Computers & Geosciences 87 (2016) 84–90

Contents lists available at ScienceDirect

Computers & Geosciences

journal homepage: www.elsevier.com/locate/cageo

Research paper

Lessons learned while building the *Deepwater Horizon* Database: Toward improved data sharing in coastal science

Anne E. Thessen [a,b,c,*], Sean McGinnis [b], Elizabeth W. North [b]

[a] The Data Detektiv, 1412 Stearns Hill Road, Waltham, MA 02451, USA
[b] University of Maryland Center for Environmental Science, Horn Point Laboratory, P.O. Box 775, Cambridge, MD 21613, USA
[c] The Ronin Institute for Independent Scholarship, Montclair, NJ, USA

ARTICLE INFO

Article history:
Received 19 June 2015
Received in revised form
26 October 2015
Accepted 2 December 2015
Available online 3 December 2015

Keywords:
Gulf of Mexico
Deepwater Horizon
Database
Data sharing
Oil spill
Hydrocarbon

ABSTRACT

Process studies and coupled-model validation efforts in geosciences often require integration of multiple data types across time and space. For example, improved prediction of hydrocarbon fate and transport is an important societal need which fundamentally relies upon synthesis of oceanography and hydrocarbon chemistry. Yet, there are no publically accessible databases which integrate these diverse data types in a georeferenced format, nor are there guidelines for developing such a database. The objective of this research was to analyze the process of building one such database to provide baseline information on data sources and data sharing and to document the challenges and solutions that arose during this major undertaking. The resulting *Deepwater Horizon* Database was approximately 2.4 GB in size and contained over 8 million georeferenced data points collected from industry, government databases, volunteer networks, and individual researchers. The major technical challenges that were overcome were reconciliation of terms, units, and quality flags which were necessary to effectively integrate the disparate data sets. Assembling this database required the development of relationships with individual researchers and data managers which often involved extensive e-mail contacts. The average number of emails exchanged per data set was 7.8. Of the 95 relevant data sets that were discovered, 38 (40%) were obtained, either in whole or in part. Over one third (36%) of the requests for data went unanswered. The majority of responses were received after the first request (64%) and within the first week of the first request (67%). Although fewer than half of the potentially relevant datasets were incorporated into the database, the level of sharing (40%) was high compared to some other disciplines where sharing can be as low as 10%. Our suggestions for building integrated databases include budgeting significant time for e-mail exchanges, being cognizant of the cost versus benefits of pursuing reticent data providers, and building trust through clear, respectful communication and with flexible and appropriate attributions.

1. Introduction

Geoscience is highly interdisciplinary and often requires integration of heterogeneous data sets for understanding large-scale processes in earth systems (Parsons et al., 2011). Because researchers tend to collect data within their area of specialization, an interdisciplinary study may require researchers to use data they have not collected and share their data with colleagues outside their immediate discipline. Sharing data across disciplines can be very difficult and in many cases acts as a barrier to collaboration (Beers and Bots, 2009, Edwards et al., 2011). The reasons for this are sociological and technical and have been discussed in detail elsewhere (Cragin et al., 2010, Edwards et al., 2011, Tenopir et al., 2011, Enke et al., 2012). Briefly, the absence of standards leads to the generation of highly heterogeneous data and metadata that require large investments of time and money to discover, access, and integrate properly. Because of this heterogeneity and the different cultures of research communities, a complete suite of tools and services to help researchers leverage modern computing in data reuse tasks does not yet exist. As a result, the integration of data within an interdisciplinary data set is often minimal.

The vast majority of data gathered by scientists are not discoverable or accessible through a repository (Heidorn, 2008). A global review of ocean data centers found that centers had half the data they should for each country (Kohnke et al., 2005). Disciplines that do have ease of data discovery and access often have many scientists using a few common pieces of equipment that were funded publicly with sharing in mind (e.g., Sloan Digital Sky Survey). In geoscience, this is similar to remote sensing data sets in

* Corresponding author at: The Ronin Institute for Independent Scholarship, Montclair, NJ, USA.
E-mail addresses: annethessen@gmail.com (A.E. Thessen), mcginnis@umces.edu (S. McGinnis), enorth@umces.edu (E.W. North).

http://dx.doi.org/10.1016/j.cageo.2015.12.001
0098-3004/

oceanography (e.g., LandSat, MODIS). Overall, these larger data sets are dwarfed by the collective body of smaller, heterogeneous data sets collected by individual researchers for a specific purpose (Heidorn, 2008).

