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

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:

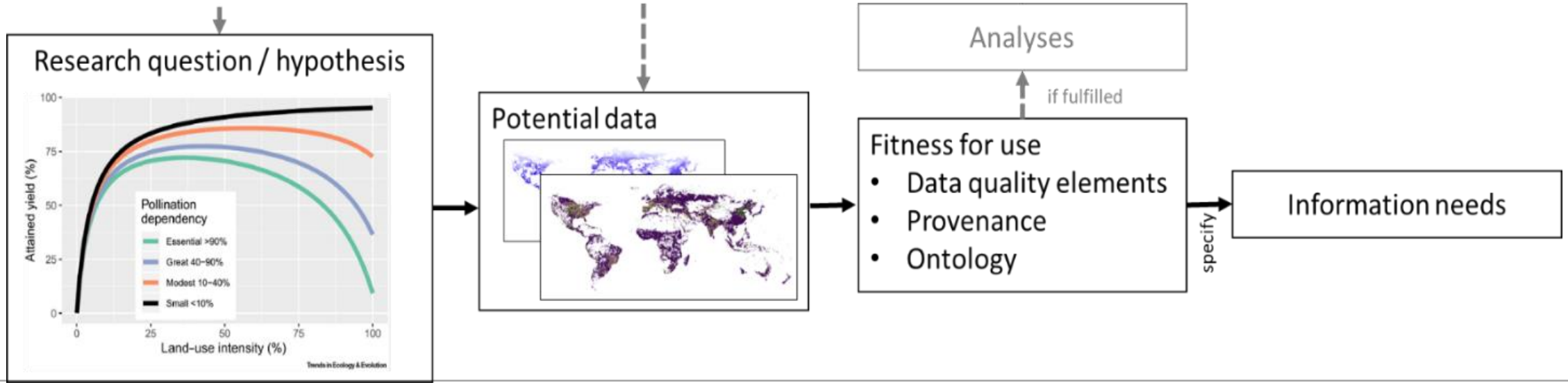
- Too many special solutions
- Lack of management plans, software and guidance for research data

Project GeoKur

Data consumer's perspective

2 Use cases on

Human-nature interactions

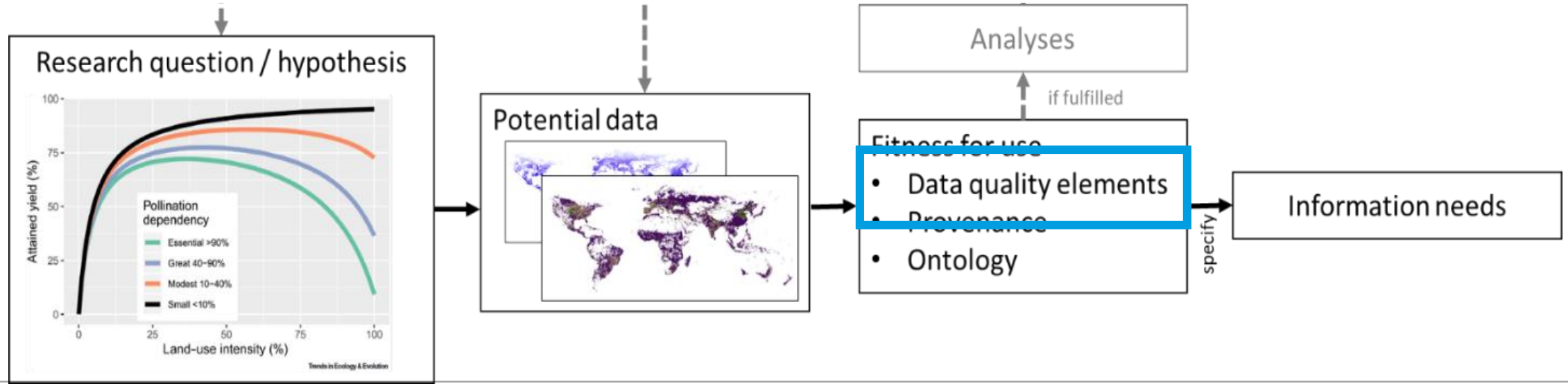


Use Cases from Different Perspectives

Data consumer's perspective

Thematic classification correctness

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)?

