TYPE Perspective
PUBLISHED 25 August 2022
DOI 10.3389/fclim.2022.958533

Check for updates

OPEN ACCESS

EDITED BY
Tiffany C. Vance,
U.S. Integrated Ocean Observing System, United States

REVIEWED BY
Ted Habermann,
Metadata Game Changers,
United States

*CORRESPONDENCE
Rosalie R. Rossi
rosalie.rossi@tamucc.edu

SPECIALTY SECTION
This article was submitted to Climate Services, a section of the journal Frontiers in Climate

RECEIVED 31 May 2022
ACCEPTED 29 July 2022
PUBLISHED 25 August 2022

CITATION
Rossi RR, LeBel DA and Gibeaut J (2022) Growing pains of a data repository: GRIIDC's evolution from environmental disaster rapid response to promoting FAIR data. *Front. Clim.* 4:958533.
doi: 10.3389/fclim.2022.958533

Growing pains of a data repository: GRIIDC's evolution from environmental disaster rapid response to promoting FAIR data

Rosalie R. Rossi*, Deborah A. LeBel and James Gibeaut

Harte Research Institute for Gulf of Mexico Studies, Texas A&M University — Corpus Christi, Corpus Christi, TX, United States

GRIIDC is a multidisciplinary data repository created in the aftermath of the Deepwater Horizon oil spill. Development of the repository occurred even as researchers collected post-spill data, and as a result, the data management system initially focused on the ingestion of data and metadata. Data sharing was not as prevalent as it is currently, and many researchers were not familiar with data sharing and data organization best practices. Implementation of data management planning, submission, citation, and distribution features required many iterations and occurred while GRIIDC was assisting researchers with managing their rapid response data. From this challenging beginning, over the decade since the Deepwater Horizon oil spill, GRIIDC has improved the data management system and the training of researchers, which has enhanced the ease of submission and quality of data submitted. The GRIIDC system has also evolved to prioritize the implementation of FAIR data principles to ensure the data are findable, accessible, interoperable, and reusable. All data are issued digital object identifiers (DOIs) through DataCite and are findable via GRIIDC's data search page, DataONE, and Google Dataset Search. Each dataset has a landing page where the data and metadata can be accessed. GRIIDC is continuously striving to add FAIR principles to the system. Although there are still many challenges including quality of data and metadata received, funding limitations, and program priorities, GRIIDC must always continue to improve its ability to meet user needs while implementing FAIR data principles.

KEYWORDS

data sharing, data management plan (DMP), FAIR data, multidisciplinary data repository, data citation, data discoverability

Introduction

The Deepwater Horizon (DWH) offshore drilling rig operated by BP, located 50 miles off the coast of Louisiana, experienced a blowout on 20 April 2010 resulting in an explosion that killed 11 workers, released an estimated 4.9 million barrels of oil (McNutt et al., 2011), and sank the rig. Approximately 2.1 million gallons of dispersant were released both at the surface and wellhead, the first time a dispersant was applied to the water column (Kujawinski et al., 2011). A disaster this large mitigated with new methods required immediate research to study the potential effects of oil and dispersant on the environment. Although previous oceanographic research had been performed in the Gulf of Mexico, the information collected proved insufficient for this spill (Shepherd et al., 2016). Data for determining effects of oil on species (Bjorndal et al., 2011) and assessing the effects of the deep-water application of dispersants were lacking (Kujawinski et al., 2011).

On 24 May 2010, while the well was still releasing oil, BP committed $500 million dollars over a 10-year period "to fund an independent research program designed to study the impact of the oil spill and its associated response on the environment and public health in the Gulf of Mexico." This program, the Gulf of Mexico Research Initiative (GoMRI), would be independent of BP's control and administered by the Gulf of Mexico Alliance (GOMA). A Master Research Agreement (MRA) between GOMA and BP stated that GoMRI-funded data should be submitted to a "Research Database" and "that all data shall be fully accessible and posted thereto with minimum time delay." The research database formed was the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC). GRIIDC would be based out of the Harte Research Institute for Gulf of Mexico Studies (HRI) at Texas A&M University—Corpus Christi as HRI's vision and mission to support a sustainable Gulf of Mexico aligned nicely with that of GoMRI.