Despite the challenges, interdisciplinary data sets do exist that are well integrated (i.e., not just a collection of data files) (e.g., Nielson et al., 2014). The challenges and levels of effort associated with building these data sets are not well documented. This type of quantitative information is needed to enhance understanding of data sharing in coastal sciences in order to better budget for creation of integrated interdisciplinary databases which would advance interdisciplinary geoscience. The goal of this research is to generate such information. This research focuses on an effort to build a database of oceanographic and hydrocarbon field data collected from the Gulf of Mexico during and after the *Deepwater Horizon* spill in order to validate a model of the fate and transport of hydrocarbons in the Gulf of Mexico. While the database was built with model development in mind, the database is anticipated to have general utility for the Gulf of Mexico research community. Although the heterogeneous nature of scientific data sets mandates specialized solutions, the general process of data discovery, access, and integration for this database provides valuable information about the data sharing culture within coastal and estuarine science. The objectives of this paper were to document the process of gathering and integrating data (both large databases and small data sets from individual researchers) into the *Deepwater Horizon* Database (Thessen et al., 2014a), present statistics on data sharing experienced in this study, and document technical challenges along with the solutions used to overcome them.

2. Methods

2.1. Process for data discovery and access

Hydrocarbon and oceanographic data were discovered by searching for relevant data sets and for projects that may have produced relevant data sets in online data and project repositories and funding agency award databases. A list of relevant projects was obtained from the Gulf of Mexico Sea Grant activities database. Awards made by the National Science Foundation and the Gulf of Mexico Research Initiative were searched based on appropriate keywords to find relevant projects. Online databases were searched, including the National Ocean Data Center (NODC), the Environmental Response Management Application (ERMA) Deepwater Gulf Response tool, the National Estuarine Research Reserve System Centralized Data Management Office (CDMO), the Biological and Chemical Oceanography Data Management Office (BCO–DMO), the Central Gulf of Mexico Observing System (CenGOOS), Gulf Science Data, and the Deep-C Consortium.

Once data sets were discovered they were either accessed online or through the data provider. Most of the data repositories offered downloadable data files through an ftp or searchable database with varying user interface sophistication. Some data sets contained in these repositories were embargoed and were either inaccessible or required interventions from data managers to gain appropriate permissions.

To obtain data sets not available through a repository, investigators associated with relevant awards and projects were contacted via email (Box S1). All communication was documented for each data set. As many as three (and at least two) follow up emails were sent to individuals who did not respond to the first email. After 200 days with no response, communication efforts ceased. Follow up emails included reference to funded awards, conference abstracts, and published manuscripts. The manuscripts were found with an online literature search of the investigator's name. Some projects were determined to be irrelevant to our model development. These data sets were no longer pursued nor were they included in this analysis. Data sets were received from individual investigators as .pdf, .doc, .xls, .csv, and .txt file formats. Effort was made to obtain actual data files rather than extracting data from figures; however, some data sets had to be obtained this way. Special effort was made to gather as much metadata as possible regarding accuracy, limits of detection, and methods for each data point.

Some data had to be extracted from published bar charts (2 out of 38 data sets). This was performed manually with pencil and ruler. Horizontal lines were drawn from the data point to the vertical axis. The exact value was interpolated from the intervals on the axis and an uncertainty value was assigned based on the precision of the intervals. This type of extraction was only performed if a data provider recommended extracting information from the Figure (1 data set) or if a data provider responded positively and then stopped communicating about data that was available in a published paper (1 data set). Data were also extracted from .pdf tables, but the actual numerical values were listed so no estimation was required.