	Positional accuracy	Temporal quality	Thematic accuracy	Completeness	Logical consistency	Usability element
Very often	○	○	○	○	○	○
Often	○	○	○	○	○	○
Sometimes	○	○	○	○	○	○
Never	○	○	○	○	○	○
Don't know	○	○	○	○	○	○

Coming soon

Are the provided information structured using a standard?

Choose one of the following answers

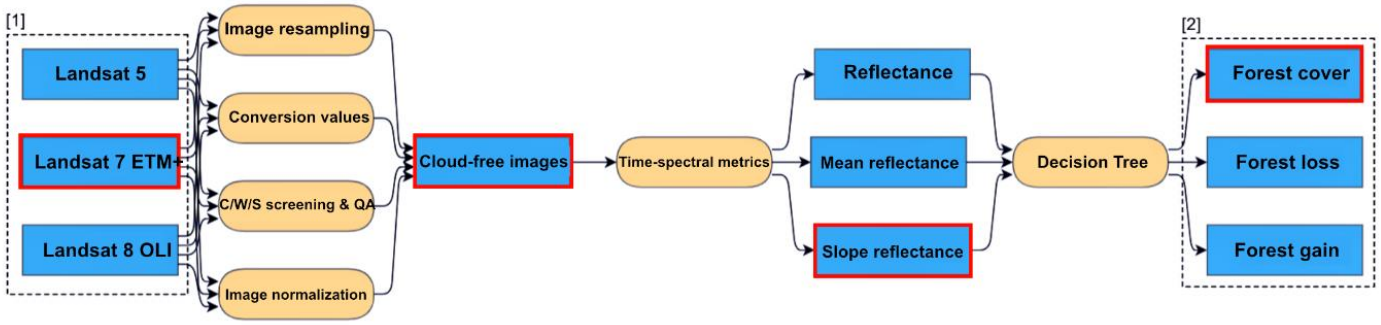
Yes. Please specify:

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

	Positional accuracy	Temporal quality	Thematic accuracy	Completeness	Logical consistency	Usability element
Metadata	○	○	○	○	○	○
Associated publication	○	○	○	○	○	○
Associated report	○	○	○	○	○	○
Directly from/within the data (e.g. grid based)	○	○	○	○	○	○

Global Forest Change 2000-2019 v1.7 (Hansen)

Provenance Graph



Quality Measures

0.6 — 1.0

	Landsat 7 ETM+	Cloud-free images	Slope reflectance	Forest cover
Representativity	---	---	---	---
Omission	0.06	0.11 (0.03)	0.03	---
Comission	0.72	0.92	0.86 - 0.89	0.98 (0.01)
Quantitative attribute accuracy	3.25 - 4.16	2.4	---	0.47
Thematic classific. correctness	0.9	0.72	0.87	0.997
Absolute external positional accuracy	193 m	120 m	98 m	67 m
Temporal consistency	---	False	True	True

