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

J**ames Gibeaut and Sandra Ellis (Harte Research Institute, Texas A&M University – Corpus Christi)**

1. **Introduction**

Research investigations funded by the Gulf of Mexico Research Initiative (GoMRI) have resulted in a large pulse of scientific data produced by studies ranging across the five research themes of the program (**Shepherd et al,** this volume). These studies have produced datasets from laboratory, field, and modeling activities describing phenomenon ranging from microscopic fluid dynamics to large-scale ocean currents, bacteria to marine mammal behavior at scales from detailed field observations to synoptic mapping. One of GoMRI’s central tenants is to ensure that all data are preserved and made publicly available. Thus GoMRI formed the Gulf of Mexico Research Initiative Data and Information Cooperative (GRIIDC, <https://data.gulfresearchinitiative.org/> ) with the mission to ensure a data and information legacy that promotes continual scientific discovery and public awareness of the Gulf of Mexico ecosystem.

The preservation and open access of scientific research data have recently received much attention by government, scientific journals, and funding organizations (McNutt et al. 2016). Open data requirements are increasing in number and enforcement. There are many noncontroversial reasons for the effective curation and sharing of data including (1) providing environmental baselines for gauging the effect 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 when used in applying and developing public policy, and (4) enabling new discoveries through data mining, to name a few.

GoMRI became a leader in open scientific data in 2011 when BP and the Gulf of Mexico Alliance developed the Master Research Agreement (MRA) governing the GoMRI program. Included in the agreement was the establishment of a Research Database where all GoMRI data were to be made “fully accessible” with “minimum time delay.” The MRA also charges the GoMRI Research Board (RB) with developing data policies and the GoMRI Administrative Unit with administering the Research Database.

The RB established that “fully accessible” meant publicly available with documentation (metadata) to make the datasets understandable and reusable. Further, the phrase “minimum time delay” was defined as within one year of data acquisition or before publication using the data. This “one-year or before publication” requirement is on the forefront of data sharing policies among research funding organizations. It has brought a focus to data management throughout the data life cycle and 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.

1. **Promoting a Data Sharing Culture**

With more than 3,000 researchers and over 200 institutions generating data subject to what the science community considered a challenging and, in some minds, over reaching data sharing policy, GRIIDC promotes a cooperative atmosphere to achieve the policy. To that effect, there are several approaches taken to maximize success meeting the data policy:

1. Providing an efficient data management and training system – To encourage researchers to share their data, it is important to minimize technical difficulties in documenting and uploading data. It is critical that users are provided training and guidance on what data to provide and how to upload that data. One key benefit of the GRIIDC repository is that for most datasets, no format transformations, other than transformations to non-proprietary formats, are required. In many cases, a group of files may be compiled into one zip file as a dataset for submission. This allows flexibility in organizing data and reduces the time required to reorganize data for submission. GRIIDC also developed a metadata editor web application that generates ISO 19115-2 compliant metadata (ISO 2009) files for GRIIDC-required core elements.
2. Storing and disseminating data – The GRIIDC repository provides data storage and dissemination for the authors, as well as distribution of datasets to appropriate National Archives. This support increases the visibility of datasets and helps ensure preservation.
3. Publishing Digital Object Identifiers (DOI) – A DOI is a unique and persistent link for resources including datasets. DOIs include metadata describing the resource and its location. DOI metadata can be updated to reflect changes, while maintaining the original identifier, facilitating the citation of datasets in publications, hence giving credit to those who develop and share datasets. All GoMRI datasets are assigned DOIs and GRIIDC provides suggested citation formats for the dataset.
4. Highlighting data sharing – Data sharing stories are highlighted on the GRIIDC website to bring recognition to scientists and their datasets and to share data management experiences and the open data process with other researchers .
5. Dataset monitoring – GRIIDC shares the status of GoMRI datasets underdevelopment with the public. The dataset monitoring website page allows visitors to see by project which datasets have been slated for acquisition and are registered, documented with metadata, reviewed, and publicly available.
6. Sharing tied to funding – GoMRI requires evidence of data sharing progress during the project period unlike other funding organizations whose data sharing requirement may be met after the research has ended and grant accounts have been closed. Inadequate data management policy compliance could jeopardize approval of continuation or future GoMRI funding opportunities.

These approaches combine to enable and reward researcher sharing, while enforcing consequences for not sharing. and it is not surprising to see increases in data submittals leading up to times of RB 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).



Figure 1: Cumulative count of datasets entering the GRIIDC system through time. Total identified refers to datasets registered and available plus those that are planned to be developed. Total registered includes available datasets and those that have been registered but not yet approved for public sharing.