2.2. Process for data attribution

An important part of gaining access to data was to ensure appropriate academic credit to data providers. Every data set and data point was assigned a unique identifier that enabled a link back to attribution information and a usage policy. The exact method of attribution and usage policy was approved by each individual data provider (except for the provider who stopped communicating about published data; in that case the publication citation was used). The data-set-level use policies were very important because no restrictions, other than reminding users to give appropriate academic credit, were placed on redistribution or publication of derived products by the overall database use policy. The attribution took one of several forms:

- If a data set was published in the peer-reviewed literature, the bibliographic citation of the paper was used as attribution.
- If a data set was published, but not in the peer-reviewed literature (such as in FigShare), the DOI, handle, or other unique identifier was cited.
- If a data set was not published in any way, a citation was generated and approved by the data provider. These citations included a URL for data access if available.
- If a data set itself was not published, but documentation was available in the form of a white paper, the bibliographic citation of the white paper was used.
- If a data provider (particularly a repository) already had citation guidelines for a particular data set, those guidelines were used.

Data sets available from existing repositories were assigned a usage policy copied directly from the repository web site. Data sets without existing usage policies were given a generic usage policy that applied to the database as a whole (Box S2). The *Deepwater Horizon* Database was constructed so that every query submitted to the database generated an attribution metadata file based on the data points that were returned in query results. This XML file included the citations and usage policies for the database as a whole and for all of the data sets returned in the query. This allowed a user to more precisely follow citation guidelines and usage policies. Fortunately, the different data-set-level use policies were not markedly contradictory.

2.3. Process for data integration

Every data point had a minimum of a 3D location (where), a date (when) and a parameter with a value (what); the integration of the data sets was based on these attributes. Data sets were very heterogeneous and were often not digital. Data sets that were digital were uploaded into the database in their original format, with original column headers and units of measure. Any modification or transformation of these data were accomplished by creating derived views that perform the necessary mathematical conversions and dictionary mapping. Then, these standardized views were incorporated into the aggregate database which users could query.

Heterogeneity of terms, units, and quality flags had to be normalized in order to effectively integrate all the data sets. Term reconciliation within the *Deepwater Horizon* Database was necessary because the many synonyms for hydrocarbon compounds made it extremely difficult to extract a unified selection of data points for a given analyte across all datasets. Term reconciliation was achieved through term mapping that listed a preferred term (for our specific project) tied to a list of synonyms (Thessen et al., 2014b). Unit reconciliation was also achieved via mapping to a preferred (for our project) unit. An additional step was required to transform the associated value based on the unit conversion. To normalize quality flags, we assigned our own definitions to individual letters and numbers and matched the original quality flags to our quality flags. Both the original value and the converted value for the terms, units, and quality flags were exposed in the final table view. This method allowed us to appropriately aggregate data without losing the terms used in the original data set. Additional information about this process and other technical details can be found in the supplementary material (Appendix S1).

3. Results

3.1. Database

At the time of this publication, the *Deepwater Horizon* Database incorporated data from 38 data sets (Table 1) and contained over 8 million data points representing *in situ* and laboratory measurements of 1442 parameters with 88 units. It was approximately 2.4 GB in size. The spatial extent of database coverage was approximately 640000 km^2 of the northern Gulf of Mexico, 4673 m altitude and 2850 m depth. Temporal coverage was 2010 to 2012. Although 74% of the data sets were from academia, the majority (92.8%) of the individual measurements came from government sources with 4.4% coming from academia, 2.6% from industry, and 0.1% from citizen science sources. It should be noted that these numbers reflect the contents of the database, not the total amount of data available. The *Deepwater Horizon* Database was opened to the general public in 2015 via web access (http://gisr.hpl.umces.edu/). Since that time, the database has been queried 153 times from 39 unique IP addresses.

Table 1
List of data sets in GISR *Deepwater Horizon* database. The 'Data Set ID' is the unique identifier assigned to each dataset before it was integrated into the database.

