AGILE: GIScience Series, 2, 25, 2021. https://doi.org/10.5194/agile-giss-2-25-2021
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 Geo-Dashboard Concept for the Interactively Linked Visualization of Provenance and Data Quality for Geospatial Datasets

Heiko Figgemeier [a] (corresponding author), Christin Henzen[a] , Arne Rümmler[a]

heiko.figgemeier@tu-dresden.de, christin.henzen@tu-dresden.de, arne.ruemmler@tu-dresden.de

[a]Chair of Geoinformatics, Technische Universität Dresden, Germany

Abstract. In Earth System Sciences, a data-driven research domain, several communities discuss the importance, guidance and implementation of making research data findable, accessible, interoperable, and reusable. To foster these principles, in particular to support reusability, users need easy-to-use user interfaces with meaningful visualizations for detailed metainformation, e.g. on dataset's origin and quality. However, visualization tools to facilitate the evaluation of fitness for use of ESS research data on domain-specific metainformation, do hardly exist.

We provide a Geo-dashboard concept for user-friendly interactive and linked visualizations of provenance and quality information using standardized geospatial metadata. A provenance graph visualization serves as overview and entry point for further evaluations. Quality information is essential to evaluate the fitness for use of data. Therefore, we developed quality visualizations on several levels of detail to foster evaluation, e.g. by enabling users to choose and classify quality parameters based on their use-case-specific needs.

Keywords: provenance, data quality, metadata, geo-dashboard

1 Introduction

In recent years, the publication of research data as a major output of research projects in Earth System Sciences (ESS) has become more established. Several communities [1] are currently discussing aspects and methods of research data management, like openness, reuse and reproducibility, making data findable, accessible, interoperable, and reusable (FAIR) (Wilkinson et al., 2016, Roos et al., 2016, Magagna et al., 2020). They strengthen the relevance of detailed descriptions of the data, in particular data quality and provenance.

Provenance information describes the workflow to generate data - the origin of data (cp. Moreau, 2010; Di and Yue, 2011; Simmhan et al. 2005). Quality information is strongly linked to provenance information, describing characteristics of the data (in the workflow) and is essential to evaluate the fitness for use of data. However, providing user-friendly interactive visualizations for provenance and quality information that facilitate the discovery, interpretation and evaluation of the data is still a pressing challenge.

Within our research project GeoKur[2], we identify (1) information needs of data consumers in terms of data quality and provenance to determine the fitness for use of global land use data for downstream analyses, and (2) develop user interfaces to provide proper visualizations for this complex information. We use two use cases to build our concepts on:

1. We focus on how land degradation caused by agricultural intensification threatens biodiversity and ecosystem services. For example, while agricultural intensification typically increases yields, this relationship may turn negative due to soil degradation or the loss of biodiversity that is beneficial to crop production.

[1] E.g. The Open Knowledge Foundation: https://okfn.org/opendata/how-to-open-data/, The Open Data Handbook http://opendatahandbook.org/guide/en/, Open Definition: https://opendefinition.org/, GEOLabel: https://www.geolabel.info/, Open Data Label: https://www.opendatainside.com/de/ German Reproducibility Network: https://reproducibilitynetwork.de/, NFDI4Earth Research Software and Reproducibility Interest Group: https://www.nfdi4earth.de/participate/get-involved-by-interest-groups/ig-geo-researchsoftware-reproducibility, OSGEO and Open Geoscience: https://www.osgeo.org/initiatives/open-geoscience/

[2] https://geokur.geo.tu-dresden.de/

2. We analyse spatio-temporal linkages between land degradation and human migration. Land degradation poses a risk to people dependent on natural resources and thus, is a potential cause for migration. Conversely, migration can also increase land degradation in destination areas, e.g. by contributing to land use changes.

