COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

Call to Action for Global Access to and Harmonization of Quality Information of Individual Earth Science Datasets

SPECIAL COLLECTION: OPEN SCIENCE FOR A GLOBAL TRANSFORMATION

ESSAY

]u[ubiquity press

GE PENG
ROBERT R. DOWNS
CARLO LACAGNINA
HAMPAPURAM RAMAPRIYAN
IVANA IVÁNOVÁ
DAVID MORONI
YAXING WEI
GILLES LARNICOL
LESLEY WYBORN
MITCH GOLDBERG
JÖRG SCHULZ
IRINA BASTRAKOVA
ANETTE GANSKE
LUCY BASTIN
SIRI JODHA S. KHALSA
MINGFANG WU
CHUNG-LIN SHIE
NANCY RITCHEY
DAVE JONES
TED HABERMANN
CHRISTINA LIEF
IOLANDA MAGGIO
MIRKO ALBANI
SHELLEY STALL
LIHANG ZHOU
MARIE DRÉVILLON
SARAH CHAMPION
C. SOPHIE HOU
FRANCISCO DOBLAS-REYES
KERSTIN LEHNERT
ERIN ROBINSON
KAYLIN BUGBEE

*Author affiliations can be found in the back matter of this article

CORRESPONDING AUTHOR:
Ge Peng
Earth System Science Center/NASA MSFC IMPACT, The University of Alabama in Huntsville, Huntsville, AL, USA
gpeng93@gmail.com

KEYWORDS:
Data Quality; Quality Dimension; Earth Science Information; Interoperability; FAIR; Stewardship

TO CITE THIS ARTICLE:
Peng, G, Downs, RR, Lacagnina, C, Ramapriyan, H, Ivánová, I, Moroni, D, Wei, Y, Larnicol, G, Wyborn, L, Goldberg, M, Schulz, J, Bastrakova, I, Ganske, A, Bastin, L, Khalsa, SJS, Wu, M, Shie, CL, Ritchey, N, Jones, D, Habermann, T, Lief, C, Maggio, I, Albani, M, Stall, S, Zhou, L, Drévillon, M, Champion, S, Hou, CS, Doblas-Reyes, F, Lehnert, K, Robinson, E, Bugbee, K. 2021. Call to Action for Global Access to and Harmonization of Quality Information of Individual Earth Science Datasets. *Data Science Journal*, 20: 19, pp. 1–15. DOI: *https://doi.org/10.5334/dsj-2021-019*

ABSTRACT

Knowledge about the quality of data and metadata is important to support informed decisions on the (re)use of individual datasets and is an essential part of the ecosystem that supports open science. Quality assessments reflect the reliability and usability of data. They need to be consistently curated, fully traceable, and adequately documented, as these are crucial for sound decision- and policy-making efforts that rely on data. Quality assessments also need to be consistently represented and readily integrated across systems and tools to allow for improved sharing of information on quality at the dataset level for individual quality attribute or dimension. Although the need for assessing the quality of data and associated information is well recognized, methodologies for an evaluation framework and presentation of resultant quality information to end users may not have been comprehensively addressed within and across disciplines. Global interdisciplinary domain experts have come together to systematically explore needs, challenges and impacts of consistently curating and representing quality information through the entire lifecycle of a dataset. This paper describes the findings of that effort, argues the importance of sharing dataset quality information, calls for community action to develop practical guidelines, and outlines community recommendations for developing such guidelines. Practical guidelines will allow for global access to and harmonization of quality information at the level of individual Earth science datasets, which in turn will support open science.

1. NEEDS FOR CURATING AND SHARING DATASET QUALITY INFORMATION

Knowledge about the quality of data and metadata at the individual dataset level, such as their accuracy, completeness, timeliness, and provenance, is imperative for establishing trust and supporting informed decisions on and accurate (re)use of data (Digital Science et al. 2019). To support effective decision- and policy-making processes, dataset quality information, such as information about the state of data, metadata, documentation, software, workflows, and tools used for producing and managing the data; and how the data being curated and serviced, should be consistently captured in the metadata and be a part of the ecosystem that supports open science.

Assessment of data quality is key for ensuring that the available data and information are credible and such assessments are essential when establishing trust for reuse of the data (Callahan 2017). Trusted data are perceived as worthy of use in decision making environments where the metadata is sufficient to adequately describe the data, e.g., information about the dataset author and data timeliness. Describing the quality of a data product and providing access to such quality information can support potential users of a particular dataset to determine whether it is appropriate for their planned usage, i.e., fitness for purpose.

A systematic analysis was carried out under European Commission to estimate the annual cost of not sharing data, which was found to be a minimum of €10.2bn per year (European Commission and PwC EU Services 2018). On average, data scientists spend 60–70% of their effort on dealing with data quality related issues (e.g., Press 2016). Thus, not sharing data quality information will compound that loss, especially on productivity lost due to redundancy in assessing data quality.

Although the importance of access to quality information is well recognized, methodologies for an evaluation framework and presentation of resultant quality information to end users may not have been comprehensively addressed. Access to this information is especially important for enabling data to be findable, accessible, interoperable, and reusable (FAIR). The FAIR data guiding principles defined by Wilkinson et al. (2016) emphasize the importance of data sharing in a machine-friendly environment. Since their inception, the FAIR data guiding principles have been adopted by global entities and have had a major impact in promoting data sharing and reuse globally (e.g., G20 Leaders 2016; Australia FAIR Access Working Group 2017; European Commission 2018, 2020; Mons 2018; U.S. Public Law 115–435 2019; CODATA 2019). However, the FAIR data guiding principles are somewhat limited in that they call for only meta(data) to be associated with detailed provenance and “richly described with a plurality of accurate and relevant attributes” (Wilkinson et al. 2016). They do not explicitly address the sharing of quality information of meta(data). For example, if the FAIRness of a dataset has been evaluated, what can be done to ensure that the method used and assessment results are readily findable, accessible, and (re)usable to end users? What can be done to ensure that the quality information can be readily integrated across different tools and systems within and out of individual organizations?

Therefore, building on the direction that the FAIR guiding principles have provided for data sharing, we would like to go one step further and call for all dataset quality information to be FAIR to improve the sharing of quality information of individual datasets. The optimal goal of this call is to allow for global access to and global harmonization of quality information of individual datasets as an important step towards open science in both machine- and human-friendly environments. FAIR dataset quality information will also help improve data sharing as discussed in more detail in Section 4.