Data Set ID	Description	Reference
0001	Dissolved organic carbon and salinity	Zhou et al. (2013)
0010	Hydrographic data from Florida shelf	Snyder and Jeffrey (2011)
0011	PAH in Louisiana marsh sediment, salinity and pH	Silliman et al. (2012)
0013	Hydrocarbons and heavy metals in sediments	Montagna et al. (2013)
0020	Measurements related to dissolved organic matter in Barataria Bay	Bianchi et al. (2011)
0023	Measurements related to carbon chemistry	Lunden et al. (2013)
0030	Hydrocarbons and salinity in water and sediment	Whitehead et al. (2012)
0036	Oceanographic data and Polycyclic Aromatic Hydrocarbons (PAH) in oyster tissue from Lake Borgne and Mississippi sound	Soniat et al. (2011)
0039	Shape files showing surface slick, for more information see www.roffs.com	Mariano et al. (2011), Muhling et al. (2012)
0044	Northern Gulf of Mexico current data	Dzwonkowski and Park (2012)
0051	Hydrocarbons in sediments, plant tissues and tar balls	Williams et al. (2010)
0054	Tarballs and dispersant from AL coast	Clement et al. (2012)
0062	Hydrocarbons in air	Ryerson et al. (2012)
0067	Oceanographic current data	Goni et al. (2014)
0070	Hydrocarbon and aerosol measurements in air collected via aircraft	NOAA/ESRL (2013)
0073	Trace metals, nutrients, hydrocarbons and oceanographic measurements	Joung and Shiller (2013), Shiller and Joung (2012)
0076	PAH in water, sediment and tissue from the AL and MS coast	Daley (2012)
0077	PAH in sediment on barrier islands	Miller and Gornish (2014)
0085	CTD data	Roman et al. (2011)
0089	Oil range organics, PAH and water quality on MS coastline	Biber et al. (2014)
0097	Conductivity-temperature-depth (CTD) data	Daly (2010)
0101	PAH in water	Allan et al. (2012)
0102	Small alkanes in water	Valentine et al. (2010)
0104	Gaseous hydrocarbons in water	Joye et al. (2011)
0105	Volatile organic compounds in water	Reddy et al. (2012)
0112	CTD data	Patterson (2010)
0114	Nutrients and chlorophyll along northern Gulf of Mexico coastline	NERRS (2012)
0115	Hydrographic and hydrocarbon data from Tampa Bay	EPCHC (2013)
0116	Northern Gulf of Mexico current data	USM (2013)
0120	PAH in sediment	Brunner et al. (2013)
0125	PAH in sediment and plant tissue from Alabama coastal lagoon	Moody and Aronson (2011)
0129	PAH and alkanes in sediment	Overton et al. (2013)
0130	kml files showing surface slick	Walker et al. (2010)
0145	Oceanographic and hydrocarbon measurements	CSIRO (2013)
0144	Oceanographic and hydrocarbon measurements	NOAA/NOS/ORR (2013)
0146	Oceanographic and hydrocarbon measurements	Lee and Ryan (2010)
0147	Oceanographic and hydrocarbon measurements	ERMA (2014)
0148	Water chemistry data	BP (2013)

3.2. Discovered and accessed data

To build this database, we identified 146 potential sources of data, 95 of which were either determined to be unique and relevant for our model development needs or have an unknown status because a response was never received (65%). Of the 95 relevant data sets, we were able to obtain, either in whole or in part, 38 of them (40%). Some data sets were partially accessed because the entire set was not yet published. Six data sets (6%) were freely available online with enough documentation that contact with a provider was not necessary. We received a response and no data for 23 (24%) data sets and no response and no data for 34 (36%) data sets. The majority of these 95 data sets (85%) were from academic sources (university or research institution). Data sets were also from government (12%), industry (2%), and volunteer citizen science networks (1%).

The majority (62%) of requests for relevant data received a response (55 of the 89 data sets that required contact). Of these 55 responses, 39 (71%) were received after the first inquiry via email and 15 (27%) were received after the second inquiry via email. Only one response was received after a third email. The majority of responses were received within seven days of first contact (67%) and 40% were received within the first 24 h (Fig. 1). For the data sets that were received, the majority (69%) required zero to ten email exchanges (Fig. 2). The maximum number of email exchanges for one data set was 24 and the average was 7.8. Academic sources made up 85% of the total number of existing relevant data sets and comprised 74% of the final number of data sets that were shared, but this is only 35% of the total academic data sets. Government sources provided 13% of the total data sets and comprised 24% of the final number of shared data sets and 0% of the refused data sets. Two data sets were from an industry source and one was not shared. One data set came from a volunteer network and it was shared. Single factor ANOVAs were conducted to examine the effect of data set source (government, academia, industry, volunteer) on the number of emails needed, the length of time to first response, and the number of inquiries before first response. No effects were found ($p=0.963$, $p=0.725$, and $p=0.408$, respectively, $\alpha=0.05$, $n=37$). The volunteer network data set came from the Environmental Protection Commission of Hillsborough County (EPCHC). We received a positive response from EPCHC to our first request for data within 24 h of asking, which was faster than most of the responses we received. Nine emails were required to communicate with EPCHC about the data, which was more than

