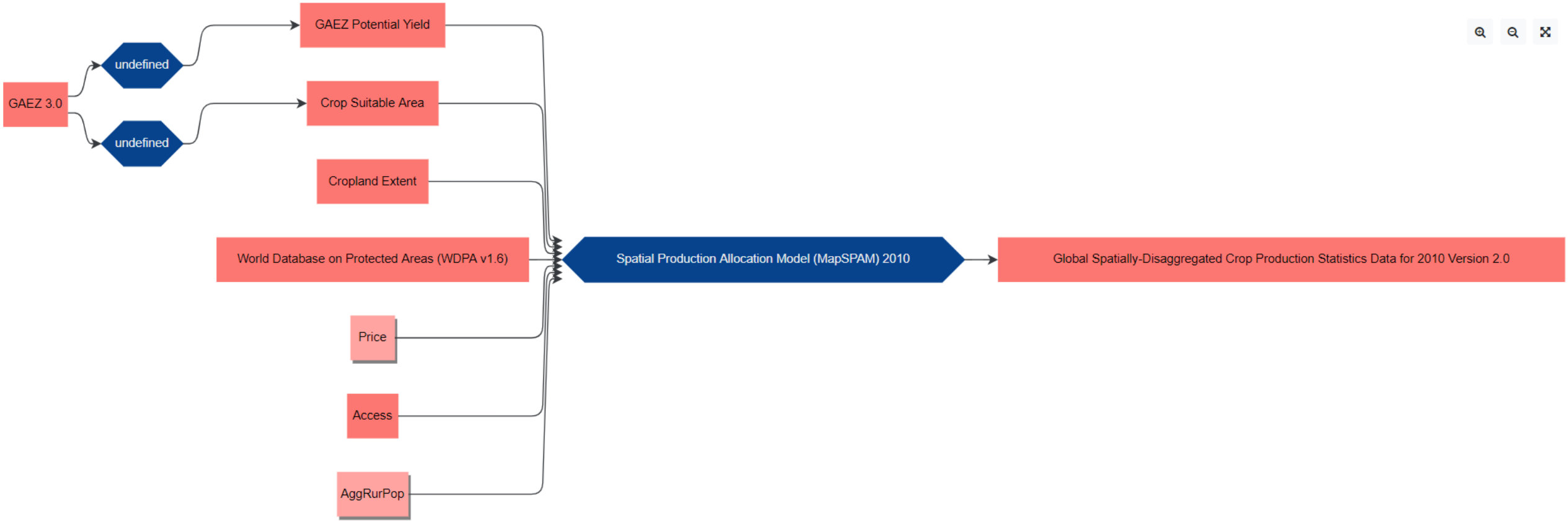
Refresh Datasets

Select Dataset:

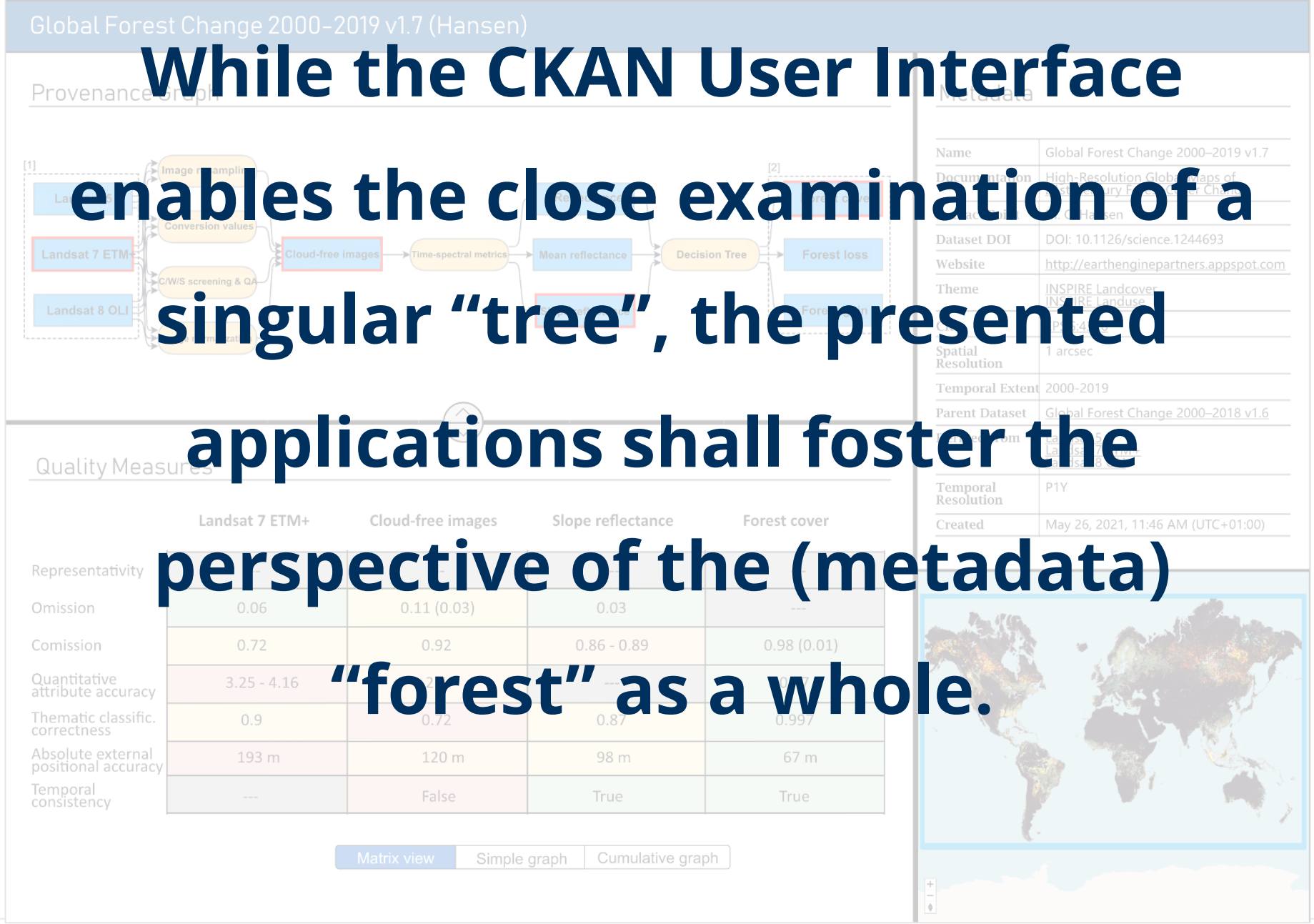
Global Spatially-Disaggregated Crop Production Statistics Data for 2010 Version

New Provenance Graph

Add To Provenance Graph



https://geokur-dmp2.geo.tu-dresden.de/provViewer/?endpoint=https://geokur-dmp2.geo.tu-dresden.de/fuseki/ckan_mirror&dataset=https://geokur-dmp.geo.tu-dresden.de/dataset/b0e5c26c-7762-4f99-8234-b793ce13d19c



While the CKAN User Interface enables the close examination of a singular “tree”, the presented applications shall foster the perspective of the (metadata) “forest” as a whole.

Your feedback is welcome!

Dear data user,

Information on data quality and provenance is essential to determine its suitability for specific geodata purposes (fitness for use). Within the GeoKur project we developed a survey to better grasp both the relevance of data quality and provenance information from the perspective of a data user and the availability and accessibility of such information.

In summer 2021 we sent out a similar, longer survey. The response rate was relatively low, apparently due to its complexity and length. However, we received a great deal of feedback that data quality is an interesting and relevant criterion! For this reason we have substantially revised and simplified the survey. The new version will only take 5 to 10 minutes.

https://www.soscisurvey.de/dataquality/

Many thanks for your participation by December 2022.

For any further information you can contact lukas.egli@ufz.de, j.fischer@ufz.de, christin.henzen@tu-dresden.de