Dataset quality can be affected by activities that are conducted throughout the data lifecycle. For example, lack of a quality assurance and control (QA/QC) procedure when generating the data product may impact the scientific quality of the data, while lack of the information about the QA/QC procedure will influence the quality of metadata and potentially reduce the level of user's confidence in trusting and using the data. In addition, data can be corrupted at any stage of the data lifecycle. In order for users and decision-makers to trust the data and the scientific findings resulting from analysis of this data, it is essential to establish and demonstrate, in a consistent and transparent way, the credibility of not only the data itself, but also the whole process of producing, managing, stewarding, analyzing, and servicing the data (Tilmes et al. 2015a). Therefore, a data lifecycle approach to data quality assessment is necessary. Furthermore, a data lifecycle approach to data quality assessment can facilitate effective recording of data quality information during various data lifecycle activities. The details of such information could be lost during later stages in the data lifecycle if they are not recorded in a timely fashion when the data quality events or assessments occurred. Moreover, managing quality throughout the entire dataset lifecycle is imperative for ensuring that the information and knowledge gained are not contaminated by inaccurate or corrupted data, as well as for facilitating accurate uncertainty estimates in the derived analyses. The value of lifecycle approaches to data quality has been recognized for various kinds of data, including remote sensing observations (Barsi et al. 2019), health services (Kahn et al. 2015), and health and biomedical citizen science (Borda et al. 2020). Data lifecycle approaches to quality assessment also could be informed by lifecycle approaches to software quality (Lenhardt et al. 2014).

Another example where verifiable and consistent quality-controlled data are becoming increasingly important is in the domain of the emerging technologies of machine learning (ML) and artificial intelligence (AI). These are flourishing as useful and effective tools to uncover or gain new knowledge from various domains of Earth science data (see an overview by Maskey et al. 2020). However, any sound analysis needs to build on reliable data (Breck et al. 2019; Shen and Sanghavi 2019). Good quality data allows people and organizations to increase trust in data and analysis, apply robust ML models, increase revenue and business performance. Poor quality data could lead to a negative economic impact and even loss of human lives. Without good definitions and procedures for working with data quality, critical operations could be at risk. Organizations should be aware of the quality of the data used for AI, and ensure that only data fit for purpose within acceptable risks be used (e.g., High-Level Expert Group on Artificial Intelligence 2018).

Therefore, it is crucial to consistently record, curate, and represent quality information of individual datasets, and make it readily available and integrable. However, dataset quality information is not routinely curated and much less represented in a human- and machine-readable manner, despite the fact that international standards for describing the quality of geographic data have been in place since 2003 (e.g., ISO 19157: 2013; ISO 19115-1:2014). The lack of adoption of one or more data quality standards may in part reflect the diversity of approaches, availability of resources, technologies, networks, and research questions of investigators (Leonelli 2017), as well as the context for the planned purpose and use of the data (Canali 2020; Illari 2014). Lack of motivation to document quality can be caused by the lack of prescriptiveness of existing standards – documentation of data quality metadata has always been optional in the ISO 19100 series and as of 2014, the ISO 19115-1 standard for metadata does not define a minimum set of discovery metadata, which used to suggest at least one data quality element (the provenance of a dataset).

Several other issues also may contribute to challenges for assessing and reporting data quality, and ultimately for the curation of dataset quality information. A frequently cited barrier against documenting the quality of spatial data is that it requires special domain-expert technical knowledge and across-domain expert knowledge integration, while documenting general metadata can be done automatically or by non-specialists (Coetzee 2018; Peng et al. 2020a). Oftentimes, knowledge of the quality information resides with domain experts who need to contribute to the information but may not be fully aware of the importance of their contributions as the impact comes later on. The benefit of such knowledge may also be substantially less at an individual level than for the common good.

Investigating barriers to assessing and reporting data quality information in medical bioinformatics, Callahan et al. (2017) identified issues, at both organizational and individual levels, that contribute to deficiencies in quality assurance. Such organizational issues include inadequate support, unclear expectations and insufficient training such as the absence of best practices. Individual ones include consequential as well as process issues, such as unresolvable conditions.

All these issues and challenges have led us to make this call-to-action statement to bring awareness about the needs and benefits of sharing quality information at the individual dataset level and rally people behind an important and challenging international and cross-disciplinary community effort.

The paper is organized as follows. Section 2 provides definitions of some key terms used in this paper. Section 3 touches on multiplicity of dataset quality attributes and dimensions, which is one of the main challenges for assessing and curating dataset quality information. Some potential benefits of having FAIR dataset quality information are described in Section 4, along with a brief summary of a real-life use case of how pre-vetted, timely and readily usable data and quality information is critical to disaster response efforts conducted by utility companies, with lives and billions of dollars at stake. A call for global community guidelines is then made in Section 5 which also includes an outline of community recommendations for developing such guidelines. A summary in Section 6 concludes the paper.

2. TERMS AND DEFINITIONS

In this paper, data are representations of observations, objects, or other entities and can refer to anything that is collected, observed, generated or derived, and used as a basis for hypothesis testing, reasoning, discussion, or calculation. Observed data include in situ and remotely sensed measurements. In situ measurements can be from weather stations, rain gauges, buoys, or autonomous vehicles/vessels, while remotely sensed data can be from instruments on satellites or aircrafts. Generated data can be results from a numerical model (e.g., a climate model) or a statistical model (e.g., a linear regression model). Model data can be analyses, predictions, or projections. Derived data can be produced from raw measurements or other data products. For example, atmospheric reanalysis data is one type of derived data that combines modeled data and observations via data assimilation to produce a dynamically consistent estimate of the atmospheric state.

Dataset refers to an identifiable collection of data (ISO 19115-1 2014), and it can be published or curated by a single agent (W3C 2020). A dataset can be the digital rendition of a data product of a given version of an algorithm or model or experiment. A dataset may contain one or many data files or records in a database in an identical format, having the same variable(s) and product specification(s).

Dataset quality information consists of information about the quality of data, metadata and documentation. Documentation can include descriptions of measurement methods and instruments, software, provenance, as well as that on the state of practices, workflows, frameworks, tools, and systems associated with the dataset and production, data and quality management, data services and usage, customer support and user engagement.

3. MULTI-DIMENSIONALITY OF DATASET QUALITY

A dataset is associated with a number of distinct quality attributes or characteristics. For example, Wang and Strong (1996) identified over 179 individual data quality attributes through a survey from a data consumer perspective, attributes such as accuracy, correctness, freedom from bias. Furthermore, dataset quality attributes can be categorized into different perspectives or dimensions with emphasis on certain quality attributes (e.g., Wang and Strong 1996; Redman 1996; Lee et al. 2002; Wilkinson et al. 2016; Ramapriyan et al. 2017). For instance, Wang and Strong (1996) prioritized the 179 quality attributes down to 15, and categorized them into four perspectives, i.e., intrinsic, contextual, representational, accessibility. Redman (1996) defined accuracy, completeness, consistency, and currency as four quality dimensions of data values. Alternatively, based on the full dataset lifecycle, Ramapriyan et al. (2017) categorized quality attributes into four quality dimensions: science, product, stewardship, and service. These various groups of quality attributes are explicitly listed to demonstrate that they can differ greatly depending on the different perspectives.

Assessment models are developed in Earth science to measure the maturity of different quality perspectives and dimensions at the dataset or collection level (see an overview by Peng 2018). However, there are very limited and sparse actionable guidelines on how to curate and represent dataset quality information in a way that is consistent with FAIR principles for improved sharing.