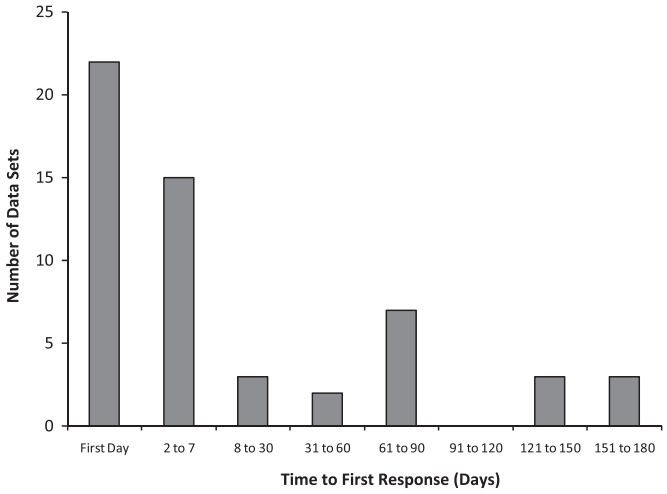

Fig. 1. Time to first response. Length of time (days) between the first request for data and the first response from the data provider for each data set. Each bar represents the number of data sets with a response time in that category.

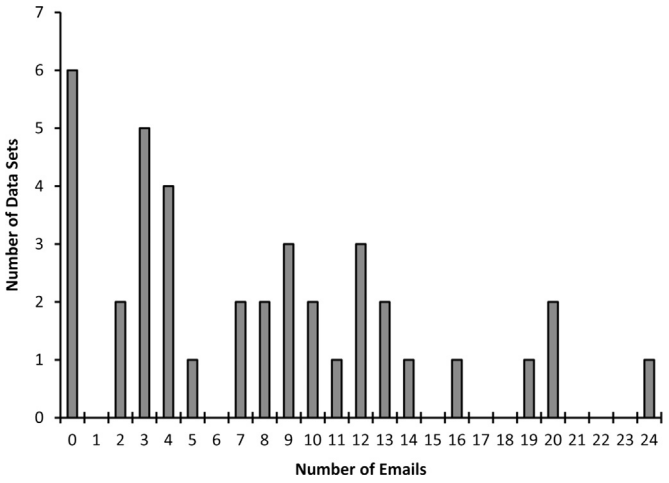

Fig. 2. Total email exchanges. Total number of email exchanges required to successfully access and integrate data and metadata for each dataset. Each bar represents the number of data sets which required that number of emails.

average. Most of the emails were questions about metadata, data use policy, and data citation. No special treatment of the data were necessary.

Of the 55 data sets for which there was a response to our request for data, 23 were not shared. The most common reasons given for not sharing were:

1) Data were not published yet (30%)
2) Keeper of the data was deceased or in poor health (9%)
3) Data/Samples were poor quality (9%)
4) Too busy (9%)

Several data sets were not obtained because contacts pointed us to another person who was unresponsive (17%). Out of the 34 data sets that received no response, 12% did not have adequate contact information. The data on which these statistics are based are available at http://dx.doi.org/10.6084/m9.figshare.

3.3. Data integration

Term and unit reconciliation was highly effective. The original data sets used 2212 different terms to describe the analytes and 122 different units. After term reconciliation, the database used 1442 terms for analytes and 88 terms for units, a 35% and a 28% decrease, respectively. The table used to normalize hydrocarbon terms can be found here: http://dx.doi.org/10.6084/m9.figshare.942516. Several examples show the effectiveness of term and unit reconciliation. Before reconciliation, a search for Isopropylbenzene and n-Decane yielded 8401 and 178 data points, respectively. After reconciliation, the same search yielded 27,312 and 40,644 data points, respectively, because the following terms were reconciled: Isopropylbenzene with i-propylbenzene and Cumene; n-Decane with Decane, nC-10 Decane, n-C10, and nC10. Observations of n-Decane were reported using seven different units: ppb, $\mu g\, g^{-1}$, $ng\, g^{-1}$, ppt, $mg\, kg^{-1}$, $\mu g^{-1}\, kg^{-1}$, and parts per trillion. All were normalized to ppb while still retaining the original unit and value in separate columns.