In this paper, we provide a user-friendly visualization concept for linked provenance and quality information to support the evaluation of ESS research data. In the proposed Geo-dashboard concept, we include established visualization concepts, like a graph-based visualization for provenance, tabular/chart views for data quality and a map, built on available, well-known and well-used, metadata schemas for ESS data.

2 Methodology

Providing FAIR data is a core principle of research data management across domains. Thus, ensuring the provision of proper visualizations for meaningful metadata is core to foster the evaluation, understandability, and reproducibility of the research data. In our approach, we focus on linked visualizations for provenance and quality information. Therefore, we first address provenance modelling and graph visualizations with a given example and then provide an overview of data quality modelling and metrics.

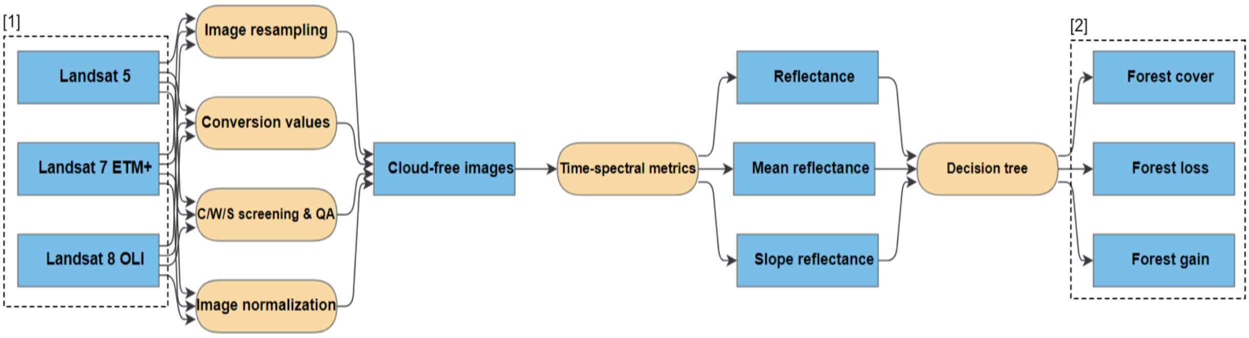


Figure 1: Provenance graph for "Global Forest Change 2000-2019" with entities (blue) and activities (yellow)

2.1 Provenance

Provenance information describes the origin or history of data and is essential to evaluate the fitness for use. In the case of ESS research data, provenance information can become complex, including scientific publications, a detailed description of all processing steps to generate the data, used inputs, intermediate results, and used tools (Magagna et al., 2020, Closa et al., 2017). Several concepts to model and formalize provenance information do already exist, e.g. provenance information for features or digital objects, described as either data-oriented or process-oriented (Henzen et al., 2013). Here, we focus on provenance on dataset level.

Existing schemas enable the description of provenance (1) as unstructured free text or as structured elements for datasets and processing steps, e.g. ISO 19115-1:2014. We can further distinguish linked data schemas and other schemas to model provenance information. The ontology PROV-O uses the linked data concept to express provenance information as (linked) entities, activities and agents as described in the provenance data model PROV-DM (Moreau and Missier, 2013).

In this paper, we use the PROV-DM, implemented as PROV-O, to describe provenance graphs and map the core concepts entity, activity, and agent as followed (Closa et al., 2017):

- An **Entity** describes a dataset used or generated during the workflow.
- An **Activity** describes a processing step within the workflow and refers to at least one dataset.
- An **Agent** implements or executes an activity.

Fig. 1 shows the provenance graph for an open accessible dataset about forest change, called "Global Forest Change 2000-2019" (see section "Software and Data Availability", Hansen et al., 2013), and being relevant for our use cases. The input datasets Landsat 5, Landsat 7 ETM+ and Landsat 8 provide satellite images taken from the related satellites (Fig. 1, [1]). In the first processing step, these datasets are used as inputs for four pre-processing scripts, e.g. resampling and normalization, resulting in a dataset with cloud-free images, which is then used to get reflectance values based on time-spectral metrics. Hansen (2013) developed a decision tree to use these datasets on reflectance to generate three outputs: percent tree cover, forest loss, and forest gain [2].