Currently dataset quality information, when available, is published in science journals that are text-based and cannot be readily integrated into data management and stewardship processes or across different systems. In addition, dataset quality information needs to be readily understandable by both machine and human end users, including those who plan to use the described data as well as by those who are trying to determine whether the data are appropriate for their intended use. Therefore, we need to converge towards harmonized approaches for curating dataset quality information in a way that is consistent with FAIR guiding principles to effectively enable global access of this information.

4. POTENTIAL BENEFITS OF FAIR DATASET QUALITY INFORMATION

The FAIR data guiding principles emphasize the importance of data sharing by ensuring that data and data descriptions (metadata) are findable, accessible, interoperable, and reusable (Wilkinson et al., 2016). Findable data are discovered and understood by a search agent (e.g., search engine or human user). Accessible data are rendered and used by machine and human end users via standard protocols within use and access constraints. Interoperable data can be readily used in conjunction with other data products or services and can also be integrated with other data to create new data products or services. Reusable data can be used, under the proper license and given well-documented dataset provenance, by diverse audiences beyond those who were initially envisioned as potential users by the original data producer(s).

Applying the FAIR guiding principles when curating and representing dataset quality information can help ensure the information is optimal for sharing and enabling global access. Successful reuse of data often depends critically on a potential user being able to access information about the quality of the data and determine its fitness for the intended application. Information about data quality also contributes to or improves the FAIRness of a dataset. For example, when data can be discovered based on information about certain quality attributes, the findability of the data is improved for users who need data that contain such attributes, and further, the quality information supports users to assess the relevance of a discovered dataset to a research or operational need. Global access to and sharing of dataset quality information will help improve data transparency and enable reproducibility, which is especially critical to highly-influential data that are used for decision-making (e.g., Tilmes et al. 2015b).

Consistently curating and representing dataset quality information by following the FAIR principles could eventually lead to standardization within and across organizations, tools, and systems, which in turn will lead to harmonization of the information. In addition, describing quality information using standardized formats, schemas, and terminology with controlled vocabularies improves the interoperability and reusability of the data.

Appendix A describes in detail a real-life use case of how trusted, timely and readily-integrable data and quality information is critical to disaster responses by utility companies with billions of dollars at stake. For disaster response managers, any information needs to be trusted and readily integrated, and understood in layman's terms in a matter of a minute. Accuracy and timeliness of data and information is extremely important, and any datasets that are selected to be a part of their decision-making processes need to be trusted. Managers will not trust just any available datasets, since their decisions can have an impact on the safety and survival of at-risk populations, can cost up to millions of dollars, and influence the reputation of their organizations.

For this use case, datasets are pre-vetted with an operational readiness level (ORL) ranking that is readily available and easily understood by decision makers who are generally non-data experts. An assigned ORL enables such decision makers to rapidly trust datasets. Data and information are integrated into a system which underpins an easy-to-understand dashboard for disaster response managers to allow them to make decisions promptly. Thus, providing quality information along with the data establishes the trust needed for supporting such potentially life-saving emergency response activities, and maximizes the benefit of sharing data. More detailed information can be found in Appendix A.

Pre-vetting datasets and developing the dashboard requires years of work and ongoing effort in addition to cultivated human relationships. Readily-available and consistently curated quality information of an individual dataset will help improve the process of establishing trust necessary to support tools and services provided to disaster responses, saving time and money. It will also support effective (re)use of the dataset for other applications, resulting in wide community utilization and therefore maximizing the value of the dataset.

5. CALL FOR GLOBAL EARTH SCIENCE COMMUNITY GUIDELINES

The voluntary sharing of meteorological observations and scientific knowledge among countries and between individual researchers and organizations has been going on for over a century (e.g., WMO 2019a). Since 1991, the World Meteorological Organization (WMO) has passed several resolutions for sharing essential and high value data, including basic weather, hydrological, and climate data through a series of WMO resolutions (WMO 1991; 1999; 2015). Guidelines on acquisition, quality assurance and control of meteorological station and model data were developed (e.g., WMO 1986; 2004; 2019b; Taylor 2012; Stockhause et al. 2012).

Sharing and reusing data is a big challenge but significant progress has been made collectively by the Earth science community over the last few decades. For the first time, WMO (2019c) issued a regulatory technical recommendation on managing climate data to be more accessible and usable. Recognizing the need to address ever increasing data volume and variety of data types across disciplines, WMO (2019d) called for one unified data policy to support global environmental data sharing and open science. United Nations Committee of Experts on Global Geospatial Information Management (UN-GGIM) developed a strategic framework with recommendations for a geospatial community with a strong focus on data, data quality and standards. These recommendations will help countries make their geospatial data and information reliable, accessible, and easy to use (UN-GGIM 2018; 2019). Success stories include greatly enhanced accessibility, usability, and interoperability of observational and climate model data through coordinated community efforts such as the Observations for Model Intercomparisons Project (Obs4MIP; Ferraro 2015) and the Coupled Model Intercomparison Projects (CMIP; Eyring 2016). CMIP6 data are published with persistent identifiers such as Digital Object Identifiers (DOIs) (Stockhause and Lautenschlager 2017). Sharing and reusing dataset quality information is an even bigger challenge that requires global community effort. Issuing a call to the community for such an effort is the first step towards achieving that goal.

To explore the needs for dataset quality information, approaches and challenges for consistently evaluating and representing the quality information, a one-day virtual workshop was held on Monday July 13, 2020. Domain experts from 9 countries across America, Europe and Oceania have participated in the workshop. It was followed by a report-out session on Wednesday July 22, 2020, during the virtual Earth Science Information Partners (ESIP) 2020 Summer Meeting (SM20), July 14–24, 2020. Additional information can be found in the workshop report by Peng et al. (2020a).

A total of 14 presentations from organizations across 9 countries were given during the two live sessions of the workshop and the ESIP SM20 report-out session. Presenters summarized the data quality assessment approaches that have been developed and/or adopted by organizations representing the scope of national and international Earth science data producers, data management stewardship programs, data and service centers as well as data and information providers. Participants also included data users from academic and private sectors.

The needs, challenges and approaches of consistently curating digital Earth science data and products were discussed during the live sessions as well as online during the following weeks. There is an overwhelming need for developing actionable community guidelines. The scope and path forward for developing such community guidelines for Earth science dataset quality information were also discussed. Key takeaways from the discussions are listed and described in the following subsections.

5.1. BUILT BY THE GLOBAL COMMUNITY

While needs are strong for practical community guidelines for curating data quality information, currently such information is very limited and sparse. Therefore, to ensure the relevance of such guidelines, it is crucial for the guidelines to be developed through a coordinated effort via an iterative process, leveraging the experiences and expertise of an international team consisting of interdisciplinary domain experts, and community best practices.

An international working group has thus been formed to develop practical community guidelines. The current members of the working group consist of data producers, publishers, managers and stewards from national science and/or data centers, repositories, as well as data users from the academic and private sectors. Collectively they bring together many years of valuable experience in production, management, services, and applications of various types of Earth science data, including satellite, in situ, and model data, along with knowledge of the challenges and best practices in their domains.