Developing a data repository in parallel with initial data collection presented several challenges. Time was a critical issue as a team of software developers was building the system while other GRIIDC personnel were working with researchers to help them organize and submit their data. Another barrier was that in 2010, data sharing and data management best practices were only just being developed. Some researchers were not familiar with or resisted data sharing. Other researchers did not identify their work as data, applying a traditional model of a physical sample collected in the field and analyzed in the laboratory. Still others valued only a publication as a product with impact, not recognizing the benefits of data sharing to the researcher and the general scientific community, including higher citation rates (Piwowar et al., 2007). A final challenge was the breadth of the research being undertaken in the aftermath of the DWH disaster. This included data collection in environmental, ecological, and sociological/public health sectors.

GRIIDC did have the benefit of an advisory committee which included members of its future research board and a number of principal investigators from the GoMRI research consortia. During initial GRIIDC planning meetings in 2011, data management topics discussed included data management plans, metadata standards, digital object identifiers (DOIs), data citations, data types to accept, levels of processed data to store, and best practices. The majority of these are features of a good data management plan. It is obvious when reviewing meeting notes that GoMRI and GRIIDC had already made a clear commitment to adopting best data and metadata practices as set by funding agencies such as National Science Foundation and National Oceanic and Atmospheric Administration, including interoperability, persistent DOIs, and promoting a different, open culture for data sharing.

In 2016, FAIR data principles were published, codifying principles which are finable, accessible, interoperable, and reusable (Wilkinson et al., 2016). GRIIDC had already established several FAIR data principles, including data management planning and issuing DOIs, and continues to learn and apply those principles in software development and data curation practices. In the 11 years since the formation of GRIIDC, the data management system has evolved to mitigate submission barriers for researchers and grow with the data sharing movement as best practices advanced.

GRIIDC has developed easy-to-use and intuitive submission and search interfaces, created useful management tools, crafted curation standards, and trained researchers, resulting in the submission of more useful and well-documented data that meets funding deadlines and adheres to FAIR data principles. The following sections present the principles GRIIDC initially identified as critical: data management, data and metadata submission, citation, and distribution.

Data management planning

A data management plan (DMP) template was one of the first items prioritized as GRIIDC needed to collect information about the data to be ingested to help determine repository development needs (see Figure 1 for a timeline of events). A DMP is a document that describes what data will be collected or generated and how those data will be organized, stored, documented, and backed up throughout the entirety of the research project. GoMRI research consortia were required to complete the DMP template and submit to GRIIDC via email at the beginning of a funding cycle to plan for data submission. At the beginning of the program, researchers were not familiar with DMPs or the concept of sharing data and needed guidance to develop these documents. GRIIDC reviewed all GoMRI proposals to help determine what data were to be collected and worked with researchers to develop and understand the importance of DMPs. GRIIDC has updated the DMP template through the years, adding more fields to account for the wide variety of data types GRIIDC receives (Figure 2). More specific details are obtained for each data type such as research cruise, field work, environmental lab analysis, microcosms/mesocosms, modeling, mapping, social surveys, images, and video. Researchers can utilize these resources for any project as many funding organizations now require DMPs when submitting proposals.