Matrix view | Simple graph | Cumulative graph

Metadata

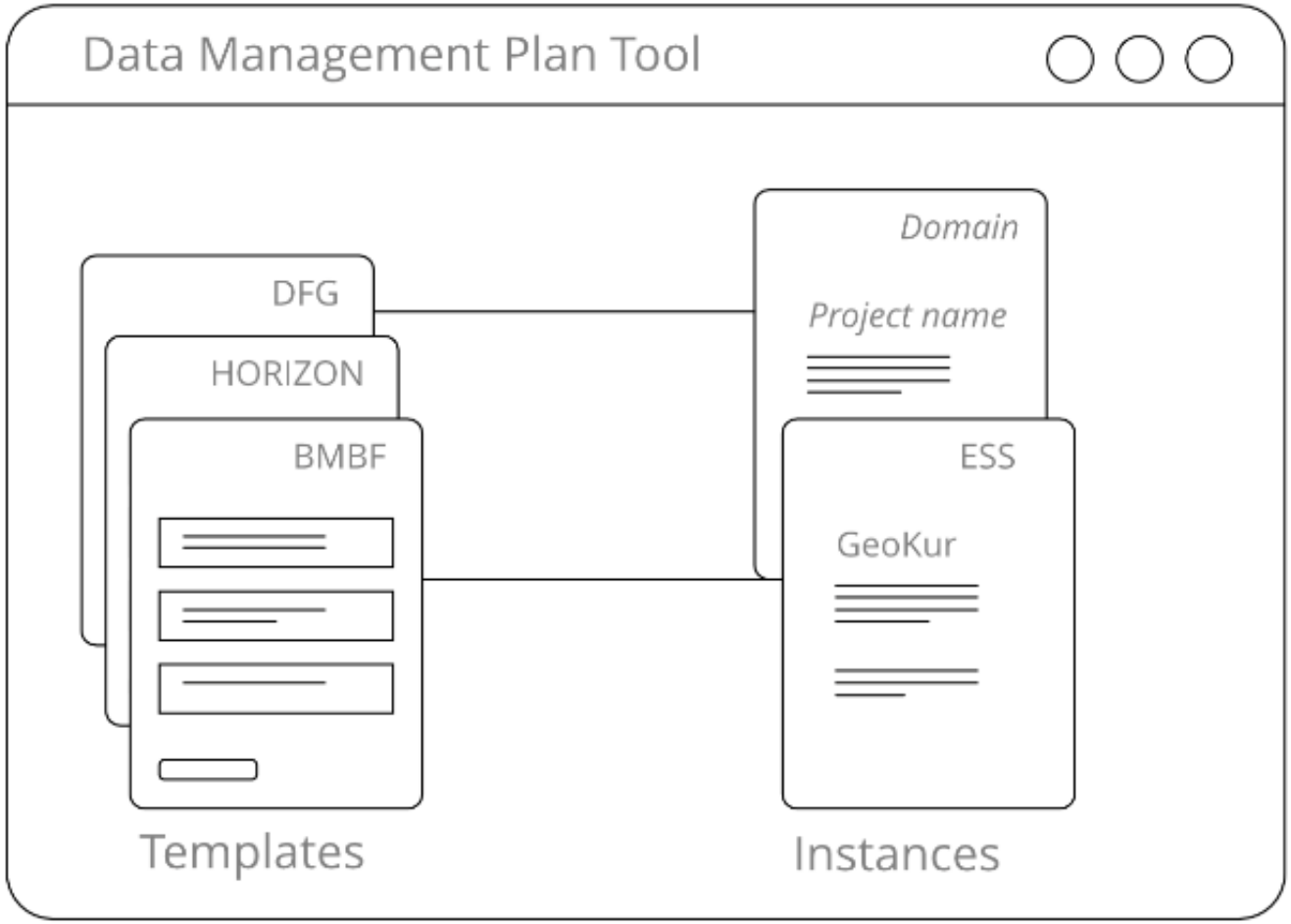
Name	Global Forest Change 2000–2019 v1.7
Documentation	High-Resolution Global Maps of 21st-Century Forest Cover Change
Contact Point	M. C. Hansen
Dataset DOI	DOI: 10.1126/science.1244693
Website	http://earthenginepartners.appspot.com
Theme	INSPIRE Landcover INSPIRE Landuse
CRS	EPSG:4326
Spatial Resolution	1 arcsec
Temporal Extent	2000-2019
Parent Dataset	Global Forest Change 2000–2018 v1.6
Derived from	Landsat 5 Landsat 7 ETM+ Landsat 8 OLI
Temporal Resolution	P1Y
Created	May 26, 2021, 11:46 AM (UTC+01:00)

[Geo-Dashboard Concept](https://doi.org/10.5194/agile-giss-2-25-2021)