The membership of the working group is open for any domain expert who is willing and able to contribute. One can also support this effort by reviewing the draft of the guidelines or providing a use case. Interested parties are encouraged to contact the corresponding author of this article.

5.2. QUALITY-ATTRIBUTE AGNOSTIC GUIDELINES

As characterized in Section 2, assessing dataset quality is a multi-dimensional problem. The selection of the relevant attributes is context-dependent and it leads to different categorizations and practical dimensions (Redman 1996). The complexity exists even within one discipline that a quality attribute can have different definitions, and be measured and represented differently. An example of this is data uncertainty as explored by Moroni et al. (2019).

The selection of the relevant attributes is context-dependent because datasets are often crafted for specific designated communities. Traditionally, designated communities of data consumers are domain literate and have some familiarity with the scientific context, data generation, or intended data use. However, with the increasing availability of data today, the existence of interested audiences with a variety of scientific backgrounds outside the domain of data collection must be taken into consideration in order for scientific knowledge to be widely conveyed and understood. Designated communities may also change over time (Baker et al. 2016). These guidelines are intended to be more general and agnostic of the quality attributes, how to tailor the dataset quality information to the designated community is left to the specific entities who serve that particular community.

Therefore, the guidelines aim to equip data consumers with readily available information that is consistently curated and disseminated, in a way that can be easily found, accessed, understood by end-users and integrated across tools and systems, regardless of the quality attribute and the assessment approach.

5.3. COMMUNITY CONSENTED TERMINOLOGY FOR ENHANCED INTEROPERABILITY

For a given quality dimension, consistency in various components and attributes across entities in each component, namely, semantic and structured consistency as defined by Redman (1996), is important to generating machine-actionable quality information. Common terminology is necessary for integrating data and information across workflows, tools and systems, as well as for curating and representing dataset quality information. Moreover, terms should be defined, and, ideally, referenced with a persistent identifier, for all stages of a dataset lifecycle.

5.4. CONTINUOUS ENGAGEMENT WITH STAKEHOLDERS

The guidelines are being developed through an iterative process to allow for feedback from the community and all stakeholders, including those who contribute to the acquisition, curation, dissemination, and application of data. Continuous community engagements are planned by means of informal updates to various working groups and formal presentations to the targeted stakeholders, including those to the American Geophysical Union (AGU) community at the 2020 AGU Fall Meeting (Peng et al. 2020b), the ESIP community at its winter meeting in January 2021 (Peng et al. 2021), and the European Geosciences Union community at its general assembly in April 2021 (Lacagnina et al. 2021). Additionally, reviews of the guidelines draft are planned prior to its being baselined. The guidelines document will be a living document to allow for evolving community requirements and best practices.

5.5. LONG-TERM SUSTAINABILITY

It has been suggested by the participants that long-term sustainability should be planned for such community guidelines. Once baselined, the guidelines will be publicly accessible via an open science platform such as ESIP Figshare (esip.figshare.com) and/or Open Science Framework (osf.io) that provides a globally unique and persistent identifier with searching and sharing capability for the document. To ensure currency and timeliness of the guidelines, curation and revision planning is essential. These approaches will help with maintaining relevancy and long-term access to the guidelines.

6. SUMMARY

In summary, there are fundamental challenges in collecting, curating, and representing dataset quality information. Those challenges include:

- Dataset quality information is multi-dimensional with many quality attributes,
- Information about dataset quality traverses different knowledge domains and curating it requires cross-disciplinary collaborations and knowledge integration,
- Requirements may be different for different user applications as well as ways to assess dataset quality.

There are also strong community needs and benefits of having community-developed guidelines that are practical to implement. At the same time, guidelines for curating and representing dataset quality information are limited and sparse.

Recognizing the needs, challenges, and benefits of sharing information on quality at a dataset/collection level, interdisciplinary domain experts around the world have come together and called for community guidelines towards global access and harmonization of information on quality of individual Earth science datasets. The guidelines will be targeted at addressing these fundamental challenges, as well as others that are identified through the development activities.

The quality-attribute agnostic guidelines will be developed under community effort via an iterative process, leveraging the experiences and expertise of an international team of interdisciplinary domain experts and best practices. Description of what the quality attributes are, how the attributes are assessed, and what assessment approaches are utilized, should be included in relevant metadata or a document, preferably in a consistent way for transparency and enhanced usability. The guidelines will call for a machine-actionable mechanism to represent assessed results for enhanced interoperability across systems and disciplines.

By adopting the FAIR principles, the guidelines will help to ensure global access to the dataset quality information. Effective sharing of structured dataset quality information will help to move towards its global harmonization, which in turn will support (re)use of the data by both human and machine end users and therefore further enhance the value of the data.

As mentioned in section 5.1, an international FAIR dataset quality information working group has been formed under the leadership of ESIP Information Quality Cluster (IQC), the Barcelona Supercomputing Center (BSC) Evaluation and Quality Control (EQC) team, and the Australian Research Data Commons (ARDC) coordinated Australia/New Zealand Data Quality Interest Group (AU/NZ DQIG). The membership of this working group is open to any domain expert who is willing to contribute to the development effort. Development of the guidelines has begun and the outcomes will be reported in a follow-up paper. The guidelines will be primarily developed for the Earth Science community. They will, however, be general enough so that other disciplines can readily adapt them, which will further promote global access and harmonization of dataset quality information in supporting open science.

APPENDIX A

A REAL-LIFE USE CASE OF HARMONIZATION DATA AND QUALITY INFORMATION

In 2012, Superstorm Sandy impacted the Northeast United States at an angle perpendicular to the coastline, a worst-case scenario for landfalling storms on the East coast, resulting in widespread electric power outages. Utility crews from more than 40 states were activated in a massive response effort, called ‘mutual assistance’, to restore power to millions of customers in New York, New Jersey, Connecticut, Pennsylvania and Delaware.

The damage done by Superstorm Sandy in October 2012 was unprecedented in its size and scope. Approximately 10 million customers lost power across 24 states in the Northeast, Mid-Atlantic, and parts of the Midwest. In response, the electric power industry deployed an army of tens of thousands of restoration workers—representing 80 companies from almost every state and Canada. The goal was to restore power as quickly and safely as possible.[1]

At its highest point, utility response through mutual assistance reached $20,000,000 per hour as mutual assistance reached across 40+ states, including Hawaii, from where military aircraft were used to transport electric utility vehicles to the US East Coast. Because information could not be shared across platforms, across state borders and along the response pathway such as to weigh stations and toll booths, as well as coordination issues across Regional Mutual Assistance Groups (RMAGS),[2] response delays of up to 2-days resulted since responding utility trucks could not reach their destination at the expected time. In some cases, police escorts were used to assist cross-state transport. A two-day delay in response at $20,000,000/hour could result in more than $900,000,000 in wasted expense if applied to all responding utilities.

