AGILE: GIScience Series, 2, 41, 2021. https://doi.org/10.5194/agile-giss-2-41-2021
Proceedings of the 24th AGILE Conference on Geographic Information Science, 2021.
Editors: Panagiotis Partsinevelos, Phaedon Kyriakidis, and Marinos Kavouras.
This contribution underwent peer review based on a full paper submission.

A Research Data Infrastructure Component for the Automated Metadata and Data Quality Extraction to Foster the Provision of FAIR Data in Earth System Sciences

Michael Wagner[a] (corresponding author), Christin Henzen[b], Ralph Müller-Pfefferkorn[a]

michael.wagner@tu-dresden.de, christin.henzen@tu-dresden.de, ralph.mueller-pfefferkorn@tu-dresden.de

[a]Centre for Information Services and High Performance Computing (ZIH), Technische Universität Dresden, Germany
[b]Chair of Geoinformatics, Technische Universität Dresden, Germany

Abstract. Metadata management is core to support discovery and reuse of data products, and to allow for reproducibility of the research data in Earth System Sciences (ESS). Thus, ensuring acquisition and provision of meaningful and quality assured metadata should become an integral part of data-driven ESS projects.

We propose an open-source tool for the automated metadata and data quality extraction to foster the provision of FAIR data (Findable, Accessible, Interoperable Reusable). By enabling researchers to automatically extract and reuse structured and standardized ESS-specific metadata, in particular quality information, in several components of a research data infrastructure, we support researchers along the research data life cycle.

Keywords: Metadata management, automated metadata extraction, data quality, ISO 19115

1 Introduction

Most Earth System Sciences (ESS) research projects are data-driven. Metadata provide descriptions for these data and are core to support discovery, evaluation and reuse of the created data products. Gathering detailed metadata, as addressed by open and reproducible science communities, can be time-consuming (cp. Devillers, 2010). Thus, there is a growing need for efficient tool-supported metadata and data management along the data life cycle. The automated acquisition of meaningful and quality assured metadata should become an essential part of research data management (RDM).

The FAIR principles provide guidelines to improve the findability, accessibility, interoperability, and reuse of digital objects, in particular data (Wilkinson et al., 2016). Thus, these principles strengthen the importance of the availability of meaningful and quality assured metadata for the research data. While the FAIR principles include domain-independent guidelines, several communities strongly encourage connecting the FAIR principles to domain-specific standards for data quality (cp. RfII, 2020). In ESS, quality information plays a major role to evaluate research data. However, in most cases, there is still a lack in providing quality information – (1) needed quality information is not available, (2) provided information lacks the required level of detail, e.g. only summarized qualified information is published, and/or (3) quality information is presented as long-texts, being not directly usable in analysis workflows.

We therefore provide an open-source RDM component for the automated extraction of ESS-specific metadata, in particular quality metadata, to foster the provision of FAIR data in ESS. Thus, we support data producers to generate and update selected metadata automatically, e.g. to improve existing metadata, building on well-known standardized formats and schemas.

2 Related Work

In ESS, data are often created and used in heterogeneous and interdisciplinary research projects. Therefore, the requirements for a suitable metadata schema differ, e.g. focussing a certain aspect like quality information or covering general aspects such as interoperability across domains. Dublin Core, for instance, provides a vocabulary to describe cross-domain resources (DCMI Usage Board, 2020). To cover ESS-specific meta information, like linked spatial, temporal, and thematic information, ISO 19115-1:2014 provides a schema for the description of geographic information, which can be implemented as XML (International Organization for Standardization, 2014 and 2016). The ESS-specific application profile (AP) GeoDCAT uses linked data concepts and extends the Data Catalog Vocabulary (DCAT) AP with ISO 19115:2003 elements (W3C, 2020; European Commission, 2016).

2.1 Data Quality Assurance and Modelling

Researchers need standardized data quality measures to foster the evaluation of data quality and effective data management (Yang et al. 2013). Thus, data quality and related quality assurance should become essential parts in research data management and should be covered during all phases of the data life cycle.

However, generating, providing and assuring structured and standardized, machine-readable quality information, is still a pressing challenge in data-driven research, addressing several aspects like big and interdisciplinary data, circular reasoning, verification and reproducibility (cp. RfII, 2019; Albertoni and Isaac, 2021, Yang et al. 2013).