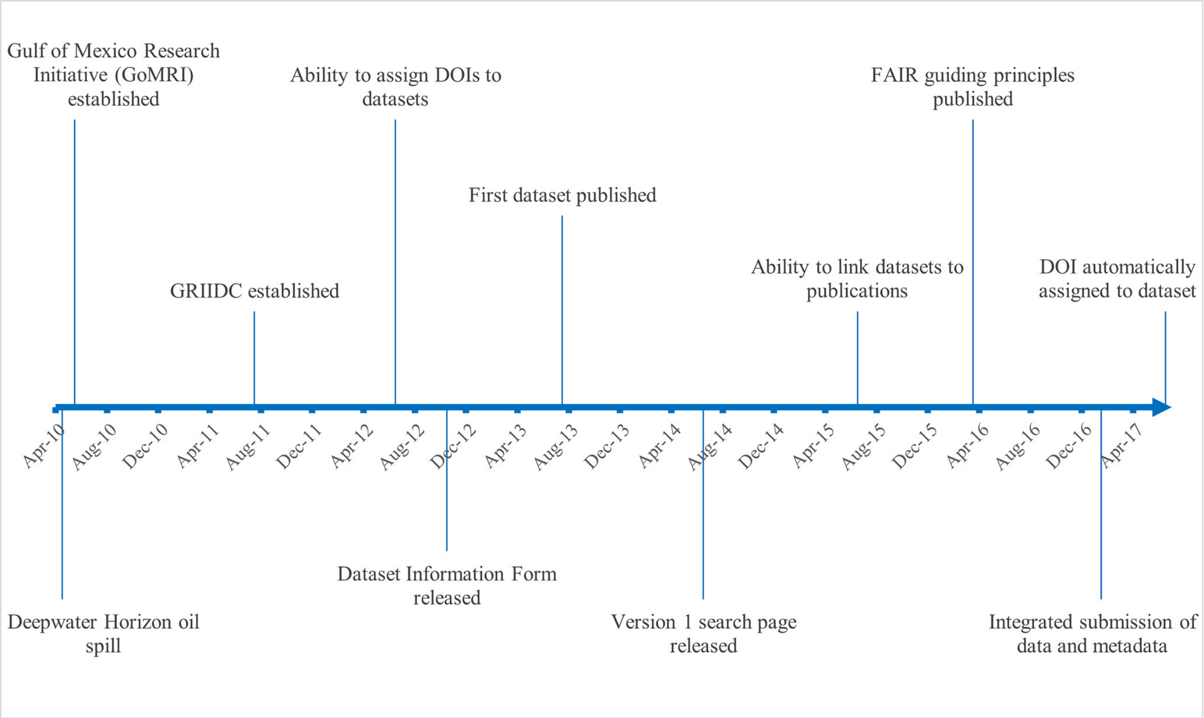
Gulf of Mexico Research Initiative (GoMRI) established
Ability to assign DOIs to datasets
FAIR guiding principles published
First dataset published
GRIIDC established
Ability to link datasets to publications
DOI automatically assigned to dataset
Apr-10
Aug-10
Dec-10
Apr-11
Aug-11
Dec-11
Apr-12
Aug-12
Dec-12
Apr-13
Aug-13
Dec-13
Apr-14
Aug-14
Dec-14
Apr-15
Aug-15
Dec-15
Apr-16
Aug-16
Dec-16
Apr-17
Dataset Information Form released
Deepwater Horizon oil spill
Version 1 search page released
Integrated submission of data and metadata

FIGURE 1
GRIIDC timeline of events.

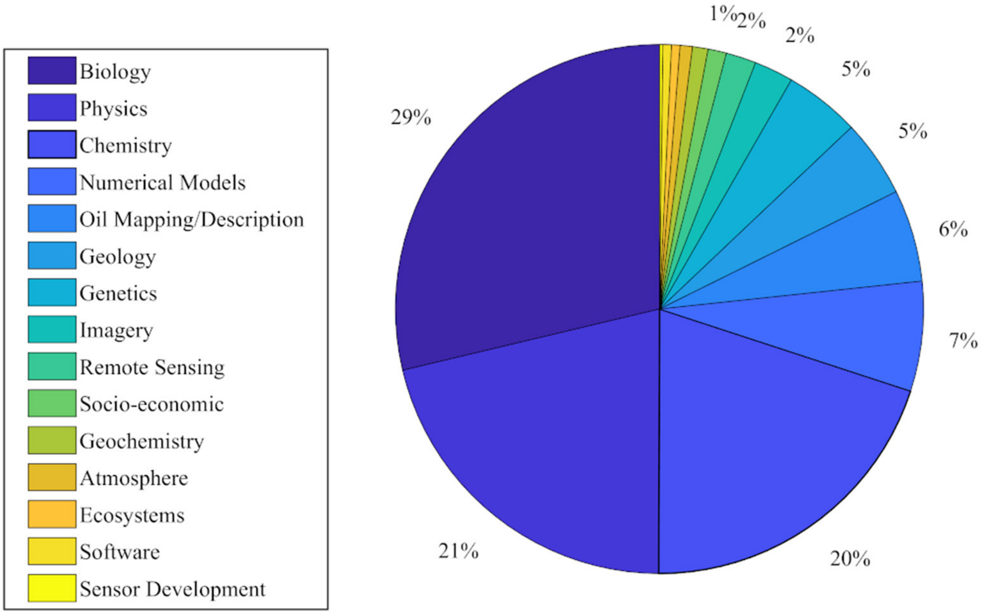

FIGURE 2
Distribution by discipline of the 3,086 GoMRI-funded datasets. Imagery and model datasets typically have a second classification indicating subject matter. Classes not labeled with a percentage comprise <1% of the total GoMRI-funded datasets.