2.2 Data Quality

Data generated and provided in research projects fulfil high quality standards (RfII, 2020). Data quality is an essential criterion to evaluate a dataset's fitness for use, describing common characteristics of data from a methodological point of view (RfII, 2020, Devillers et al., 2010, Devillers et al., 2007, Ślusarski and Jurkiewicz, 2020).

ISO 19157:2013 provides a metadata schema to structure quality information for geographic data and describes seven data quality elements: positional accuracy, temporal quality, metaquality, logical consistency, completeness, thematic accuracy and usability (ISO19157:2013). Each element is structured in several sub elements, e.g. completeness includes omission and commission.

Following our use cases (Section 1), we performed semi-structured interviews with three environmental researchers to gather information needs, research data and related quality information as candidates for further analysis, and to prove our concept. Tab. 1 shows quality information for three of the proposed datasets - focusing on ISO quality elements that have been indicated as highly or medium relevant for the use cases. After evaluating provided quality information, we can summarize the following challenges:

(1) Quality of information: In some cases, quality information for a certain dataset is not available or do not provide meaningful information.

- Data producers lack guidance on how to describe quality information and do not meet data consumer's requirements. Thus, even quality elements, which can be automatically extracted, are not published, e.g. commission and omission (Tab. 1).
- Data consumers need meaningful and detailed quality information, but provided information only summarizes data quality, e.g. thematic accuracy for the Crop production (apro_cp) dataset by EUROSTAT "is assessed to be good" (Tab. 1)

(2) Information sources: Quality information is extractable from metadata, scientific publications or supplemental methodology, or can be automatically extracted from the data.

Table 1: Selected quality information for three yield and forest cover data sets – gathered from metadata or further sources (for download and description see Section "Software and Data Availability")

Quality element	Sub elements	MapSPAM	Crop production (apro_cp)	Global Forest Change 2000–2019
		IFPRI	EUROSTAT	University of Maryland
Thematic accuracy	Thematic classification correctness	Not available	"The accuracy for the final data delivery is assessed to be good."	There is no explicitly mentioned user accuracy for 2000-2019. Hansen et al. (2013) states user accuracy for 2000 – 2012 (global level): • Forest loss: 87.0 (2.8) • Non-forest loss: 99.8 (0.1)
	Quantitative attribute accuracy	Validation with Cropland Data Layer 2010 dataset in the United States (R^2 = 0.71-0.91, Root Mean Square Error=231-307)	Data quality flags (e.g. estimated, low reliability); sampling error thresholds	Not available
Completeness	Commission	Not available	Not available	Not available
	Omission	Not available	"Most of the requested data are available, but there are some missing data in the older time series."	*No missing data pixel was indicated using a data mask layer.*
Metaquality	Representativity	Not available	Not available	"All pre-processing steps were tested at national scales around the globe using a method prototyped for the Democratic Republic of Congo."

Source of quality information: □ Research paper 🟦 Supplementary Material / Methodology ▨ Metadata file 🟧 Measure not available

- Information from the different sources often complement each other and need to be combined for the evaluation.
- The levels of detail of quality information often differ - quality information is often summarized using qualitative descriptions in publications and are described in detail in supplementary material, e.g. thematic accuracy for the Global Forest Change dataset (Tab. 1).