Sandy's impact on major population centers caused widespread interruption to critical water / electrical services and also caused 159 deaths (72 direct, 87 indirect). Sandy also caused the New York Stock Exchange to close for two consecutive business days. Such a closure was the first since March 12–13, 1888, when it was closed due to a major winter storm. Sandy's CPI-adjusted damage cost has been estimated at $74.8 Billion.[3]

In the aftermath of Sandy, Edison Electric Institute (EEI)[4] members also recognized the need to enhance and formalize the mutual assistance program for national events. In September 2013, EEI's Board of Directors approved a framework to institutionalize the lessons learned and best practices from Sandy to optimize restoration efforts following events that impact a significant population or several regions across the U.S. and require resources from multiple RMAGs.[5]

Having access to quality trusted information and the ability to share that information collaboratively across platforms and users in real-time was also found to be critical by the All Hazards Consortium, a 501(c)3 organization in Maryland that focuses on driving solutions through public-private sector cooperation. Tom Moran, Executive Director of AHC stated, "The sharing of trusted information drives decision making and shortens response time and is critical to improving efficiencies and lowering costs of response and recovery." In 2012, neither the technology, nor the focus on data quality were as mature as they had become in 2017 (and continue to improve) when CAT5 Hurricane Maria struck Puerto Rico.

2017 CATEGORY 5 HURRICANE MARIA IN PUERTO RICO

When Hurricane Maria made landfall in Puerto Rico in 2017 utilities responded under mutual assistance request to expedite the restoration of power in Puerto Rico. This required massive logistical hurdles to be overcome because of Puerto Rico's remote location when it comes to moving massive amounts of equipment. In 2019, the All Hazard Consortium's official report to DHS included operational impacts that a trusted information sharing environment was able to deliver to its users, including the utility sector and emergency managers. The development of the Sensitive Information Sharing Environment (SISE) has proven to be a valuable evolution since Sandy in 2012.

The SISE private sector operated RCOP (Regional Common Operating Picture, a.k.a. Daily Dashboard) has already had national impacts. This dashboard was developed to address several use cases in the electric sector. It is openly accessible to any user across any state or any sector in the United States. It is one of the few, if not the only, platforms operated by the private sector that allows the private sector to coordinate more effectively with multiple states and federal agencies during a regional multistate disaster.

[1] *https://www.eei.org/issuesandpolicy/electricreliability/mutualassistance/documents/ma_101final.pdf* (pg.4, accessed on February 20, 2021).

[2] *https://www.eei.org/issuesandpolicy/electricreliability/mutualassistance/documents/ma_101final.pdf*.

[3] *https://www.ncdc.noaa.gov/billions/events/US/2012* (Accessed on February 20, 2021).

[4] *https://www.eei.org/pages/default.aspx*.

[5] *https://www.eei.org/issuesandpolicy/electricreliability/mutualassistance/documents/ma_101final.pdf* (pg.5).

This has large operational impacts on business continuity, supply chain recovery, restoration, and overall infrastructure resilience. Many of the products that have been produced from SISE Use Case Committees had a dramatic impact operationally.

- It virtually eliminated power sector delays across the US during Hurricane Maria response;
- It saved thousands of private-sector manhours searching to find validated disaster documentation;
- It reduced confusion, wasted trips, storms costs, and improved operational coordination w/ multiple states;
- It Uses the GeoCollaborate/ESRI technology, now installed in the DHS NICC.[6]

STEPS THAT WERE TAKEN FROM 2012 TO 2017 TO LOWER THE COST OF THE RESPONSE

Because of the high impact event of 'Sandy' in 2012, steps were taken to improve efficiencies in mutual assistance and information sharing.

To prepare for severe storms and outage events that cross RMAG boundaries, such as Superstorm Sandy, we developed guidelines for responding to large, multi-RMAG or industry-wide National Response Events (NREs). The hurricane Sandy resulted in the single biggest post-storm restoration the electric power industry had ever undertaken. The damage was catastrophic and widespread. All RMAGs were impacted or involved in the restoration effort. Prior to Sandy, there was not a national framework in place to respond to storms of this magnitude. Determined to enhance the restoration process, EEI members are institutionalizing best practices based on the lessons learned from Sandy. The electric power industry is prepared for significant outage events and continues to improve its coordination and response and recovery efforts. Customers have increasing expectations and electricity dependence, and EEI is committed to making the mutual assistance process safe, efficient, equitable, and scalable.

Data is currently trusted in the emergency response sector because it is delivered by vetted organizations with established relationships. In many emergency response situations, data that is accessed quickly or through pathways not established prior to the event, may not have been of high quality. This is where decisions can be made based on seemingly trustworthy but actually low-quality data and information and resulting decisions and impacts can cost millions of dollars.

In addition, the system that delivers the 'pre-vetted' trusted data based on Operational Readiness Levels[7] also gains high trust levels from its users. StormCenter Communications has been a leader in this approach along with the ESIP and the All Hazards Consortium (AHC). The web based GeoCollaborate technology is designed to enable rapid generation of real-time trusted data sharing and collaborative environments that work across platforms to drive decision-making and is a core of the SISE 'Daily Dashboard'.

HUMAN RELATIONSHIPS, TRUST AND THE EVOLUTION OF DATA QUALITY PRINCIPLES TO ACCELERATE TRUST IN DATA TO ACCELERATE HIGH CONFIDENCE SITUATIONAL AWARENESS AND DECISION MAKING

With the development of the FAIR data principles, ORLs and other evolving data quality standards, it is critical to provide ample meta data that can drive confidence in the data. ORLs are an evolving use-case-based approach for indicating trustworthiness of the data, thus enabling decision makers, who are generally not data experts, to use datasets and make rapid decisions. Developed by the ESIP and AHC, ORLs are becoming a widely-adopted approach to validating the fitness for use of digital information for the mobilization of disaster response efforts.

The human relationship between information providers and decision makers has been essential to building trust. As Craig Fugate, former FEMA Administrator used to say, "You do not want to share business cards during a disaster," trusted relationships also connote service quality.

[6] *https://www.ahcusa.org/uploads/2/1/9/8/21985670/ahc_report_nipp_sise_operational_results__impacts_of_the_sise_2019.pdf* (pg.4; Accessed on February 20, 2021).

[7] *https://www.esipfed.org/orl* (accessed February 20, 2021).

You are more likely to trust the quality of the information provided by a familiar and trusted individual or organization than from someone you have never met. Trusted data providers can further elevate their trust levels by documenting the maturity of data quality processes and adhering to evolving FAIR data principles. This, hopefully, will be reflected in the metadata which can inform users of the lineage of the data.