An important advance that GRIIDC made in data management planning was the development of the Dataset Information Form (DIF), which initiates metadata collection for a dataset expected to be developed. Although a DMP for the project has important information on the project level, GRIIDC determined that more detail on specific datasets to be submitted was needed to initiate tracking (Gibeaut, 2016) and to organize dataset submissions. The DIF also helps GRIIDC prepare to ingest the data. The DIF is implemented through an online tool that GRIIDC developed, and it is integrated into the data submission workflow on the GRIIDC website. The DIF collects basic metadata such as title, abstract, data parameters and units, size of dataset, estimated data sampling period, and spatial extent. It also provides the opportunity for a researcher to indicate if the data are already located at a national data archive or if they are governed under the Institutional Review Board (IRB) or Health Insurance Portability and Accountability Act (HIPAA). When researchers are ready to submit data, the submission form is pre-filled with information provided in the DIF, thereby reducing work. GRIIDC's dataset monitoring page displays the status of a dataset through the data management workflow allowing submitters, managers, journals, and funding organizations to monitor its status. Requiring data management planning prepares a researcher for the data management lifecycle and provides a document to describe how data will be FAIR.

Submitting data and metadata

Gathering information about GoMRI-funded projects and data that were collected before GRIIDC was well established was difficult as most researchers had never prepared to share data before. GRIIDC recognized that the data submission process would need to be straightforward to accommodate researchers' various levels of technical experience, time, and patience. However, with data already being collected, a submission interface would need to be developed quickly. The first interface included a "registration" page where users could upload data and metadata. GRIIDC developed a metadata editor with which users created ISO 19115-2 metadata xml files. Users had to save the file locally and then submit the xml file to the GRIIDC system. GRIIDC encountered issues with this process as researchers would submit the data but not the metadata, causing delays in the review of the dataset or prohibiting publishing an incomplete dataset. Additionally, the submission interface could only accept a single file, requiring users to create an archive for multi-file datasets. GRIIDC would have to mitigate issues with corrupt archives and files that could not be opened.

Following user feedback and software development improvements, GRIIDC has developed an easy-to-use dataset submission form that integrates metadata and data submission into one interface (Figure 1). The form is pre-filled with information previously collected in the DIF. Users simply enter metadata such as abstract, keywords, data parameters and units, methods, spatial extent, and other descriptive information. An ISO 19115-2 compliant metadata file is automatically generated from this information and also includes other attributes such as suggested citation, data usage license, and distribution information. GRIIDC has added these fields to ensure data are findable, interoperable, and reusable. GRIIDC provides metadata in a human-readable format along with the ISO-19115-2 xml version, allowing access to users with different levels of technicality (Gries et al., 2018). Once the metadata is provided, a user can submit the data by direct upload. If data are large (over 25 gigabytes), the researcher may transfer the data via SFTP, GridFTP, Globus, or an external hard drive. If data are already located at a national data archive, a user can provide the DOI URL for the data at that location. Providing multiple methods for data submission allows researchers to choose the best option for upload given the size of their data, connection quality, location of data, and technical experience.

Due to GRIIDC's unique beginning in which researchers were studying various effects of the Deepwater Horizon oil spill, a wide range of data types were submitted to the repository including biology, chemistry, physical oceanography, sociology, political science, and public health (Figure 2). The varied documentation and metadata presented another challenge for GRIIDC. To provide more information to researchers, GRIIDC to date has created 12 guidance documents that describe recommendations for each data type. These are constantly evolving as data standards are continuously being developed and improved. For example, in 2018, to complement the required metadata and facilitate submission of data to the National Centers for Environmental Information (NCEI), GRIIDC requested researchers submitting data acquired on research vessels complete a cruise data documentation template. This template provides supplemental information, including cruise platform, dates, chief scientist, and cruise designation. This allows identification of related data housed at other data repositories such as Rolling Deck to Repository (R2R) and NCEI and assists in obtaining additional documentation such as cruise reports.

