THE OFFICIAL MAGAZINE OF THE OCEANOGRAPHY SOCIETY

Oceanography

Enabling Data Sharing Through the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC)

By James Gibeaut

INTRODUCTION

Our ability to improve society's understanding of the Gulf of Mexico ecosystem, including humans, and to ensure the Gulf's long-term environmental and public health requires access to a wide array of data. Understanding the impacts of petroleum pollution and related stressors on marine and coastal ecosystems and human populations calls for the ability to integrate and analyze data from diverse sources, across disciplines, and from varied spatial and temporal scales. One of the more frequent observations hindering determination of Deepwater Horizon spill ecosystem impacts revolved around the lack of baseline data from many disciplines. Data are required to make informed decisions about the management of complex systems, particularly relating to impacts, future response, mitigation, and restoration following spills and natural disasters.

Changes in the ways scientists gather, manage, and analyze data are driven, in some cases, by the availability of innovative new data gathering tools and new low-cost computing capabilities. Other changes are driven by how and what data, particularly public health data, are collected and accessed. Society, however, is also demanding change (McNutt et al., 2016). The public wants increased transparency. Decision-makers from all sectors are calling for reproducibility and validation. As public and environmental health become increasingly interconnected, health professionals and policymakers require timely access to reliable and robust monitoring data that provide a baseline for informed decision-making to promote the health and well-being of ecosystems and the people who live and work in these systems. The science community is beginning to recognize and address this need for large, accessible, integrated data sets. Recently, the National Oceanic and Atmospheric Administration (NOAA) announced it will be partnering with five Web organizations—Amazon Web Services, Microsoft Azure, IBM, Google, and the Open Cloud Consortium—through a Cooperative Research and Development Agreement (CRADA) to organize and make NOAA's data more easily accessible and usable (https://www.commerce.gov/news/press-releases/2015/04/us-secretary-commerce-penny-pritzker-announces-new-collaboration-unleash).

Access to data generated by the Gulf of Mexico Research Initiative (GoMRI) can make a direct difference to understanding, responding to, and mitigating future oil spills. GoMRI recognized this early on in the program's development, and the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC; https://data.gulfresearchinitiative.org) serves as an excellent example of how data integration and consistency can provide exponential added value to the research community as a whole.

Research investigations funded by GoMRI have resulted in a large pulse of scientific data produced by studies ranging across the program's five research themes (Shepherd et al., 2016, in this issue). Data sets from laboratory, field, and modeling activities describe phenomena ranging from microscopic fluid dynamics to large-scale ocean currents, from bacteria to marine mammals, and from detailed field observations to synoptic mapping. One of GoMRI's central tenets is to ensure that all data are preserved and made publicly available, and GRIIDC ensures a data and information legacy that promotes continual scientific discovery and public awareness of the Gulf of Mexico ecosystem.

Open data requirements are increasing in number and enforcement. There are many reasons for the effective curation and sharing of data, including (1) providing environmental baselines for gauging the effects of episodic events such as storms or oil spills, (2) increasing the efficiency of the scientific process through reuse of data and providing direction for future data acquisitions, (3) increasing public trust by making data available that are used in applying and developing public policy, and (4) enabling new discoveries through data mining.

GoMRI became a leader in the move toward open scientific data in 2011 when BP and the Gulf of Mexico Alliance established in their Master Research Agreement (MRA) a research database from which all data are to be made "fully accessible" with "minimum time delay." The MRA also charges the GoMRI Research Board with developing data policies and the GoMRI Administrative Unit with administering the research database.

The Research Board established that "fully accessible" meant publicly available with documentation (metadata) to make data sets understandable and reusable. Further, the phrase "minimum time delay" was defined as within one year of data acquisition or before publications appear that use the data. This "one-year or before publication" requirement is ambitious and on the forefront of data-sharing policies of research funding organizations. It has caused the program to focus on data management throughout the data life cycle and requires a commitment of time and resources by researchers. It has also created the need for GRIIDC to develop processes and resources for data planning, tracking, and archiving as well as training for researchers. This article describes the structure of GRIIDC and the approach to meeting a stringent open data requirement.