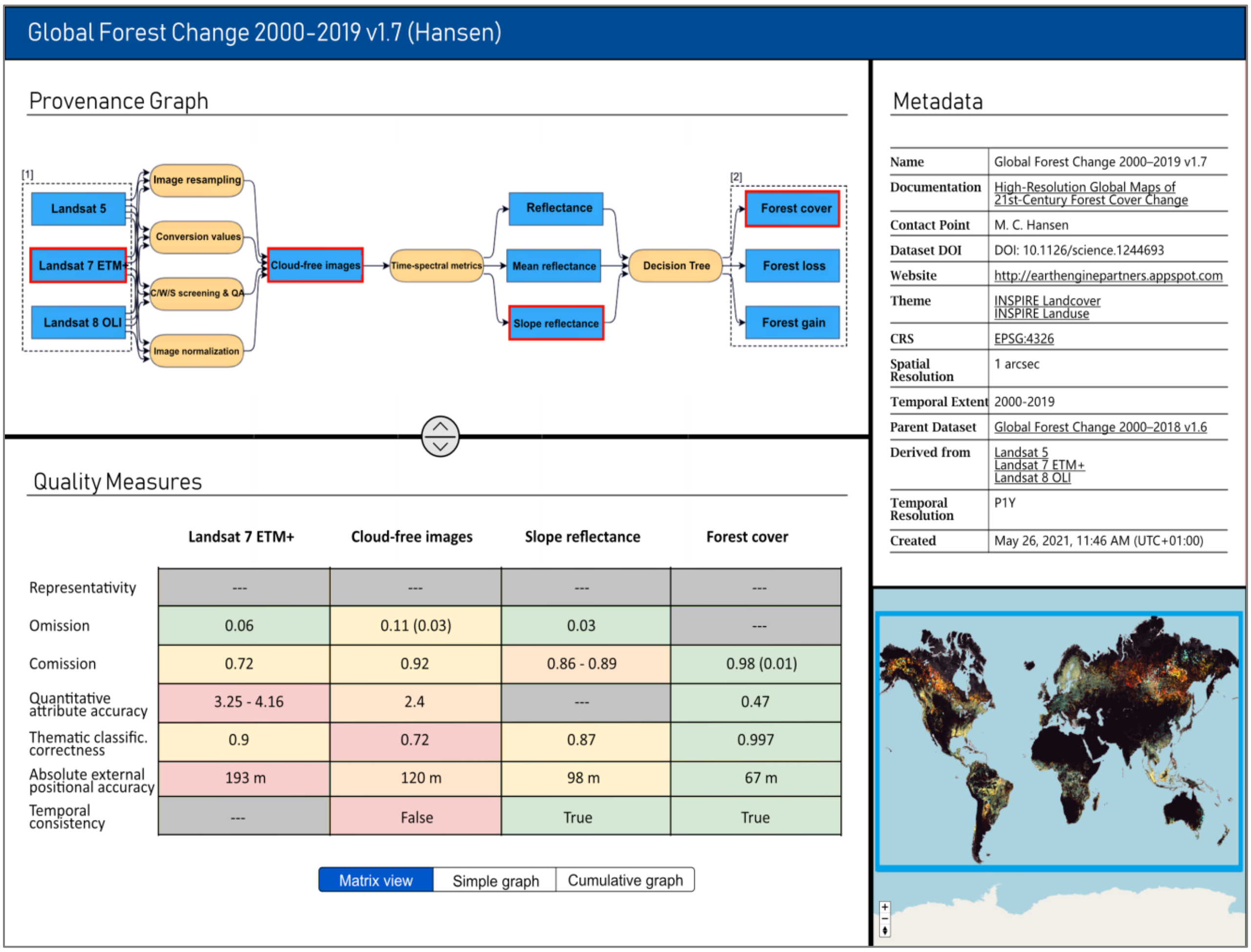
(3) Specific measures: Even within a certain domain, quality information uses different measures for the related quality elements.

- Quality information can hardly be compared, e.g. thematic classification correctness for Crop Production (apro_cp) and Global Forest Change 2000–2019 (Tab. 1).
- For automated processing, further information is needed, e.g. mapping of the measures based on ontologies or controlled vocabularies.

3 Geo-dashboard visualization concept

We provide an interactive and user-friendly Geo-dashboard concept for the linked visualization of general metadata, provenance information, quality information and geospatial data to support the efficient evaluation of fitness for use for geospatial data.