1. **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 Florida Fish and Wildlife Research Institute (FFWRI), and the Northern Gulf Institute (NGI). GCOOS provides expertise in data distribution and subject matter expertise for physical oceanographic datasets. FFWRI helped identify and collect datasets from research institutions funded directly by BP after the Deepwater Horizon spill, prior to the formation of GoMRI and the establishment of GRIIDC.

totracks all facets of GoMRI funded projects, as well as the institutions and people funded through the research program. The RIS also maintains a bibliography of all the peer reviewed publications resulting from GoMRI funding. The system supports GoMRI program administration, project compliance monitoring, and provides information through public-facing web pages. Additionally, the relationships between people, projects, and publications in the RIS are accessed by for use in associating datasets with GoMRI projects, people, and publications

Advisory Committee

GRIIDC is steered by an Advisory Committee (AC) that includes (1) the Data Management Committee of the RB, (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 two, in-person meetings during the year. RC Data Managers are expected to contribute as members of the AC, however, all GoMRI researchers are invited to participate. The AC meetings are a forum for GRIIDC staff to present new processes, resources, and guidance for data management and the RC Data Managers to present their challenges and successes for discussion and problem solving. It is also significant that data management and policy is a topic 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 dataset 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, using the GRIIDC Data Management System (DMS), and guidance on dataset submissions. The communications function develops training videos and guidance documents for researchers through the website, training webinars, in-person training events, informational booths at conferences, and individual support 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 dataset 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.

1. **GRIIDC Processes**

Data Management Planning

A section on data management is required in research proposals 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 datasets, dataset characteristics, how the data will be documented and backed up, any ethical issues, and existing repositories or archive centers that could also store the datasets; and (3) Dataset Information Forms (DIFs).

DIFs are a key element in the GRIIDC data management process. They identify who is responsible for each dataset and ask researchers to estimate characteristics of the expected datasets providing what is essentially pre-acquisition metadata using an online form. Each dataset 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 data management planning tool that helps researchers consider elements of data management early in their project. Additionally, identifying datasets 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 DIF also starts the dataset tracking process, and the identified datasets and their progress toward availability is revealed on the website.

Dataset Tracking

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

Metadata Creation

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

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 aid the user to provide detailed information and develop proper metadata.

Data Ingestion

Data packages, typically one zipped file containing multiple files, and the associated metadata file are submitted through the GRIIDC online submission and registration process. Users have a variety of options for dataset file uploads depending on the size and location of the data. Very large datasets maybe transferred using GridFTP or sent to GRIIDC on portable hard drives. Dataset 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 dataset download page exists. In this case, the URL is provided during the registration process. To submit and register a dataset, the submitter identifies the DIF that refers to the dataset, 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 dataset file transfer information to complete the process.

Once a data package, with accompanying metadata, is submitted to GRIIDC, GRIIDC staff performs a thorough review of the complete package. While this review does not include quality assurance or quality control of data points themselves, it does verify that the dataset file contains data and that these data are completely and accurately described in the metadata file. The dataset 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 datasets are complete, discoverable, and well-documented to support future use.

Data Dissemination

Datasets are made available for download using geographic or text searches on the GRIIDC website. For datasets with a geographic context, GRIIDC requires footprints, preferably generated from actual data point locations, be provided in the metadata. These footprints maybe 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 datasets, 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.org) in fall of 2016. DataONE is a distributed network that links repositories to provide discoverability and access of environmental datasets across all member nodes. GRIIDC is also submitting appropriate datasets to NOAA’s National Centers for Environmental Information (NCEI). NCEI is a national data archive and copying datasets to it will increase the probability that GoMRI data will be preserved as well as increasing discoverability.

1. **GoMRI Research Database**

As of July, 2016, a total of 1,450 datasets 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 datasets. Figure 3 shows the distribution of the number of datasets by type. This distribution shows the wide reach across disciplines in GoMRI research.



Figure 3: Numbers of datasets by type as of July, 2016, as defined by metadata theme keywords. Some datasets are counted more than once because they span multiple categories.

1. **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 RESTORE Act. 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.

1. **References**

ISO (2009). ISO 19115-2:2009 Geographic Information – Metadata – Part 2: Extensions for imagery and gridded data, International Organization for Standardization.

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(6277) 1024 – 1026. DOI:10.1126/science.aad7048.

Sayogo, D. S. and Pardo, T. A. (2013). “Exploring the determinants of scientific data sharing: understanding the motivation to publish research data”. Government Information Quarterly **30**(Supplement 1) S19-S31. DOI:10.1016/j.giq.2012.06.011.