PROMOTING A DATA SHARING CULTURE

With what would eventually become more than 3,000 researchers and more than 200 institutions involved in generating data subject to what much of the science community considered a challenging and, in some cases, an overreaching data sharing policy, GRIIDC has sought to promote a cooperative atmosphere to effectively implement the policy. To that end, several approaches are taken to maximize the success of meeting the data policy requirements:

1 Provide an efficient data management system and training. To entice researchers to share their data, it is important to make it easier for them to document and upload data. It is also critical that users have training and guidance to know what data to provide and how to upload their data. One key aspect of the GRIIDC repository is that for most data sets, no format transformations, other than transformations to non-proprietary formats, are required and that a group of files may be compiled into one zip file as a data set for submission. This gives researchers flexibility in organizing data and reduces time needed to reorganize data for submission. GRIIDC also developed a metadata editor as a Web application that generates ISO 19115-2 compliant metadata (ISO, 2009) files for GRIIDC-required core elements.

2 Store and disseminate data from the GRIIDC repository and distribute data sets to appropriate national archives on the researcher's behalf, when possible, to increase the visibility of the data sets and ensure their preservation.

3 Publish digital object identifiers (DOIs). A DOI is a unique and persistent link for resources, including data sets. DOIs have associated metadata that describe the resource. For online resources, this may include the URL to the object. If the URL of the resource changes, the DOI metadata are updated to reflect the change, and the DOI will remain the same. This allows the effective citation of data sets in publications, hence giving credit to those who develop and share data sets. All GoMRI data sets are assigned DOIs that resolve to the GRIIDC data set landing pages, which also include suggested citation formats for the data sets.

4 Highlight data sharing. Data sharing is highlighted through stories on the GRIIDC website to bring recognition to scientists and their data sets and to relay experiences to other researchers regarding data management and the sharing process.

5 Monitor data sets. GRIIDC exposes the status of data sets under development for each GoMRI research project to the public. The data set monitoring website allows visitors to explore the lists of data sets and their descriptions. Anyone can readily see by project which data sets have been slated for acquisition, registered, documented with metadata, reviewed, and made publicly available.

6 Tie data sharing to funding. Unlike many other funding organizations whose data-sharing requirement may be met after the research has ended and grant accounts have been closed, GoMRI requires proof of ongoing data sharing progress during the project period. Inadequate progress toward compliance with the data management policy could jeopardize approval of continuation or future GoMRI funding opportunities.

These approaches work together by enabling researchers to share, by rewarding sharing, and by enforcing consequences for not sharing. It is not surprising that we have seen increases in the rate of data submittal leading up to times of Review Board site visits, extension requests, and new proposal deadlines (Figure 1). It is clear that even if there is a data repository and help available to use it, there is still not enough incentive built into the environmental research community to share data (Sayogo and Pardo, 2013). The near-real-time enforcement of the data-sharing requirement has been critical for building the GoMRI research database.

GRIIDC ORGANIZATION

Partners

GRIIDC is operated from the Harte Research Institute at Texas A&M University–Corpus Christi, where staff, software, and the data repository are located. GRIIDC partners include the Gulf of Mexico Coastal Ocean Observing System (GCOOS), the Northern Gulf Institute (NGI), and the Florida Fish and Wildlife Research Institute (FFWRI). GCOOS provides expertise in data distribution and subject matter expertise for physical oceanographic data sets. NGI developed and maintains the Research Information System (RIS), a database that includes information on GoMRI personnel, institutions, funded projects, and publications, for access by the GRIIDC data management system. FFWRI sought out data sets from research institutions funded directly by BP after the Deepwater Horizon spill and prior to the formation of GoMRI and the establishment of GRIIDC.

Advisory Committee

GRIIDC is steered by an advisory committee (AC) that includes (1) the Data Management Committee of the Research Board, (2) the GoMRI Chief Science Officer, (3) a GoMRI Administrative Unit representative, (4) the GRIIDC Director, (5) designated data managers from each research consortium (RC), and (6) a representative from the NOAA Coastal Data Development Center, now part of the NOAA National Centers for Environmental Information. The inclusiveness of the AC promotes cooperation in meeting GoMRI data policies. The AC meets bimonthly via teleconferences and in semiannual in-person meetings. RC data managers are expected to contribute as members of the AC, although all GoMRI researchers are invited to participate. The AC meetings are a forum for GRIIDC staff to present new data management processes, resources, and guidance and for RC data managers to present their challenges and successes for discussion and problem solving. It is also significant that data management and policy are topics of discussion at most GoMRI science meetings and teleconferences, reinforcing the importance of data sharing and curation in the program.