4. Discussion

This data discovery and integration effort has resulted in the extensive *Deepwater Horizon* Database which has strong utility for advancing basic knowledge about physical oceanography and hydrocarbons in the environment and for validating three-

dimensional oil spill prediction models. Building this database took 2 full-time months of a highly trained database programmer collaborating with a highly trained data informatics scientist who spent 6 full-time months on the project. These months of effort were spread out over three years because there was significant time waiting for data sets and requests for information.

The size and heterogeneity of the data were the largest barriers to integration and query performance. Heterogeneity was managed through several reconciliation tables and algorithms that required approximately two full-time months of manual curation. The multiple transformations of the data for normalization could not be done in real time without severely reducing performance. To address this, several views were created manually for the user to run queries against. The drawback of the curated reconciliation tables and the manually-prepared views is that any time a table was updated or a new data set added, several algorithms had to be rerun before users could see improvements. This could be partially alleviated through the use of common standards by data providers; thereby eliminating the need for reconciliation, but convincing the community to comply would be a difficult and slow process. An alternative strategy for improvement would be to find an alternate way to store provenance information, which would greatly reduce the size of the database on-disk and make the update process faster. In addition, parallel processing would have facilitated reconciliation.

The effort of assembling this database allows us to make recommendations for efficient data-gathering techniques. Building these types of databases can strain limited resources. Based on efforts described here, we recommend the following best practices for gathering data for integrated interdisciplinary data sets:

1) *Search for grants and projects that are likely to create data, not only the data sets themselves.* The vast majority of scientific data are not published or directly connected to a published study (Heidorn, 2008). Finding a project or an awarded grant that purports to produce relevant data can give information not found elsewhere and data providers can be asked specifically about the given project or award, increasing the likelihood of a response.
2) *Weigh costs of effort versus benefits when providers take longer than one week or need more than one follow-up email.* Effort spent after seven days and one follow-up email will see diminishing returns. These data sets should be high value in order to justify the resources needed to continue to pursue them.
3) *Manage email exchanges carefully and respectfully.* This can be accomplished by keeping good records of exactly what was said, by whom, and when. Make each email count by reviewing the email history and asking specific questions with specific answers. Keep emails clear, concise, and polite. Answer provider questions promptly and thank people for responding to requests.
4) *Address provider concerns about attribution and usage, which is important for engendering trust*. We recommend conferring with each data provider to develop a citation and usage policy that he or she approves. This requires extra effort, but helps establish trust and fosters a positive relationship with providers.

This research also shed light on the data sharing culture in coastal science disciplines. Based on this effort, which took place four years after samples were collected, it appears that data sharing in coastal sciences compares favorably with other disciplines; however, because availability of data declines over time, it is likely that sharing would have been lower if more time had passed since data collection (Vines et al., 2014). We received responses from 58% of contacts initiated and were able to obtain 40% of the relevant data sets that were identified. This compares with 70% responding and 10% sharing in medicine (Savage and Vickers, 2009).

Policies for data sharing dictated by funders may be the reason sharing was so high in this effort and our results may not be indicative of coastal science as a whole. Sharing data is often regulated by institutions, publishers and funding agencies that may restrict or require sharing (Field et al., 2009, Thessen and Patterson, 2011 see Table 3 therein). Many of these policies applied to the data sets in the *Deepwater Horizon* Database (Table S1). The National Science Foundation now requires data management plans, but many of the NSF-funded projects that contributed data to the *Deepwater Horizon* Database were awarded before the data management plan requirement went into effect (January 18, 2011; http://www.nsf.gov/bfa/dias/policy/dmp.jsp). Many projects were funded by British Petroleum through the Gulf of Mexico Research Initiative (GoMRI). GoMRI has a data management plan implemented by the Gulf of Mexico Research Initiative Information and Data Cooperative for making data available and defining metadata standards (http://data.gulfresearchinitiative.org/docs/DMP/Data-Management-Plan-Version-1.0.pdf). GoMRI researchers are required to be in compliance with data sharing requirements in order to receive continued funding. These requirements, originating from funding agencies that sponsored most of the work in this area, likely contributed to the relative uniformity of data set usage policies encountered during database assembly.