We use established software concepts to enable discovery of complex metainformation: (1) The dashboard concept facilitates the visual presentation of information in several widgets. In particular, the Geo-dashboard concept provides widgets for geospatial data (Bernasocchi et al., 2012). (2) The linked-view principle enables users to interact with multiple views (Jing et al., 2019). (3) Shneiderman’s usability mantra: Overview first, zoom and filter, then details-on-demand (Shneiderman, 1997) supports the evaluation of fitness for use on several levels of detail.



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

Provenance Graph

Quality Measures

	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

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)

Figure 2: Geo-dashboard user interface with four widgets: provenance graph (upper left), several quality views (lower left), general metadata table (upper right), and map view (lower right)

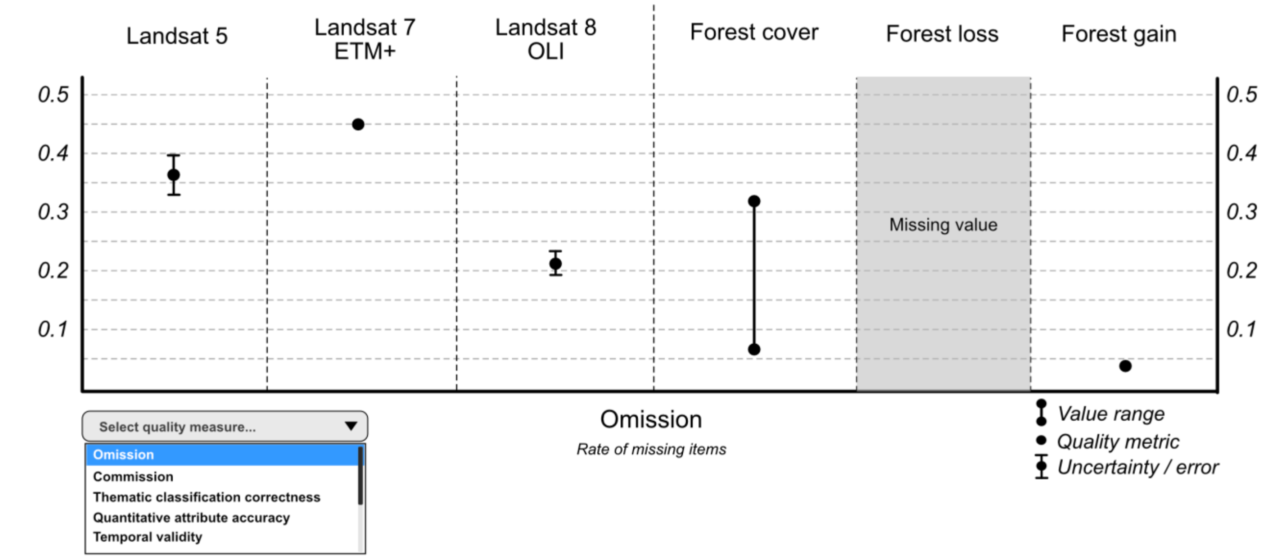


Figure 3: Visualizing the quality element *Omission* for six geospatial datasets

The dashboard consists of four linked sections, so-called widgets, to display the provenance graph, the quality information, general metadata and the data itself in a map (Fig. 2). The widget's size and position are fully customizable to the user's need.

The **provenance widget** (Fig. 2, top-left) shows the origin of a geospatial dataset as interactive provenance graph, including all datasets and processing steps of the workflow, using PROV-O. The graph serves as starting point for the evaluation, providing an overview as stated in Shneiderman's Mantra. Implemented as interactive graph, it facilitates users to select a certain dataset (click on entity), for which general metadata will be displayed in the metadata widget and to choose several datasets, for which quality information should be displayed in the quality widget below.