Quality assurance (QA) as part of the quality management focuses on the fulfilling of quality requirements (ISO 9000:2015). Hence, QA comprises a set of activities, roles, models and measures to ensure the quality (and fitness-for-use) of data. Here, we focus on models to structure domain-specific descriptions of quality information. The data quality vocabulary (DQV, Albertoni and Isaac, 2021), for instance, covers interdisciplinary quality aspects implemented as linked data. ISO 19157:2013 (International Organization for Standardization, 2013) describes seven geo-domain-specific categories of data quality including spatial, temporal and thematic aspects (Tab. 1).

Table 1: Data quality categories in ISO 19157:2013

Completeness	Missing and excess data
Logical consistency	Adherence to logical rules of data structure, attribution and relationships
Positional accuracy	Accuracy of the position; needs ground truth data
Temporal quality	Quality of temporal attributes and relationships of features
Thematic accuracy	Accuracy of quantitative and non-quantitative attributes and classifications of features; benefits from ground truth data
Usability	Based on user requirements
Quality of Metadata	Confidence in, representativity of and homogeneity of data quality evaluations

Several tools for extracting metadata from different file types do exist. Apache Tika[1], for instance, provides several ESS-specific parsers, such as GDALParser (using GDAL library[2]), and GeoParser[3], which enable the automated generation of spatial metadata, e.g. reference system, raster's origin, and cell size. The File Information Toolset (FITS[4]), acts as a wrapper for several open-source toolsets, like Tika, supporting the metadata extraction from several file types. Some tools enable metadata analysis from data published in a certain repository, like pangaeapy[5]. Other existing tools only extract certain ESS-specific metadata elements, e.g. do not cover temporal and thematic extent, and, in particular, they do lack in extracting and providing structured data quality elements.

3 A Metadata and Data Quality Extraction Tool for Geospatial Data

We propose an open-source Java tool, called metadataFromGeodata[6], for the automated metadata extraction and quality assurance to foster the provision of FAIR geospatial data. Our tool uses geospatial data, provided in well-known formats, as input to automatically extract meta information and to generate machine-readable ESS-specific general and quality metadata.

3.1 Using our tool in the data life cycle

The research data life cycle, and existing community versions of the cycle, describe the phases of data management (de la Hidalga, 2020; GFBio, 2021; RfII, 2020). We propose to use our tool during several phases of the life cycle (Fig. 1). During the *collection*

[1] http://tika.apache.org
[2] https://gdal.org/programs/gdalinfo.html
[3] https://tika.apache.org/1.26/api/org/apache/tika/parser/geo/topic/GeoParser.html
[4] https://projects.iq.harvard.edu/fits
[5] https://github.com/pangaea-data-publisher/pangaeapy
[6] https://github.com/GeoinformationSystems/MetadataFromGeodata

phase, our tool enables researchers (1) to facilitate the evaluation of the fitness-for-use by generating (complementing) metadata for collected data, and (2) to foster quality assurance, in particular quality analysis, by providing structured machine-readable quality information to be used during other phases, e.g. as analysis input. In the *analyze phase*, in particular during the iterative development of an analysis workflow, researchers can use our tool to evaluate the analysis output. The extracted quality information indicates if further improvements of the workflow are needed. During the *publication and archiving phases*, our tool supports generating and updating structured metadata for research data products, enabling researchers to use the generated metadata in other research data infrastructure components, e.g. data management systems.

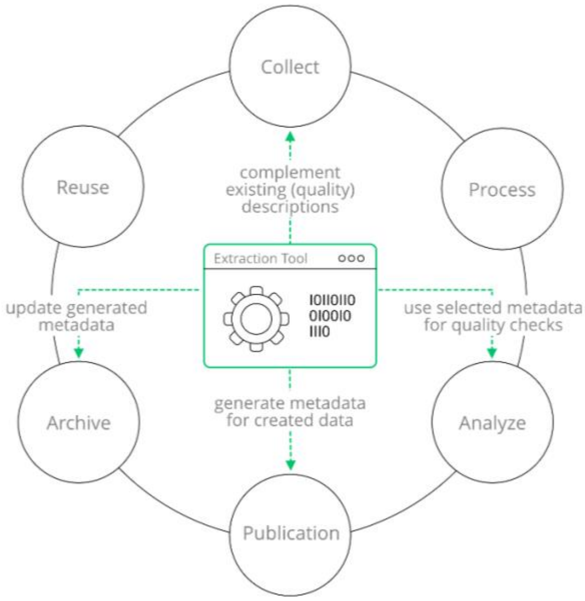
Collect
Reuse
Process
complement existing (quality) descriptions
Extraction Tool
IOIIOIIO
OIOOIO
IIIO
update generated metadata
use selected metadata for quality checks
Archive
Analyze
generate metadata for created data
Publication

Figure 1: Using the metadata extraction tool to reduce efforts in the phases of the research data life cycle