ACKNOWLEDGEMENTS

The virtual pre-ESIP workshop held on July 13, 2020 was sponsored by ESIP and co-organized by the ESIP IQC and the BSC EQC team, in collaboration with the ARDC AU/NZ DQIG. An additional community engagement event was carried out by the AU/NZ DQIG prior to the pre-ESIP workshop. ESIP is primarily supported by the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA) and the United States Geological Survey (USGS). The technological and infrastructural support during the preparation and conduct of the workshop was invaluable. In particular, we thank Megan Carter, ESIP Community Director, for supporting us throughout the workshop and providing helpful advice during the planning stage of the virtual workshop, and ESIP Community Fellow, Alexis Garretson, for supporting the ESIP SM20 report-out session. We thank all participants for attending the pre-ESIP workshop and the ESIP SM20 session and contributing to productive discussions during the live sessions and the two weeks of the ESIP SM20 period. Portions of this work have been extracted from Peng et al. (2020a), which reported on the workshop and the ESIP SM20 report-out session. The Australian participants acknowledge the support of the ARDC. The constructive suggestions from two anonymous reviewers of Data Science Journal have helped improve the quality of the paper.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Ge Peng *orcid.org/0000-0002-1986-9115*
Earth System Science Center/NASA MSFC IMPACT, The University of Alabama in Huntsville, Huntsville, AL, USA

Robert R. Downs *orcid.org/0000-0002-8595-5134*
Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, NY, USA

Carlo Lacagnina *orcid.org/0000-0001-9434-9809*
Barcelona Supercomputing Center (BSC), Spain

Hampapuram Ramapriyan *orcid.org/0000-0002-8425-8943*
Ramapriyan, Science Systems and Applications, Inc., Lanham, MD, USA and NASA Goddard Space Flight Center, Greenbelt, MD, USA

Ivana Ivánová *orcid.org/0000-0001-6836-3463*
Curtin University, AUS

David Moroni *orcid.org/0000-0003-2994-557X*
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

Yaxing Wei *orcid.org/0000-0001-6924-0078*
Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Gilles Larnicol
Barcelona Supercomputing Center (BSC)/Magellium, Spain/France

Lesley Wyborn *orcid.org/0000-0001-5976-4943*
National Computational Infrastructure, Australian National University, ACT, AUS

Mitch Goldberg
National Environmental Satellite, Data, and Information Service (NESDIS), Silver Spring, MD, USA

Jörg Schulz *orcid.org/0000-0002-2744-5590*
European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT), Darmstadt, Germany

Irina Bastrakova *orcid.org/0000-0002-4643-7289*
Geoscience Australia, ACT, AUS

Anette Ganske *orcid.org/0000-0003-1043-4964*
Technische Informationsbibliothek (TIB), Hannover, Germany

Lucy Bastin *orcid.org/0000-0003-1321-0800*
Aston University, UK

Siri Jodha S. Khalsa *orcid.org/0000-0001-9217-5550*
Cooperative Institute for Research in Environmental Sciences, NSIDC, Boulder, CO, USA

Mingfang Wu *orcid.org/0000-0003-1206-3431*
Australian Research Data Commons, Melbourne, Australia

Chung-Lin Shie *orcid.org/0000-0002-1115-1029*
University of Maryland at Baltimore County, Baltimore, MD, USA and NASA Goddard Space Flight Center, Greenbelt, MD, USA

Nancy Ritchey *orcid.org/0000-0003-3939-6287*
NOAA's National Centers for Environmental Information (NCEI), Asheville, NC, USA

Dave Jones *orcid.org/0000-0003-4573-2400*
StormCenter Communications | GeoCollaborate, Halethorpe, MD, USA

Ted Habermann *orcid.org/0000-0003-3585-6733*
Metadata Game Changers, Boulder, CO, USA

Christina Lief
NOAA's National Centers for Environmental Information (NCEI), Asheville, NC, USA

Iolanda Maggio *orcid.org/0000-0001-9409-1289*
Rhea GROUP, La Piramide, Via di Grotte Portella 6/8, 00044 Frascati, Italy

Mirko Albani
European Space Agency, Frascati, Italy

Shelley Stall *orcid.org/0000-0003-2926-8353*
American Geophysical Union, Washington, DC, USA

Lihang Zhou *orcid.org/0000-0001-6232-2871*
National Environmental Satellite, Data, and Information Service (NESDIS), Silver Spring, MD, USA

Marie Drévillon *orcid.org/0000-0003-4586-8957*
Mercator Ocean International, France

Sarah Champion *orcid.org/0000-0002-5080-6286*
North Carolina Institute for Climate Studies, North Carolina State University, Asheville, NC, USA

C. Sophie Hou *orcid.org/0000-0002-8087-1775*
Ronin Institute, USA

Francisco Doblas-Reyes *orcid.org/0000-0002-6622-4280*
ICREA and Barcelona Supercomputing Center (BSC), Spain

Kerstin Lehnert *orcid.org/0000-0001-7036-1977*
Lamont-Doherty Earth Observatory of Columbia University, Palisades, NY, USA

Erin Robinson *orcid.org/0000-0001-9998-0114*
Metadata Game Changers, Boulder, CO, USA

Kaylin Bugbee *orcid.org/0000-0001-6733-5698*
NASA Marshall Space Flight Center, Huntsville, AL, USA

REFERENCES

Australia FAIR Access Working Group. 2017. Policy Statement on FAIR Access to Australia's Research Outputs. Version: Jan 2017. Available at: *https://www.fair-access.net.au/fair-statement*.

Baker, KS, Duerr, RE and **Parsons, MA.** 2016. Scientific Knowledge Mobilization: Co-evolution of Data Products and Designated Communities. *International Journal of Digital Curation*, 10(2): 110–135. DOI: *https://doi.org/10.2218/ijdc.v10i2.346*

Barsi, Á, Kugler, Z, Juhász, A, Szabó, G, Batini, C, Abdulmuttalib, H, Huang, G and **Shen, H.** 2019. Remote sensing data quality model: from data sources to lifecycle phases. *International Journal of Image and Data Fusion*, 10(4): 280–99. DOI: *https://doi.org/10.1080/19479832.2019.1625977*

Borda, A, Gray, K and **Fu, Y.** 2020. Research data management in health and biomedical citizen science: practices and prospects. *JAMIA Open*, 3(1): 113–25. DOI: *https://doi.org/10.1093/jamiaopen/ooz052*

Breck, E, Polyzotis, N, Roy, S, Whang, SE and **Zinkevich, M.** 2019. Data Validation For Machine Learning. *Proceedings of the 2nd SysML Conference*, Palo Alto, CA, USA.

Callahan, T, Barnard, J, Helmkamp, L, Maertens, J, Kahn, M. 2017. Reporting data quality assessment results: identifying individual and organizational barriers and solutions. *eGEMs*, 5(1). DOI: *https://doi.org/10.5334/egems.214*

Canali, S. 2020. Towards a Contextual Approach to Data Quality. *Data*. 5(4): 90. DOI: *https://doi.org/10.3390/data5040090*

Digital Science, Fane, B, Ayris, P, Hahnel, M, Hrynaszkiewicz, G, Baynes, G and **others.** 2019. The State of Open Data Report 2019. *Digital Science*. Report. DOI: *https://doi.org/10.6084/m9.figshare.9980783*

CODATA. 2019. The Beijing Declaration on Research Data. Version: 7 November 2019. Available at: *http://www.codata.org/uploads/Beijing%20Declaration-19-11-07-FINAL.pdf*.

Coetzee, S. 2018. Implementing Geospatial Data Quality Standards – Motivators and Barriers, *2nd International Workshop on Spatial Data Quality*, Valletta, Malta 6–7 February 2018, *https://eurogeographics.org/wp-content/uploads/2018/06/4-SDQ2018_Coetzee_V1e.pdf*.