Functional Areas

GRIIDC has three main functions: (1) data management, (2) communications and training, and (3) assessment. The data management function includes development and maintenance of a system that includes the adoption of data set requirements and procedures to document, catalog, and host data. This function also maintains the hardware and software elements of GRIIDC. The communications section provides training and technical support to researchers regarding data management and use of the GRIIDC Data Management System (DMS) as well as guidance on data set submissions. The communications function develops training videos and guidance documents for researchers that are made available through the website, training webinars, in-person training events, and informational booths at conferences, and individual support is provided through email and phone. Stories publicizing the data management and sharing efforts of researchers are produced for the website to provide recognition and highlight data as a research product. The assessment function serves GoMRI by tracking the state of data set development and providing information to the RB. The assessment function is also working on visualizations of the research database holdings and may eventually provide assistance to researchers conducting aggregate analyses using the database.

GRIIDC PROCESSES

Data Management Planning

A section on data management is required in research proposals submitted to GoMRI, but the creation of detailed plans begins shortly after projects receive funding. In cooperation with GRIIDC staff, research consortia write data management plans with the goal of having an approved plan within 180 days of the start of the project. GoMRI data management plans have three sections: (1) research consortium information, including the designation of a data manager; (2) detailed research task information regarding who is involved in acquiring data sets, data set characteristics, how the data will be documented and backed up, any ethical issues, and existing repositories or archive centers that could also store the data sets; and (3) data set information forms (DIFs).

DIFs are a key element in the GRIIDC data management process. They identify who is responsible for each data set and ask researchers to estimate characteristics of the expected data sets, providing what is essentially pre-acquisition metadata using an online form. Each data set that is expected to be developed should have a DIF in the system to complete the data management planning process. The DIF has proven to be an important planning tool that helps researchers consider elements of data management early in their project. Additionally, identifying data sets early helps GRIIDC plan the design of the data management system and its infrastructure, including estimating the amount and type of storage the system requires. Importantly, the

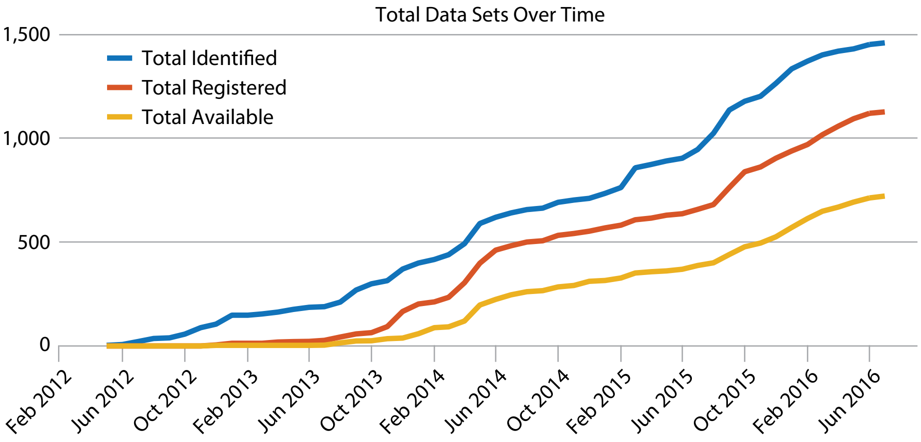

FIGURE 1. Cumulative count of data sets entering the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC) system through time. Total identified refers to data sets registered and available plus those that are planned but not yet developed. Total registered includes available data sets and those that have been registered but not yet approved for public sharing.

DIF also starts the data set tracking process, and the identified data sets and their progress toward availability is revealed on the website.

Data Set Tracking

The GRIIDC DMS allows all visitors to see data sets under development and to track progress. Data sets may be filtered by research award and project title. The monitoring table quickly shows which data sets have been identified with an approved DIF, which have been registered and submitted to GRIIDC, if metadata have been submitted and approved, and whether the data set is publicly available. Overall statistics on the numbers of data sets are also provided.