There are many reasons a researcher may not share his/her data and several published manuscripts have discussed the issue in detail (Costello, 2009; Cragin et al., 2010; Enke et al., 2012; Froese et al., 2003; Tenopir et al., 2011). Some of these published reasons overlap with the reasons we encountered in this endeavor. Two of the most important reasons found in other studies were a lack of time and funding (Enke et al., 2012; Tenopir et al., 2011). This reason was not explicitly stated to us, but is at the root of a "too busy" (9%) or a "see this other person" (17%) response. A lack of tools and services that enable easy data sharing has been cited in previous studies as an important reason for not sharing (Enke et al., 2012; Tenopir et al., 2011), but was not specifically mentioned to us. Good tools and services could decrease the amount of time necessary for data sharing and thus help to address the time and funding issue. Our finding that 30% of the denied requests were for unpublished data sets suggests that researchers also want time to publish results from their data analysis before sharing the data. There may be more reasons for not sharing data; our information on why data sets were not shared was derived from 40% of the data sets that were not obtained. The other 60% had no response and we do not know why they were not shared. In addition, a large body data relevant to this study are legally restricted and cannot be shared.

Two data sets were extracted from a published figure, one because a provider instructed this and the other because the provider stopped communicating after giving permission. Without obtaining permission, this practice could raise concerns because a provider might feel that their data were wrongfully obtained and because this method of extraction could introduce errors. On the other hand, extracting data from a figure is very common practice when gathering data for meta analyses and, while potentially problematic, can be the only way to "rescue" a data set (e.g., when a published figure is all that is left of a historical data set). Algorithms are being and have been developed for automated extraction of data from published photographs, 2D plots, and 3D plots with varying levels of sophistication (e.g., Lu et al., 2009; Rohatgi, 2015). These are likely to be more accurate than the traditional method of enlarging a graph and using a straight-edge to find the appropriate numbers on the axes.

A final challenge, and opportunity, with assembling a database is often the discrete end to funding for the effort while new data,

and the availability of old data, continues beyond the effort (Merali and Giles, 2005). Several relevant data sets have been identified that are not currently available, but will be in the future. We plan to repeat our database discovery effort in the future to incorporate more data into the *Deepwater Horizon* Database and to investigate how data sharing, access, and integration changes over time. This will enable us to assess how different data use policies influence sharing outcomes, which will be useful for funding agencies. In addition, we plan to survey users to determine if the existence of the *Deepwater Horizon* Database and its utility to them is a factor in their willingness to share data. This and future efforts to enhance understanding of data sharing and its incentives and impediments will provide important information to promote the creation and maintenance of synthetic databases in the geosciences. These will, in turn, offer important opportunity for scientific discovery and cost-effective knowledge generation.

Acknowledgments

The authors would like to thank the many researchers who answered our requests for, and questions about, data. This research was made possible by a Grant from BP/The Gulf of Mexico Research Initiative to the GISR Consortium (Grant no. 02-5140213 02-462273-19001). This is UMCES-HPL publication number 5126.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.cageo.2015.12.001.

References

Allan, S., Smith, B., Anderson, K., 2012. Impact of the Deepwater Horizon oil spill on bioavailable polycyclic aromatic hydrocarbons in Gulf of Mexico coastal waters. Environ. Sci. Technol. 46 (4), 2033–2039. http://dx.doi.org/10.1021/es202942q.

Beers, P.J., Bots, P.W.G., 2009. Eliciting conceptual models to support interdisciplinary research. J. Inf. Sci. 35 (3), 259–278. http://dx.doi.org/10.1177/0165551508099087.

Bianchi, T., Cook, R., Perdue, E., Kolic, P., Green, N., Zhang, Y., Smith, R.W., Kolker, A. S., Ameen, A., King, G., Ojwang, L.M., Schneider, C.L., Normand, A.E., Hetland, R., 2011. Impacts of diverted freshwater on dissolved organic matter and microbial communities in Barataria Bay, Louisiana, USA. Mar. Environ. Res. 72 (5), 248–257. http://dx.doi.org/10.1016/j.marenvres.2011.09.007.