The metadata widget (Fig. 2, top-right) shows general metadata for the selected dataset, based on ISO 19115:2014 / GeoDCAT elements (Perego et al., 2017), e.g. spatial and temporal extent and resolution. We implement several columns as interactive links, supporting the users to navigate to (1) external information sources, e.g. documentations, (2) to registries or ontologies, e.g. to get further descriptions for theme or CRS, or (3) to navigate through dataset's

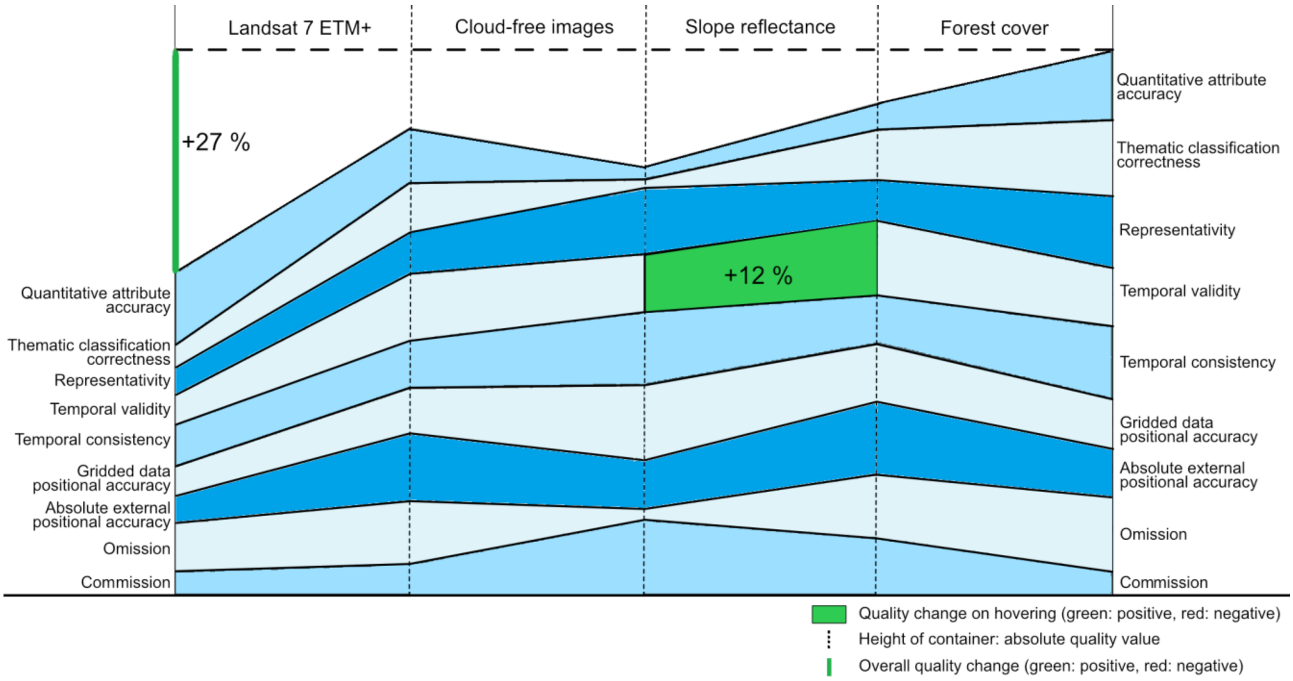


Figure 4: Cumulative graph for nine normalized quality measures

hierarchy (via parent) or provenance (via derived datasets).

The quality widget (Fig. 2, bottom-left) facilitates the evaluation of quality information, e.g. by classifying datasets according to their quality measures. This complex widget provides three different quality views for different user needs.

The **matrix view** (Fig. 2, bottom-left) displays quality information as colored table. It provides a compact view for quality measures using different scales, e.g. nominally or cardinally scaled, and different levels of detail: a certain data value, a range or errors (e.g. Fig. 2, commission). Each row of the matrix displays values for a certain quality measure for the selected datasets. Users can adapt the visualization to meet their use cases by defining individual class breaks for the color encoding. Missing values are shown as grey boxes.