Metadata Creation

Metadata are pieces of information that describe the contents and context of data set files. Their main purposes are to help users find the data they need and then determine how to use it. Metadata are also used to support data management, archiving, and preservation. Metadata standards have been developed to allow the automated cataloging and discovery of data sets. Generally, standards define what information is to be included in metadata and how it should be structured. GRIIDC uses the International Standards Organization (ISO) 19115-2 standard. GRIIDC chose this ISO standard because of its wide acceptance and use by many data repositories, including national data centers.

> “One of GoMRI’s central tenets is to ensure that all data are preserved and made publicly available, and GRIIDC ensures a data and information legacy that promotes continual scientific discovery and public awareness of the Gulf of Mexico ecosystem.”

GRIIDC provides a metadata editor tool on its website. The metadata editor creates ISO 19115-2 compliant metadata files in Extensible Markup Language (XML) format, which is readable by both people and computers. No knowledge of XML is needed to complete metadata using the interactive forms of the editor. Embedded help tips assist the users striving to provide detailed information and develop proper metadata.

Data Ingestion

Data packages, which typically consist of one zipped file containing multiple files, and associated metadata files are submitted through the GRIIDC online submission and registration process. Users have a variety of options for uploading data set files, depending on their sizes and locations. Very large data sets maybe transferred using GridFTP or sent to GRIIDC on portable hard drives. Data set files that are available through a national data archive do not need to be transferred to GRIIDC if a stable link that takes a user directly to the data set download page exists. In this case, the URL is provided during the registration process. To submit and register a data set, the submitter identifies the DIF that refers to the data set, and the DIF information automatically populates the submission and registration form. The submitter then updates this information in the submission and registration form and provides the data set file transfer information to complete the process.

Once a data package, with accompanying metadata, is submitted to GRIIDC, a thorough review of the complete package is performed by GRIIDC staff. While this review does not include quality assurance or quality control of data points themselves, it does verify that the data set file contains data and that these data are completely and accurately described in the metadata file. The data set package is verified by a subject matter expert to ensure the contents of the data file are those that a colleague in the same field of study would expect to be included. Whenever issues arise, follow up with investigators ensures that issues are documented and resolved in a timely manner. By reviewing data packages and working closely with investigators, GRIIDC ensures that data sets are complete, discoverable, and well documented to support future use.

Data Dissemination

Data sets are made available for download using geographic or text searches on the GRIIDC website. For data sets with a geographic context, GRIIDC requires that footprints, preferably generated from actual data point locations, be provided in the metadata. These footprints may be polygons, polylines, or point features, but simplified bounding boxes are not acceptable. This greatly increases the usefulness of the geographic search filter and will help with data gap analyses.

To improve discoverability of GoMRI data sets, GRIIDC is making its metadata catalog available to other data search facilities. GRIIDC expects to be a member node in the Data Observation Network for Earth (DataONE, https://www.dataone.org) in fall of 2016. DataONE is a distributed network that links repositories to provide discoverability and access of environmental data sets across all member nodes. GRIIDC is also submitting appropriate data sets to NOAA’s National Centers for Environmental Information (NCEI). NCEI is a national data archive, and copying data sets to it will increase their discoverability and the probability that GoMRI data will be preserved.

GoMRI RESEARCH DATABASE

As of July 2016, a total of 1,450 data sets had been identified for development. Of those, over 700 were publicly available. GoMRI research will continue to 2020 and will fund another round of projects to begin in 2018. Thus, we expect the database to grow to as many as 2,000 data sets. Figure 2 shows the distribution of the number of data sets by type. This distribution shows the wide reach across disciplines in GoMRI research.