3.2 Extracting General Metadata for ESS Metadata Profiles

Geoinformation are stored in a variety of file formats, e.g. following the open data and reproducibility movements by using open formats, like the CSV format, or community-driven open formats, like the Open Geospatial Consortium Geopackage[7]. Our metadataFromGeodata tool supports widely used open-source or community-specific file formats enabling the information extraction for vector and raster data from: (1) GeoPackages with multiple vector layers, (2) ESRI Shapefiles, (3) GeoTIFFs with several bands, and (4) CSV files including several related attributes, e.g. temporal information and/or commodities. When analyzing a CSV file, we distinguish content columns used for extracting quality information, spatial columns to map the geometry and to link to GeoPackages or Shapefiles, and ignored columns.

The extracted meta information varies and is strongly related to the input file format, e.g. a raster file typically includes one temporal/thematic attribute, a CSV file can include several temporal and thematic information.

We use two datasets as example: (1) *Crop production in EU standard humidity* provided by Eurostat[8] as CSV file and (2) *Database of Global Administrative Areas* (GADM[9]) providing country administrative areas as GeoPackage file, and made minor changes in both datasets, e.g. correcting country names.

Tab. 2 summarizes the extracted meta information for the currently supported file formats, providing examples based on the Eurostat and GADM datasets, following the ISO 19115:2014 structure, which we use as one output format.

3.3 Extracting Data Quality Information for ESS-specific Metadata Profiles

Following our approach to automatically extract meta information from different file formats, we can obtain subcategories of the following quality information: completeness, logical consistency, temporal quality, thematic accuracy and quality of metadata (Tab. 3).

We obtain completeness information as absolute or relative values by counting missing and excess items. For format consistency, we check, if the file format is included in a user-defined list (then true). We derive the temporal consistency by evaluating the duration of given time steps – being true, if the duration is constant, e.g. one year. Further, we obtain the non-quantitative attribute correctness by checking, if attributes are included in a thematic classification / ontology given by the user.

[7] http://www.geopackage.org

[8] https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=apro_cpsh1&lang=en

[9] https://gadm.org/data.html

Table 2: Extracted general ESS metadata for certain geospatial file formats

Metadata collection	**Metadata element**	**Extracted values from examples** ***Crop production in EU standard humidity*** **and** ***GADM*** **(marked with *)**	**GeoPackage**	**Shapefile**	**GeoTIFF**	**CSV**
General information	User name	mwagner (using system login)	x	x	x	x
	Role	resourceProvider (defined by code list in ISO 19115)	x	x	x	x
	UUID	d5939c5a-99c0-4339-a694-ef72a73d3314	x	x	x	x
	File creation date	2021-04-19T05:19:44.237Z	x	x	x	x
	Last file update	2021-04-16T15:47:02.961Z	x	x	x	x
Reference system	Spatial reference system (SRS) defining organization	EPSG *	x	x	x	
	SRS identifier	EPSG:4326 *	x	x	x	
	SRS description	WGS 84 geodetic *	x	x	x	
	Metadata creation date	2021-04-19T05:19:44.237Z	x	x	x	
Structure of spatial data	Spatial representation type	vector (defined by code list in ISO 19115)	x	x	x	x
	Environmental description	filename:examples/apro_cpsh1_1_Data_woEUl27.csvfile size: 245164 B geographical file: examples/gadm36_level0_extended.gpkg layer name: level0	x	x	x	x
	Geographical extent in source SRS	-180, 180, -90, 83.658 *	x	x	x	
	Geographical extent in standard SRS (WGS84, EPSG:4326)	-180, 180, -90, 83.658 *	x	x		
	Spatial resolution	0.048	x	x	x	
	Temporal resolution	1 (year)				x
	Thematic keywords	Cereals for the production of grain (including seed), Dry pulses and protein crops for the production of grain (including seed and mixtures of cereals and pulses), Fresh vegetables (including melons), Permanent crops for human consumption				x
Metadata contact	Link to ISO standard	ISO 19115-1, first edition	x	x	x	x

Table 3: Extracted quality information for ESS metadata for certain geospatial file formats

Quality category	Metadata element	Extracted values from examples *Crop production in EU standard humidity*	GeoPackage	Shapefile	GeoTIFF	CSV
Completeness	Counts and rates of missing items per attribute/band	651, 32.55 %			x	x
	Counts and rates of excess items	0, 0 %				x
Logical consistency	Format consistency	True	x	x	x	x
Temporal quality	Temporal consistency	True				x
Thematic accuracy	Non-quantitative attribute correctness	0 (number of incorrect attributes)				x
Quality of metadata	Polygons per area	0.003417 per 1000 km^2	x			x
	Count of temporal units	10				x
	Count of thematic units	4				x
	Empirical distribution parameters of various combinations of spatial-temporal-thematic units	mean=8.9, min=7.0, max=10.0, div. quantiles				x

For the metadata quality, we provide various information, e.g. count of polygons per area, number of geometries, temporal units, or thematic values, which indicate the representativity and homogeneity of the metadata. We calculate complex quality parameters by analyzing the distribution of temporal and thematic elements per geographical unit, and thematic elements per temporal unit.