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

Data Management Plans

Data provider's perspective

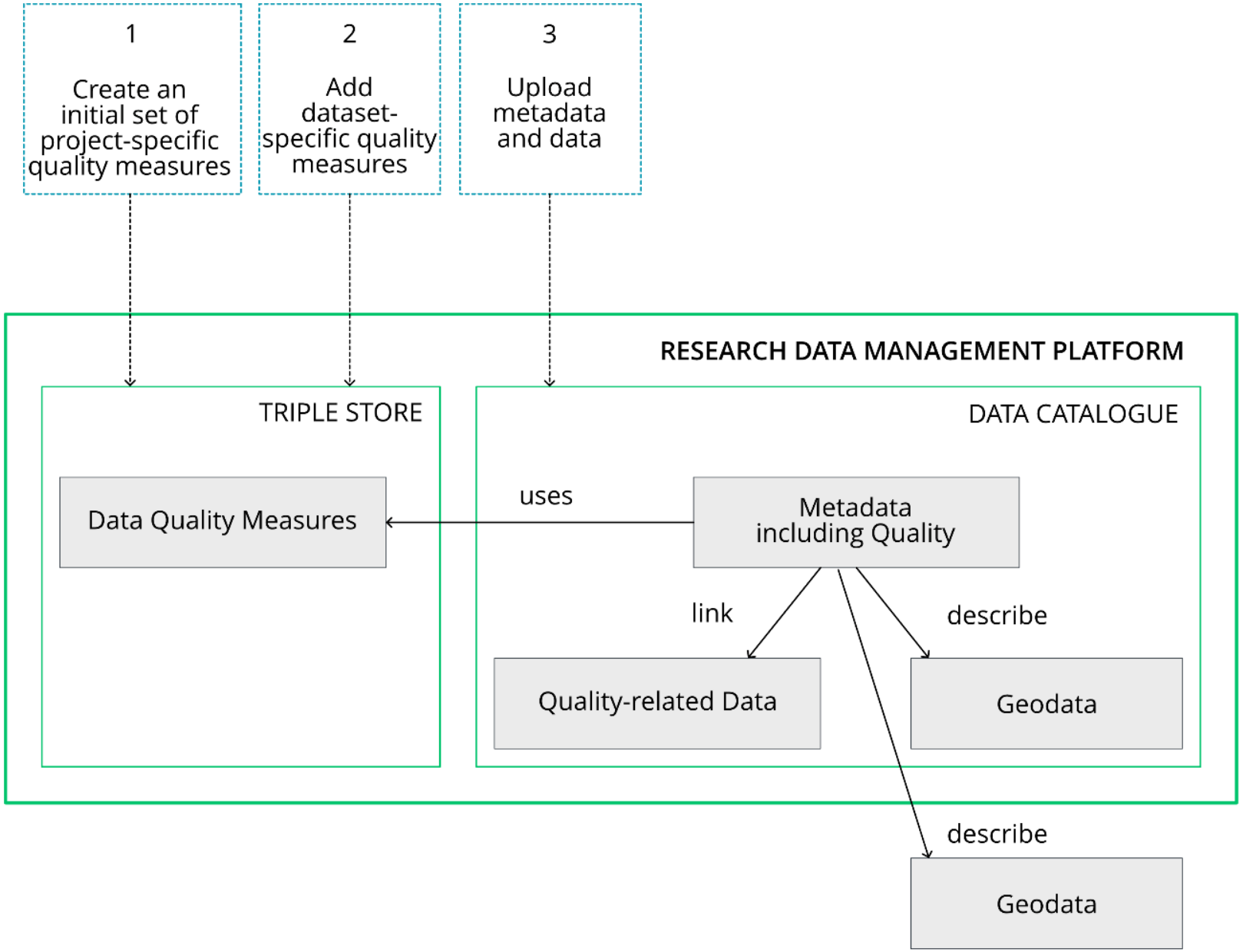


Recommendations for Future Data Management Plans in Earth System Sciences

https://doi.org/10.5194/agile-giss-2-31-2021

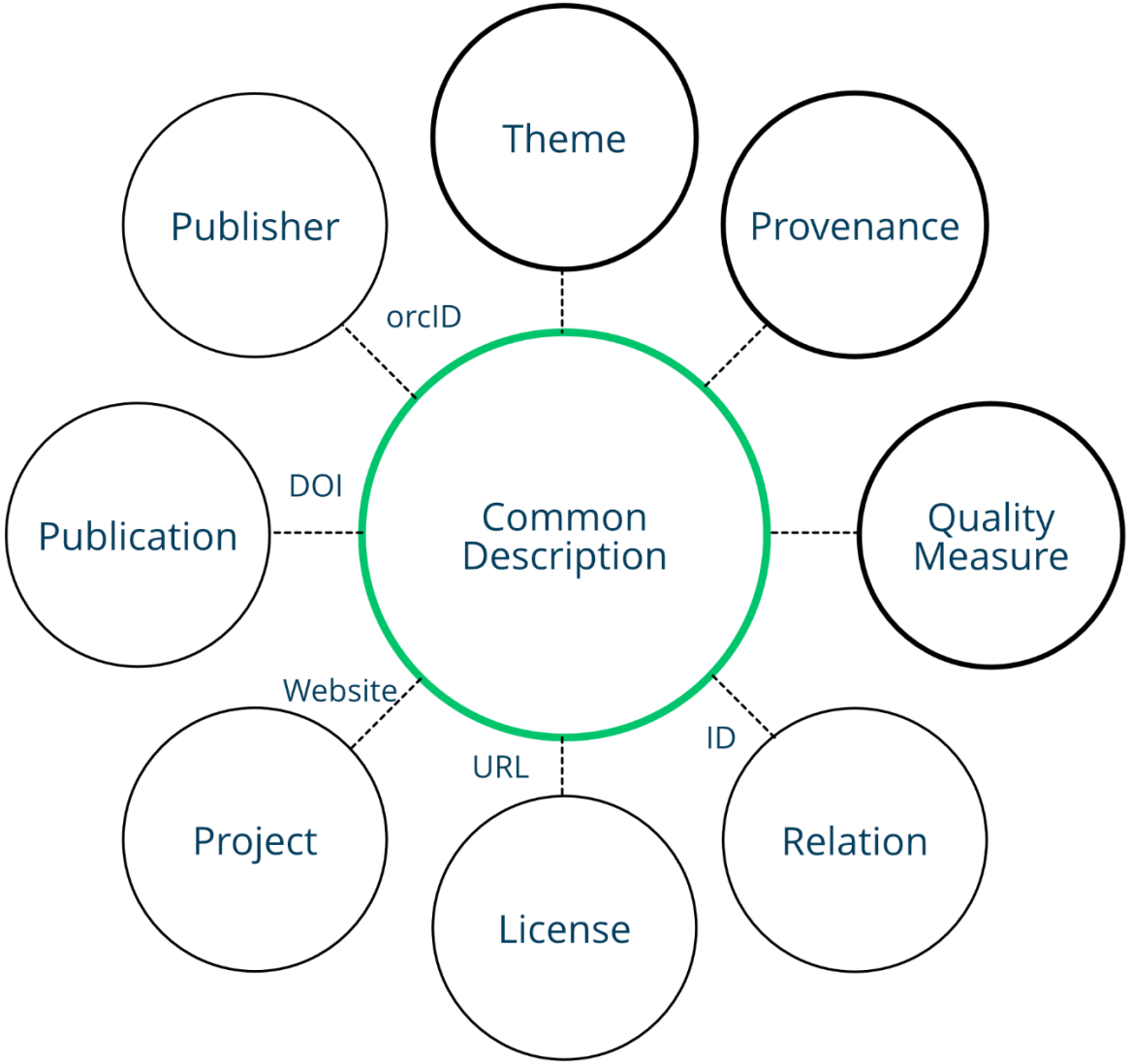
http://doi.org/10.5281/zenodo.4916856

Generic Linked Data Quality Concept



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 Ommission as Number of Missing Items

https://geokur-dmp.geo.tu-dresden.de/pages/quality-elements#completenessOmmissionAsNumberOfMissingItems