FUTURE OF GRIIDC BEYOND GoMRI

GRIIDC plans to expand beyond GoMRI science and continue to promote preservation and sharing of data from other Gulf of Mexico studies, such as those stemming from the Resources and Ecosystems Sustainability, Tourist Opportunities, and Revived Economies of the Gulf Coast States (RESTORE) Act (https://www.restorethegulf.gov). GRIIDC is a unique data program in both its breadth of data types and its full service approach. GRIIDC helps researchers with data management and the data sharing process while promoting recognition for sharing. GRIIDC can assist other funding programs by monitoring compliance with data-sharing requirements, providing researcher assistance to improve data preservation, and serving as a repository to make data more widely accessible. With the increasing emphasis on the need to preserve and effectively share data, GRIIDC will become a legacy of the GoMRI program.

REFERENCES

ISO. 2009. "ISO 19115-2:2009 Geographic Information–Metadata–Part 2: Extensions for imagery and gridded data." International Organization for Standardization, http://www.iso.org/iso/catalogue_detail.htm?csnumber=39229.

McNutt, M., K. Lehnert, B. Hanson, B.A. Nosek, A.M. Ellison, and J.L. King. 2016. Liberating field science samples and data. *Science* 351:1,024–1,026, http://dx.doi.org/10.1126/science.aad7048.

Sayogo, D.S., and T.A. Pardo. 2013. Exploring the determinants of scientific data sharing: Understanding the motivation to publish research data. *Government Information Quarterly* 30(Supplement 1):S19–S31, http://dx.doi.org/10.1016/j.giq.2012.06.011.

Shepherd, J., D.S. Benoit, K.M. Halanych, M. Carron, R. Shaw, and C. Wilson. 2016. Introduction to the special issue: An overview of the Gulf of Mexico Research Initiative. *Oceanography* 29(3):26–32, http://dx.doi.org/10.5670/oceanog.2016.58.

AUTHOR

James Gibeaut (james.gibeaut@tamucc.edu) holds the Endowed Chair for Geospatial Sciences, Harte Research Institute for Gulf of Mexico Studies, Texas A&M University–Corpus Christi, Corpus Christi, TX, USA.

ARTICLE CITATION

Gibeaut, J. 2016. Enabling data sharing through the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC). *Oceanography* 29(3):33–37, http://dx.doi.org/10.5670/oceanog.2016.59.

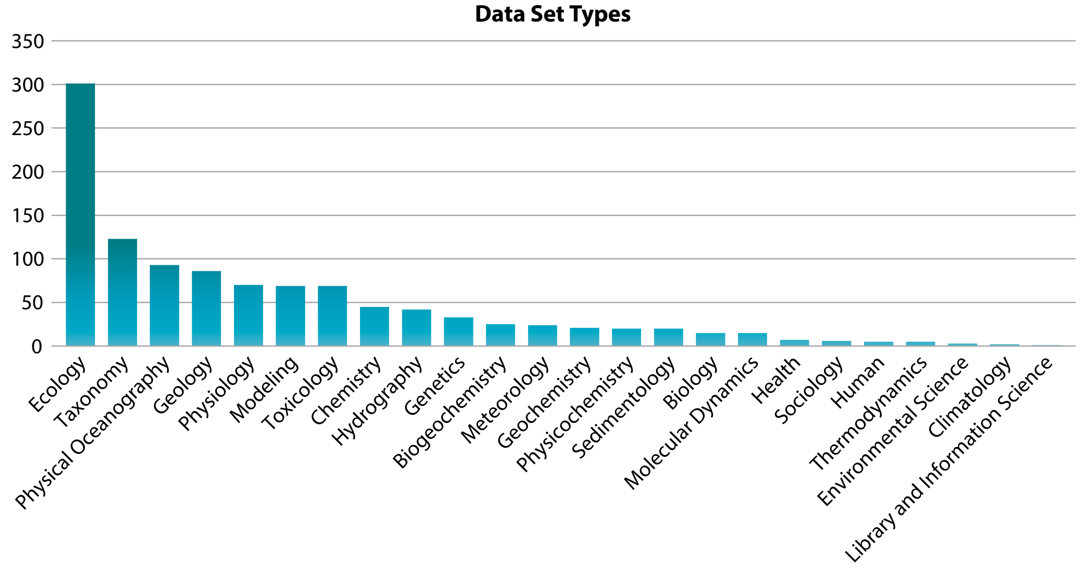

FIGURE 2. Numbers of data sets by type as of July 2016, as defined by metadata theme keywords. Some data sets are counted more than once because they span multiple categories.