The **simple graph view** (Fig. 3) and the **cumulated graph view** (Fig. 4) facilitate the visual comparison of quality information for selected datasets and combined with the provenance graph the change of quality values for selected measures along the workflow. Both interactive graphs provide detailed information as percentages or raw values when hovering.

The **simple graph view** (Fig. 3) supports users to get detailed information measures (see Shneiderman's Mantra) for a selected quality measure, in particular for comparison. Here, we can visualize quality measures in different levels of detail, e.g. a value as point, a range as line, errors as whiskers, and missing values / qualitative values as grey column.

The **cumulative graph view** (Fig. 4) shows the overall quality change along the workflow and thus, supports the user to evaluate the impact of certain dataset's on the overall quality. To provide such visualization, quality information for each quality measure and each dataset need to be available. Furthermore, all values for the selected quality measures need to be normalized. For our concept, we define to (1) remove quality measures with missing values from the graph and (2) assume that normalized values, at least for some measures, like omission and commission, do exist – otherwise the graph will not be generated.

The **map widget** (Fig. 2, bottom-right) visualizes the geospatial data facilitating the users to evaluate the quality of a certain region using quality-based previews. The interactive map widget uses OpenStreetMap[3] as basemap, showing the geospatial data – or, if not available, the spatial extent - on top and supports interactive map navigation as well as feature information by clicking on the map.

4 Conclusion and future work

To make research data in ESS findable and reusable (focus F and R of the FAIR principles), structured metainformation needs to be provided. We propose a user-friendly interactive Geo-dashboard concept to facilitate the evaluation of fitness for use of ESS research data by providing linked visualizations for metadata and data.

Quality information is essential to evaluate the fitness for use of data. As described in section 2.2, quality information is often not available or does not meet the data consumer's needs in terms of level of detail. With our Geo-dashboard, we make data producers more sensitive to collect and provide quality information. Therefore, we built our Geo-dashboard based on structured metadata and facilitate linking to tools for the automated extraction of metadata.

Quality information is often provided in heterogeneous sources. By using standard interfaces and implementing the linked data concept, we support linked quality information from several sources and provide interactive links to relevant sources, e.g. supplemental material or websites, in the Geo-dashboard. Thus, we provide with our Geo-dashboard a central access point for data consumers to evaluate quality information.

Even within a certain domain, quality information is described with different measures. In our Geo-dashboard, we partly address this by focussing on measures that can be automatically extracted from the data first, ensuring comparison for datasets. We further design a linked data quality model, enabling the mapping of the measures in our Geo-dashboard. However, enabling the comparison of qualitative and quantitative values is still future work.

Further, we are going to develop visualization concepts for complex quality measures, e.g. thematic resolution or uncertainty scores, which require complex visualization types, such as network or hierarchical graphs and views for related ground truth data.

[3] https://www.openstreetmap.org/

With our Geo-dashboard, we sensitize data producers and consumers for the importance of provenance and quality information. We guide data consumers on how to evaluate data by using an interactive provenance graph to get an overview, and classifying quality information for interactive visualizations on several levels of detail to meet specific use case requirements.

Software and Data Availability

Dataset *Global Forest Change 2000–2019:*

- Download: http://earthenginepartners.appspot.com/science-2013-global-forest/download_v1.7.html.
- Supplementary Material: https://science.sciencemag.org/content/suppl/2013/11/14/342.6160.850.DC1

Dataset *Global Spatially-Disaggregated Crop Production Statistics Data for 2010 Version 2.0:*

- Download: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PRFF8V
- Publication: https://essd.copernicus.org/articles/12/3545/2020/essd-12-3545-2020-discussion.html

Dataset *Crop production in EU standard humidity:*

- Download: https://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=apro_cpsh1&lang=en.
- Metadata: https://ec.europa.eu/eurostat/cache/metadata/en/apro_cp_esms.htm

Dataset *CropScape - Cropland Data Layer*:

- Download: https://nassgeodata.gmu.edu/CropScape/

Quality measures in the geo-dashboard screenshot (Fig. 2) are generated randomly for visualization purpose.