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

GEOKUR

Completeness Ommission as Number of Missing Items

https://geokur-dmp.geo.tu-dresden.de/pages/quality-elements#com

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/

Quantitive Attribute Accuracy as Attribute Value Uncertainity at 99% Significanc

https://geokur-dmp.geo.tu-dresden.de/pages/quality-elements#quantitiveAttributeAccuracyAsAttributeValueUncertainityAt99SignificanceLevel

value of quality metric:	R^2 = 0.71-0.91; 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 production maps
type of quality source:	Publication
link to quality source:	https://doi.org/10.5194/essd-12-3545-2020

Generic Linked Data Quality Concept

GEOKUR

Completeness Ommission as Number of Missing Items

https://geokur-dmp.geo.tu-dresden.de/pages/quality-elements#com

value of quality metric:	14
ground truth dataset:	
confidence term:	
confidence value:	
thematic representativity:	supp_info
spatial representativity	
temporal representativity	
name of quality source	
type of quality source	
link to quality source	

Quantitive Attribute Accuracy as Attribute Value Uncertainity at 99% Significanc

https://geokur-dmp.geo.tu-dresden.de/pages/quality-elements#quantitiveAttributeAccuracyAsAttributeValueUncertainityAt99SignificanceLevel

value of quality metric:	R^2 = 0.71-0.91; RMSE = 231-307
ground truth dataset:	CDL2010 dataset
confidence term:	
confidence value:	

Absolute External Positional Accuracy as Bias

https://geokur-dmp.geo.tu-dresden.de/pages/quality-elements#absoluteExternalPositionalAccuracyAsBias

value of quality metric:	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.
ground truth dataset:	
confidence term:	
confidence value:	
thematic representativity:	

bal gridded ag

Management and Extraction Tools

IGEOKUR

IGEOKUR Datasets Organizations Process Register Quality Register

Organizations / TUD / World Database on...

World Database on Protected Areas (WDPA v1.6)

Followers
0

Follow

Organization

TECHNISCHE UNIVERSITÄT DRESDEN

TUD

There is no description for this organization

Social

Twitter

Facebook

License

https://www.protectedplanet.net/en/legal

Dataset Groups Activity Stream Manage

World Database on Protected Areas (WDPA v1.6)

The World Database on Protected Areas (WDPA) is the most comprehensive global database of marine and terrestrial protected areas. It is a joint project between UN Environment Programme and the International Union for Conservation of Nature (IUCN), and is managed by UN Environment Programme World Conservation Monitoring Centre (UNEP-WCMC), in collaboration with governments, non-governmental organisations, academia and industry.The WDPA is updated on a monthly basis.

Dataset Provenance | Dataset as JSON

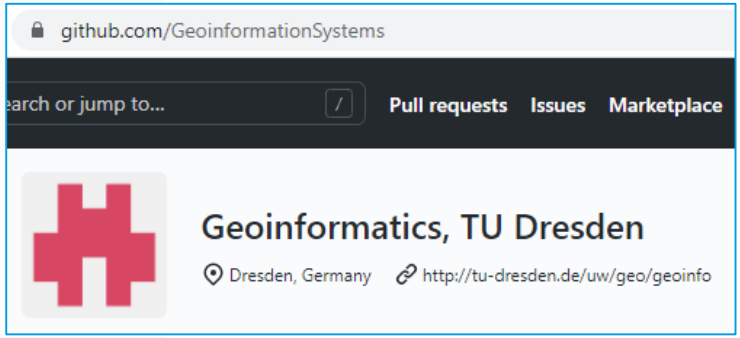
Data and Resources

This dataset has no data, why not add some?

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/WDP



Java Tool MetadataFromGeodata
https://doi.org/10.5194/agile-giss-2-41-2021

Why do we need guidelines for quality in GeoKur?

Guidance	Discipline-specific DMP Template
	Metadata Profile
	Generic Linked Data Quality Concept
	Extraction Tool
	Geo-dashboard Concept