European Commission. 2018. Turning FAIR into reality – Final Report and Action Plan from the European Commission Expert Group on FAIR data, European Commission: Brussels. DOI: *https://doi.org/10.2777/1524*

European Commission. 2020. Recommendations on FAIR Metrics for EOSC, European Commission: Brussels. DOI: *https://doi.org/10.2777/70791*

European Commission and PwC EU Services. 2018. Cost-benefit analysis for FAIR research data: Cost of not having FAIR research data. Version: March 2018. Available at: *https://op.europa.eu/en/publication-detail/-/publication/d375368c-1a0a-11e9-8d04-01aa75ed71a1/language-en*.

Eyring, V, Bony, S, Meehl, GA, Senior, CA, Stevens, B, Stouffer, RJ and **Taylor, KE.** 2016. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) Experimental Design and Organization. *Geoscientific Model Development*, 9(5): 1937–1958. DOI: *https://doi.org/10.5194/gmd-9-1937-2016*

Ferraro, R, Waliser, DE, Gleckler, P, Taylor, KE and **Eyring, V.** 2015. Evolving Obs4MIPs to Support Phase 6 of the Coupled Model Intercomparison Project (CMIP6). *Bull. Amer. Meteor. Soc.*, 96: ES131–ES133. DOI: *https://doi.org/10.1175/BAMS-D-14-00216.1*

G20 Leaders. 2016. G20 Leaders' Communique Hangzhou Summit. Version: 5 September 2016. Available at: *https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_16_2967*.

Illari, P. 2014. IQ: Purpose and Dimensions. In The Philosophy of Information Quality; Floridi, L, Illari, P, Eds.; Springer: Berlin, Germany pp. 281–302. DOI: *https://doi.org/10.1007/978-3-319-07121-3_14*

ISO 19115-1. 2014. Geographic Information—Metadata – Part 1: Fundamentals. Version: 2014–04. *International Organization for Standardization*. Geneva, Switzerland. Available at: *https://www.iso.org/standard/53798.html*.

ISO 19157. 2013. Geographic information—Data quality. Version: 2013–1. *International Organization for Standardization*. Geneva, Switzerland. Available at: *https://www.iso.org/standard/32575.html*.

Kahn, MG, Brown, JS, Chun, AT, Davidson, BN, Meeker, D, Ryan, PB, Schilling, LM, Weiskopf, NG, Williams, AE and **Zozus, MN.** 2015. Transparent reporting of data quality in distributed data networks. *Egems*, 3(1). DOI: *https://doi.org/10.13063/2327-9214.1052*

Lacagnina, C, Peng, G, Downs, RR, Ramapriyan, H, Ivanova, I, Moroni, DF, Larnicol, G, Wei, Y, Bastin, L, Ritchey, NA, Wyborn, LA, Shie, C-L, Habermann, T, Ganske, A, Champion, SM, Wu, M, Bastrakova, I, Jones, D and **Berg-Cross, G.** 2021. Towards Developing Community Guidelines for Sharing and Reusing Quality Information of Earth Science Datasets. *EGU General Assembly 2021*, Virtual, 19–30 April 2021, EGU21–23. DOI: *https://doi.org/10.5194/egusphere-egu21-23*

Lee, YW, Strong, DM, Khan, BK and **Wang, RY.** 2002. AIMQ: a methodology for information quality assessment, *Information & Management*, 40: 133–146. DOI: *https://doi.org/10.1016/S0378-7206(02)00043-5*

Leonelli, S. 2017. Global Data Quality Assessment and the Situated Nature of "Best" Research Practices in Biology. *Data Science Journal*, 16: 32. DOI: *https://doi.org/10.5334/dsj-2017-032*

High-Level Expert Group on Artificial Intelligence. 2018. Ethics guidelines for trustworthy AI. FUTURIUM – European Commission. Version: December 17, 2018. Available at: *https://ec.europa.eu/futurium/en/ai-alliance-consultation*.

Lenhardt, W, Ahalt, S, Blanton, B, Christopherson, L and **Idaszak, R.** 2014. Data management lifecycle and software lifecycle management in the context of conducting science. *Journal of Open Research Software*, 2(1). DOI: *https://doi.org/10.5334/jors.ax*

Maskey, M, Alemohammad, H, Murphy, KJ and **Ramachandran, R.** 2020. Advancing AI for Earth Science: A Data Systems Perspective. *EOS*, 101. DOI: *https://doi.org/10.1029/2020EO151245*

Mons, B. 2018. Data Stewardship for open science: implementing FAIR principles. 1st Edition. Chapman and Hall/CRC Press, *Taylor & Francis*, New York. 244 pp. Available at: *https://www.taylorfrancis.com/books/9781315380711*. DOI: *https://doi.org/10.1201/9781315380711-1*

Moroni, DF, Ramapriyan, H, Peng, G, Hobbs, J, Goldstein, JC, Downs, RR, Wolfe, R, Shie, C-L, Merchant, CJ, Bourassa, M, Matthews, JL, Cornillon, P, Bastin, L, Kehoe, K, Smith, B, Privette, JL, Subramanian, AC, Brown, O and **Ivánová, I.** 2019. Understanding the Various Perspectives of Earth Science Observational Data Uncertainty. *Figshare*. DOI: *https://doi.org/10.6084/m9.figshare.10271450*

Peng, G. 2018. The state of assessing data stewardship maturity – an overview. *Data Science Journal*, 17. DOI: *https://doi.org/10.5334/dsj-2018-007*

Peng, G, Lacagnina, C, Downs, RR, Ivanova, I, Moroni, DF, Ramapriyan, H, Wei, Y and **Larnicol, G.** 2020a. Laying the Groundwork for Developing International Community Guidelines to Effectively Share and Reuse Digital Data Quality Information – Case Statement, Workshop Summary Report, and Path Forward. *Open Science Framework*. DOI: *https://doi.org/10.31219/osf.io/75b92*

Peng, G, Lacagnina, C, Downs, RR, Ramapriyan, H, Ivanova, I, Moroni, DF, Larnicol, G, Wei, Y, Bastin, L, Ritchey, NA, Wyborn, LA, Shie, C-L, Habermann, T, Ganske, A, Champion, SM, Wu, M, Bastrakova, I, Jones, D and **Berg-Cross, G.** 2020b. Towards Developing Community Guidelines for Sharing and Reuse of Digital Data Quality Information. *AGU 2020 Fall Meeting*. Abstract 674372. Available at: *https://agu.confex.com/agu/fm20/meetingapp.cgi/Paper/674372*.

Peng, G, Lacagnina, C, Downs, RR, Ramapriyan, H, Ivanova, I, Moroni, DF, Larnicol, G, Wei, Y, Bastin, L, Ritchey, NA, Wyborn, LA, Shie, C-L, Habermann, T, Ganske, A, Champion, SM, Wu, M, Bastrakova, I, Jones, D, Hou, C-Y and **Berg-Cross, G.** 2021. An update on a community effort to promote global sharing of dataset quality information. *ESIP 2021 Winter Meeting*. Virtual.