Due to the ISO structure, we cannot automatically obtain quality information for the ISO categories positional accuracy and usability elements, as this would require an analysis with a ground truth dataset resp. individual user input.

3.4 Integrating the Extraction Tool into Research Data Infrastructures

Research data infrastructures aim to manage research data and metadata systematically, supporting researchers during all phases of the data life cycle. In a research data infrastructure, our proposed metadataFromGeodata tool can be used as a standalone component or included resp. linked to existing components, such as data management systems (DMS).

To foster the integration into other components and the reuse of the extracted information, we currently offer two options for information storage: metadata is stored as (1) structured ISO19115 and ISO19157-compliant XML file or as (2) an SQLite database. Future work is to provide DQV metadata, which can be mapped to ISO metadata.

The generated XML file can be used to link our tool to DMS, like CKAN[10], Dataverse[11] or DSpace[12], and to data archives (Fig. 2, right). Researchers can publish or update the XML metadata via a DMS/archive API or the user interface for uploads. Furthermore, we provide extension points to support implementing export modules for DMS specific file formats. With the new quality information available in the DMS or archive,

[10] https://ckan.org
[11] https://dataverse.org
[12] https://duraspace.org/dspace/

the selection of data for re-use can then also be based on the data quality criteria.

Linking our tool to analysis tools fosters the integration of quality information in the analysis workflows from the beginning, e.g., for data input or interim results, and the quality assurance of the created research data (Fig. 2, lower left).

Data management plan tools manage formal and structured documents that outline data handling, related roles, specifications and workflows in data management plans (DMP). Some DMP tools do already link to DMS, enabling the researchers to capture data descriptions via the DMP tool and automatically publish these descriptions as structured and standardized metadata in a DMS. By adding the option to update metadata generated by our extraction tool in the DMP tool, the metadata management would become more efficient.

4 Vision and Outlook

Metadata management is core to support discovery and reuse of data products. However, the automatic extraction of quality parameters for geospatial data is still a pressing challenge. In ESS, several file formats do exist for raster and vector data, enabling the extraction of quality parameters. To enable complementing metadata, we plan to implement the extraction by analyzing coupled files and generate a common metadata set.

Quality measures and metrics can become complex. Providing descriptions for the measures and metrics as linked data, similar to thematic ontologies published in a knowledge hub (Fig. 2), fosters semantic querying during analysis and discovery, when the generated metadata with the links are published in DMS. Therefore, we plan to enable researchers to register ontologies as input, which will then be included in the generated metadata files. We are currently implementing modules for GeoDCAT metadata stored in RDF files to facilitate the inclusion of metadata in other linked data components.

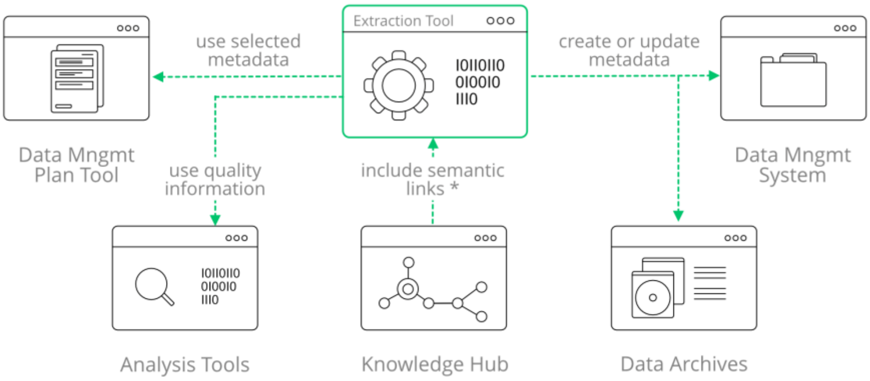

Figure 2: Linking the metadata extraction tools to research data infrastructure components

Currently, the quality category "usability" is implemented as a generic set of measures. In the future, we plan to enable researchers to define their own usability measures, e.g. based on formal descriptions like applied in model-driven development. ESS communities are discussing usage metrics (Lowenberg et al., 2021), which could be included in the usability category as well.