Data citation

GRIIDC determined at the beginning of the program that assigning DOIs was a vital component of the data submission process to make sure data were findable and reusable (Figure 1). The University of California's California Digital Library EZID service was initially used to create DOIs for GRIIDC datasets. GRIIDC developed a DOI request form that users would submit as a separate process from data submission. The DOI at EZID would automatically have an "unavailable" status, meaning that the DOI would resolve to a tombstone page with the citation's metadata and reason for not being available. GRIIDC personnel would review the request; once the dataset had passed the data package review process, the DOI would be changed to "public" and would resolve to a dataset landing page. The researcher would then have to return to the registration page and enter the DOI to include it as part of the dataset. This process required multiple steps from the user and GRIIDC personnel. Additionally, it did not ensure that all datasets were assigned DOIs as it relied on the user to request one. In 2017, GRIIDC integrated DOI assignment with the dataset submission process and switched to DataCite for DOI minting services. Upon submission of a dataset, a DOI is assigned which automatically displays on the dataset landing page where the data can be downloaded, as well as a map displaying the spatial extent (if applicable), author information, a suggested citation, number of files, file size, file format, and the collected metadata. The DOI will not resolve to the landing page if the dataset has not completed the data package review process or if there is an embargo on the dataset. Automating this process has ensured that each GRIIDC dataset is assigned a DOI and eliminates additional steps for the user and GRIIDC personnel.

Displaying a DOI on a dataset landing page upon data submission facilitates the user providing the DOI to journals that require data be made publicly available. The dataset landing page contains a suggested citation, which makes it convenient for users of the data to properly cite the resource. Citation provides credit to the researcher, helps in data access and findability, and can track impact (Ball and Duke, 2015). Also found on the dataset landing page is a link to associated publications. GRIIDC has linked 1,358 publications to GRIIDC datasets. Pairing the linking of dataset to publication and referencing the dataset DOI within its associated publication maximizes the findability and impact of the data.

Distributing data

Data can be found and downloaded using GRIIDC's search page. In keeping with the rapid response nature of GRIIDC's origin, the search functionality was originally quite minimal, returning a simple listing of datasets. Improvements were made with new software releases. Users can now enter advanced search terms and narrow down to specific fields such as dataset title, abstract, author, or theme keywords. Facets can be used to further filter results by dataset status, funding organizations, and research groups. Data may be downloaded by anyone with no requirement of a GRIIDC account. Improvements to the user interface in 2021 allow a dataset to be downloaded in its entirety as a zip file or as individual files. Upon download, a SHA256 checksum hash is calculated for compressed files to confirm transfer integrity.

Reflecting GRIIDC's commitment to FAIR data principles and long-term data archival, GRIIDC data is also available from additional sources. Increased discoverability of data is provided by participation in the Data Observation Network for Earth (DataONE) where metadata of GRIIDC datasets can be found. GRIIDC also submits GoMRI-funded oceanographic data to NCEI for long-term archival. The use of standardized National Oceanographic Data Center (NODC) vocabulary terms or the National Aeronautics and Space Administration's (NASA) Global Change Master Directory (GCMD) vocabulary terms for data types and instruments enhances data discovery.

GRIIDC is currently improving an Environmental Research Division Data Access Program (ERDDAP) server, initially developed in 2015, to further serve its oceanographic data (hydrographic data, current measurements, underway sensor measurements, and drifter/float trajectories). An ERDDAP server provides additional search functionality and online map and graph creation. It also provides the ability to download data in a single format of the user's choice, adding flexibility and reducing the extraction/translation/load (ETL) burden.