Press, G. 2016. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. Forbes. Version: March 23, 2016. Available at: *https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=1ee368c06f63*.

Ramapriyan, H, Peng, G, Moroni, D and **Shie, C-L.** 2017. Ensuring and Improving Information Quality for Earth Science Data and Products. *D-Lib Magazine*, 23. DOI: *https://doi.org/10.1045/july2017-ramapriyan*

Redman, CT. 1996. Data quality of the information age. *Artech House*, Boston. 303 pp.

Shen, Y and **Sanghavi, S.** 2019. Learning with Bad Training Data via Iterative Trimmed Loss Minimization. Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, USA.

Stockhause, M, Höck, H, Toussaint, F and **Lautenschlager, M.** 2012. Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. *Geosci. Model Dev. 5*. DOI: *https://doi.org/10.5194/gmd-5-1023-2012*

Stockhause, M and **Lautenschlager, M.** 2017. CMIP6 Data Citation of Evolving Data. *Data Science Journal*, 16. DOI: *https://doi.org/10.5334/dsj-2017-030*

Taylor, KE, Stouffer, RJ and **Meehl, GA.** 2012. An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, 93(4): 485–498. DOI: *https://doi.org/10.1175/BAMS-D-11-00094.1*

Tilmes, C, Privette, AP, Chen, J, Ramachandran, R, Bugbee, KM and **Wolfe, RE.** 2015a. Linking from observations to data to actionable science in the climate data initiative. Proc. 2015 IEEE Geosci. and Remote Sensing Symposium, 26–31 July 2015, Milan, Italy. DOI: *https://doi.org/10.1109/IGARSS.2015.7326027*

Tilmes, C, Wolfe, RE, Duggan, B, Aulenbach, S, Goldstein, JC, Ma, X and **Zednik, S.** 2015b. Supporting trust with provenance of the findings of the national climate assessment. METHOD 2015: The 4th Intl. Workshop on Methods for Establishing Trust of (Open) Data. 11 Oct. 2015, Bethlehem, PA, USA. [Available at: *http://www.few.vu.nl/~dceolin/method2015/papers/METHOD_2015_paper_2.pdf*].

UN-GGIM. 2018. Integrated Geospatial Information Framework Part 1. United Nations Committee of Experts for Global Geospatial Information Management. Available at: *https://ggim.un.org/IGIF/part1.cshtml*.

UN-GGIM. 2019. Integrated Geospatial Information Framework Part 2. United Nations Committee of Experts for Global Geospatial Information Management. Available at: *https://ggim.un.org/IGIF/part2.cshtml*.

U.S. Public Law 115-435. 2019. Foundations for Evidence-Based Policymaking Act of 2018. Title II OPEN Government Data Act. Version: 14 January 2019. 115th U.S. Congress. Available at: *https://www.govinfo.gov/content/pkg/PLAW-115publ435/pdf/PLAW-115publ435.pdf*.

Wang, RY and **Strong, DM.** 1996. Beyond Accuracy: What Data Quality Means to Consumers. *Journal of Management Information Systems*, 12(4). DOI: *https://doi.org/10.1080/07421222.1996.11518099*

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A and **others.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: *https://doi.org/10.1038/sdata.2016.18*

WMO. 1986. Guidelines On The Quality Control Of Surface Climatological Data. *WMO/TD-No. 111*. Geneva, Switzerland: World Meteorological Organization. Available at: *https://library.wmo.int/doc_num.php?explnum_id=9205*.

WMO. 1991. Resolution 40 (Cg-XII) – WMO policy and practice for the exchange of meteorological and related data and products including guidelines on relationships in commercial meteorological activities. *WMO-No. 827*. Geneva, Switzerland: World Meteorological Organization. Available at: *https://www.wmo.int/pages/prog/hwrp/documents/wmo_827_enCG-XII-Res40.pdf*.

WMO. 1999. Resolution 25 (Cg-XIII) – Exchange of Hydrological Data and Products. Geneva, Switzerland: World Meteorological Organization. Available at: *https://www.wmo.int/pages/prog/hwrp/documents/Resolution_25.pdf*.

WMO. 2004. Guidelines on Quality Control Procedures for Data from Automatic Weather Stations. *Expert Team on Surface Technology and Measurement Techniques*, Geneva, Switzerland: World Meteorological Organization. Available at: *https://www.wmo.int/pages/prog/www/IMOP/meetings/Surface/ET-STMT1_Geneva2004/Doc6.1(2).pdf*.

WMO. 2015. Resolution 60 (Cg-17) – WMO Policy for the International Exchange of Climate Data and Products to Support the Implementation of the Global Framework for Climate Services. Geneva, Switzerland: World Meteorological Organization. Available at: *https://library.wmo.int/doc_num.php?explnum_id=4192*.

WMO. 2019a. Origin, impact and aftermath of WMO resolution 40. WMO-no 1244. Geneva, Switzerland: World Meteorological Organization. Available at: *https://library.wmo.int/doc_num.php?explnum_id=10140*.

WMO. 2019b. WMO Guidelines on Surface Station Data Quality Assurance for Climate Applications. Draft: April 5, 2019. Geneva, Switzerland: World Meteorological Organization. Available at: *https://www.wmo.int/pages/prog/wcp/wcdmp/hq-gdmfc/documents/QC_QAguidelines-April2019.pdf*.

WMO. 2019c. Manual on the high-quality global data management framework for climate. WMO-No. 1238. Geneva, Switzerland: World Meteorological Organization. 43 pp. Available at: *https://library.wmo.int/doc_num.php?explnum_id=10197*.

WMO. 2019d. WMO data policy statement. Draft 1.0. *Study Group on Data Issues and Policies. WMO Data Conference*. 16–19 November 2020, Virtual. Available at: *https://meetings.wmo.int/WMO-Data-Conference/Documents/Flyer%20for%20Res.42%2011%2015.pdf*.

W3C (World Wide Web Consortium). 2020. Data Catalog Vocabulary (DCAT), Version 2. Available at: *https://www.w3.org/TR/vocab-dcat-2/#Class:Dataset*.

TO CITE THIS ARTICLE:
Peng, G, Downs, RR, Lacagnina, C, Ramapriyan, H, Ivánová, I, Moroni, D, Wei, Y, Larnicol, G, Wyborn, L, Goldberg, M, Schulz, J, Bastrakova, I, Ganske, A, Bastin, L, Khalsa, SJS, Wu, M, Shie, CL, Ritchey, N, Jones, D, Habermann, T, Lief, C, Maggio, I, Albani, M, Stall, S, Zhou, L, Drévillon, M, Champion, S, Hou, CS, Doblas-Reyes, F, Lehnert, K, Robinson, E, Bugbee, K. 2021. Call to Action for Global Access to and Harmonization of Quality Information of Individual Earth Science Datasets. *Data Science Journal*, 20: 19, pp. 1–15. DOI: *https://doi.org/10.5334/dsj-2021-019*

Submitted: 14 December 2020
Accepted: 18 April 2021
Published: 04 May 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.