Acknowledgement

The authors thank Lukas Egli, Julia Fischer, and Juliane Groth, Julius Schmiedt (Helmholtz Centre for Environmental Research) for their inputs on information needs and for supporting quality and provenance information gathering.

The presented recommendations have been developed in the project GeoKur. GeoKur is granted by the BMBF (Federal Ministry of Education and Research) under number 16QK04A.

References

Bernasocchi, M., Coltekin, A. and Gruber, S.: An Open Source Geovisual Analytics Toolbox for Multivariate Spatio-Temporal Data in Environmental Change Modelling, ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. I-2, 123–128. https://doi.org/10.5194/isprsannals-I-2-123-2012, 2012.

Closa, G., Masó, J., Proß, B. and Pons, X.: W3C PROV to describe provenance at the dataset, feature and attribute levels in a distributed environment, Computers, Environment and Urban Systems 64, 103–117.https://doi.org/10.1016/j.compenvurbsys.2017.01.008, 2017.

Devillers, R., Bédard, Y., Jeansoulin, R. and Moulin, B.: Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data, International Journal of Geographical Information Science, 21 (3), 261–282. https://doi.org/10.1080/13658810600911879, 2007.

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.

Di, L. and Yue, P.: Provenance in Earth Science Cyberinfrastructure, A White Paper for NSF EarthCube, 2011.

Hansen, M. C., Potapov, P. V., Moore, R.; Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., Townshend, J. R. G.: High-resolution global maps of 21st-century forest cover change, Science (New York, N.Y.) 342 (6160), 850–853. https://doi.org/10.1126/science.1244693, 2013.

Henzen, C., Mäs, S. and Bernard, L.: Provenance Information in Geodata Infrastructures, Vandenbroucke, D., Bucher B. and Crompvoets, J.: Geographic Information Science at the Heart of Europe. Springer International Publishing, 133–151, 2013.

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.

Jing, C., Du, M., Li, S. and Liu, S.: Geospatial Dashboards for Monitoring Smart City Performance, Sustainability 11 (20), 5648. https://doi.org/10.3390/su11205648, 2019.

Magagna, B., Goldfarb, D., Martin, P., Atkinson, M., Koulouzis, S. and Zhao, Z.: Data Provenance. In: Zhiming Zhao und Margareta Hellström: Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, Bd. 12003. Springer International Publishing, 208–225, 2020.

Moreau, L. and Missier, P.: PROV-DM: The PROV Data Model. W3C Recommendation 30 April 2013, https://www.w3.org/TR/prov-dm/#prov-notation, 2013.

Perego, A., Cetl, V., Friis-Christensen, A., Lutz, M.: GeoDCAT-AP: Representing geographic metadata by using the "DCAT application profile for data portals in Europe". In: Smart Descriptions & Smarter Vocabularies (SDSVoc) workshop, Amsterdam. https://publications.jrc.ec.europa.eu/repository/handle/JRC107410, 2020.

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

Shneiderman, B.: Information visualization. White Paper.http://www.cs.umd.edu/hcil/members/bshneiderman/ivwp.html, 1997.

Simmhan, Y. L., Plale, B. and Gannon, D.: A survey of data provenance in e-science, SIGMOD Rec. 34 (3), 31–36. https://doi.org/10.1145/1084805.1084812, 2005.

Ślusarski, M. and Jurkiewicz, M.: Visualisation of Spatial Data Uncertainty. A Case Study of a Database of Topographic Objects, IJGI 9 (1), 16. https://doi.org/10.3390/ijgi9010016, 2020.

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