Discussion

GRIIDC has a unique origin story as a data repository. Due to the urgency of its initial development and the rapidly evolving climate of data sharing, GRIIDC has faced challenges since its inception. As GRIIDC was at the forefront of the data sharing movement (Gibeaut, 2016), data standards were still being developed and researchers' knowledge of what constitutes data, data organization, and data sharing data was limited. However, involving an advisory committee during developmental stages of the program helped to address these challenges and develop data management best practices that would set the program up for success well into the future. The data sharing culture has vastly changed since the origination of the GoMRI program. Many funding agencies and journals now require that data be shared, and researchers are accepting the numerous benefits of sharing data: open data can be used to discover errors, create new questions, or be combined with other data (McNutt et al., 2016). GRIIDC has prepared researchers for success in this data sharing culture as they have been trained in data organization and management and are now familiar with submitting data and creating descriptive metadata.

GRIIDC is always striving to support FAIR data practices and contribute to the ever-growing collection of open data. GRIIDC now hosts data not only from GoMRI but also from the Florida RESTORE Act Centers of Excellence Program; the Mississippi Based Center of Excellence; the Harte Research Institute; the National Academies of Sciences, Engineering, and Medicine Gulf Research Program; as well as others.

While its inception was based on an environmental disaster, GRIIDC has come a long way, developing a data repository that strives to follow the FAIR data principles and will continue to ensure a data and information legacy for the Gulf of Mexico.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

RR was the primary author of the manuscript. DL wrote sections of the manuscript, provided edits, and created figures. JG provided edits and feedback. All authors contributed to the article and approved the submitted version.

Funding

This work was funded by a grant from the Gulf of Mexico Research Initiative.

Acknowledgments

We would like to acknowledge the Gulf of Mexico Research Initiative (GoMRI), the GoMRI Research Board Data Committee, researchers, and research consortia data managers for their assistance and feedback over the years that contributed to the development and improvement of GRIIDC. We would also like to acknowledge the Harte Research Institute for Gulf of Mexico Studies at Texas A&M University—Corpus Christi for their support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Ball, A., and Duke, M. (2015). *How to cite datasets and link to publications. DCC How-to Guides. Edinburgh: Digital Curation Centre. [online]*. Available online at: https://www.dcc.ac.uk/guidance/how-guides/cite-datasets (accessed March 25, 2022).

Bjorndal, K. A., Bowen, B. W., Chaloupka, M., Crowder, L. B., Heppell, S. S., Jones, C. M., et al. (2011). Better science needed for restoration in the Gulf of Mexico. *Science*. 331, 537–538. doi: 10.1126/science.1199935

Gibeaut, J. (2016). Enabling data sharing through the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC). *Oceanography*. 29, 33–37. doi: 10.5670/oceanog.2016.59

Gries, C., Budden, A., Laney, C., O'Brien, M., Servilla, M., Sheldon, W., et al. (2018). Facilitating and improving environmental research data repository interoperability. *Data Science J*. 17, 22. doi: 10.5334/dsj-2018-022

Kujawinski, E. B., Kido Soule, M. C., Valentine, D. L., Boysen, A. K., Longnecker, K., and Redmond, M. C. (2011). Fate of dispersants associated with the Deepwater Horizon oil spill. *Environ. Sci. Technol*. 45, 1298–1306. doi: 10.1021/es103838p

McNutt, M., Camilli, R., Guthrie, G., Hsieh, P., Labson, V., Lehr, B., et al. (2011). "Assessment of flow rate estimates for the Deepwater Horizon / Macondo well oil spill," in *Flow Rate Technical Group report to the National Incident Command, Interagency Solutions Group*, March 10, 2011.

McNutt, M., Lehnert, K., Hanson, B., Nosek, B. A., Ellison, A. M., and King, J. L. (2016). Liberating field science samples and data. *Science*. 351, 1024–1026. doi: 10.1126/science.aad7048

Piwowar, H. A., Day, R. S., and Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE* 2, e308. doi: 10.1371/journal.pone.0000308

Shepherd, J., Benoit, D. S., Halanych, K. M., Carron, M., Shaw, R., and Wilson, C. (2016). Introduction to the special issue: an overview of the Gulf of Mexico Research Initiative. *Oceanography*. 29, 26–32. doi: 10.5670/oceanog.2016.58

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*. 3, 160018. doi: 10.1038/sdata.2016.18 Erratum in: (2019) *Sci. Data*. 6, 6.