By enabling ESS researchers to automatically generate quality metadata, we hope to foster the provision of FAIR data and to support an efficient metadata management as well as quality assurance in data-driven research projects.

Software and Data Availability

The tool metadataFromGeodata is published as an open-source project on GitHub: https://github.com/GeoinformationSystems/MetadataFromGeodata under the GNU LGPL license. The Eurostat dataset used as an example for tabular data is available on https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=apro_cpsh1&lang=en. The GADM dataset, serving as example for geometries stored in Geopackage files can be obtained on https://gadm.org/data.html. In a pre-processing step, we made minor adaptions on country names to facilitate the linking of geometry and thematic files. Both modified datasets are published on

GitHub in the subfolder examplesAGILE2021. Further files in that folder comprise the properties file to run metadataFromGeodata on Eurostat and GADM examples, and the generated metadata files.

Acknowledgement

The presented metadataFromGeodata tool was implemented in the project GeoKur. GeoKur is granted by the BMBF (Federal Ministry of Education and Research) under number 16QK04A.

References

Albertoni, R. and Isaac, A.: Introducing the Data Quality Vocabulary. Semantic Web, Vol. 12, No. 1, 81-97, https://doi.org/10.3233/SW-200382, 2021.

DCMI Usage Board: DCMI Metadata Terms. http://dublincore.org/specifications/dublin-core/dcmi-terms/2020-01-20/, last access: 7 April 2021, 2020.

Devillers, R., Stein, A., Bédard, Y, Chrisman, N., Fisher, P., and Shi, W.: Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. Transactions in GIS, 14 (4), 387-400, https://doi.org/10.1111/j.1467-9671.2010.01212.x, 2010.

European Commission: GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe. Version 1.0.1, 2016.

German Federation for Biological Data: GFBio Training Materials: Data Life Cycle Fact-Sheet: Data Life Cycle: Plan. https://www.gfbio.org/training/materials/data-lifecycle/plan, last access: 7 April 2021.

de la Hidalga, A.N., Hardisty, A., Martin, P., Magagna, B., and Zhao, Z.: The ENVRI Reference Model. In: Zhao, Z. and Hellström, M. (eds) Towards Interoperable Research Infrastructures for Environmental and Earth Sciences. Lecture Notes in Computer Science, vol 12003. Springer, Cham. https://doi.org/10.1007/978-3-030-52829-4_4, 2020.

International Organization for Standardization: Geographic Information – Data quality (ISO 19157:2013). First edition, Geneva, 2013.

International Organization for Standardization: Geographic Information – Metadata – Fundamentals (ISO 19115-1:2014). First edition, Geneva, 2014.

International Organization for Standardization: Quality management systems – Fundamentals and vocabulary (ISO 9000:2015). Fourth edition, Geneva, 2015.

International Organization for Standardization: Geographic Information – Metadata – XML schema implementation for fundamental concepts (ISO 19115-3:2016). First edition, Geneva, 2016.

Lowenberg, D., Jouneau, T., and Bruno, I.: RDA Data Usage Metrics WG Recommendations. Research Data Alliance. https://doi.org/10.15497/RDA00062, 2021.

Nightingale, J., Boersma, K.F., Mulller, J.-P., Compernolle, S., Lambert, J.-C., Blessing, S., Giering, R., Gobron, N., De Smedt, I., Coheur, P., George, M., Schulz, J., and Wood, A.: Quality Assurance Framework Development Based on Six New ECV Data Products to Enhance User Confidence for Climate Applications. Remote Sens. 2018, 10(8), 1254; https://doi.org/10.3390/rs10081254, 2018.

RfII – Rat für Informationsinfrastrukturen: The Data Quality Challenge - Recommendations for Sustainable Research in the Digital Turn. Göttingen, 2020.

Wagner M.: metadataFromGeodata [code]. Retrievable from https://github.com/GeoinformationSystems/MetadataFromGeodata, 2021.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018, https://doi.org/10.1038/sdata.2016.18, 2016.

W3C: Data Catalogue Vocabulary (DCAT) – Version 2. Eds: Albertoni, R., Browning, D., Cox, S., Beltran, A.G., Perego, A., and Winstanley, P., https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/, 4 February 2020.

Yang, X., Blower, J. D., Bastin, L., Lush, V., Zabala, A., Masó, J., Cornford, D., Díaz, P., and Lumsden, J.: An integrated view of data quality in Earth observation. Phil Trans R Soc A 371: 20120072, https://doi.org/10.1098/rsta.2012.0072,2013.