Biber, P., Wu, W., Peterson, M., Liu, Z., Pham, L., 2014. Oil contamination in Mississippi saltmarsh habitats and the impacts to Spartina alterniflora photosynthesis. In: Alford, J., Peterson, M., Green, C. (Eds.), Impacts of Oil Spill Disasters on Marine Habitats and Fisheries in North America. CRC Press, Boca Raton, Florida, USA.

BP, British Petroleum, 2013. Gulf Science Data Water Chemistry Data File. Reference no. W-01v01-01. Last modified November 12, 2013. ⟨http://gulfsciencedata.bp.com/go/doc/6145/1942326/⟩.

Brunner, C., Yeager, K., Hatch, R., Simpson, S., Keim, J., Briggs, K., Louchouran, P., 2013. Effects of Oil from the 2010 Macondo Well Blowout on Marsh Foraminifera of Mississippi and Louisiana, USA. Environ. Sci. Technol. 47 (16), 9115–9123. http://dx.doi.org/10.1021/es401943y.

Clement, T.P., Hayworth, J., Mulabagal, V., Yin, F., 2012. Research Brief-II Impact of Hurricane Isaac on Mobilizing Deepwater Horizon Oil Spill Residues along Alabama's Coastline-a Physicochemical Characterization Study. Auburn University, Auburn, AL, USA.

CSIRO, Commonwealth Scientific and Industrial Research Organization (2013) Completed dataset from the Commonwealth Scientific and Industrial Research Organization (CSIRO) collected during the response to the Deepwater Horizon incident in the Gulf of Mexico onboard the M/V Ryan Chouest Cruise 1 through Cruise 15 from 2010-06-05 t. Available online from the US National Oceanographic Data Center at ⟨http://accession.nodc.noaa.gov/0086283⟩ (Last accessed 30.09.13).

Costello, M.J., 2009. Motivating Online Publication of Data. Bioscience 59 (5), 418–427. http://dx.doi.org/10.1525/bio.2009.59.5.9.

Cragin, M.H., Palmer, C.L., Carlson, J.R., Witt, M., 2010. Data sharing, small science and institutional repositories. Philos. Trans. R. Soc. A 368 (1926), 4023–4038. http://dx.doi.org/10.1098/rsta.2010.0165.

Daley, M., 2012. Temporal and Spatial Assessment of PAHs in Water, Sediment, and Oysters as a result of the Deepwater Horizon Oil Spill. The University of Mississippi, Oxford, MS, USA.

Daly, K.L., 2010. CTD Measurements from the Northeast Gulf of Mexico and West Florida shelf. Collected in 2010. University of South Florida. UDI is R1. x135.120:0003 Support provided by the University of South Florida Research Foundation and a FIO-Deep Water Horizon Oil Spill Grant, 2010.

Dzwonkowski, B., Park, K., 2012. Subtidal circulation on the Alabama shelf during the Deepwater Horizon oil spill. J. Geophys. Res.: Oceans 117 (C3). http://dx.doi.org/10.1029/2011JC007664.

Edwards, P.N., Mayernik, M.S., Batcheller, A.L., Bowker, G.C., Borgman, C.L., 2011. Science friction: data, metadata, and collaboration. Soc. Stud. Sci. 41 (5), 667–690. http://dx.doi.org/10.1177/0306312711413314.

Enke, N., Thessen, A.E., Bach, K., Bendix, J., Seeger, B., Gemeinholzer, B., 2012. The user's view on biodiversity data sharing—investigating facts of acceptance and requirements to realize a sustainable use of research data. Ecol. Inform. 11, 25–33. http://dx.doi.org/10.1016/j.ecoinf.2012.03.004.

ERMA, Environmental Response Management Application, 2014. Web application. Gulf of Mexico, Natural Resources Damage Assessment. National Oceanic and Atmospheric Administration. National Oceanic and Atmospheric Administration ⟨http://response.restoration.noaa.gov/erma/⟩ (Last accessed 3.06.13.

EPCHC, 2013. Surface Water Quality in Tampa Bay. Environmental Protection Commission of Hillsborough County, United States ⟨http://www.epchc.org/⟩.

Field, D., Sansone, S.A., Collis, A., Booth, T., Dukes, P., Gregurick, S.K., Kennedy, K., Kolar, P., Kolker, E., Maxon, M., Millard, S., Mugabushaka, A.M., Perrin, N., Remacle, J.E., Remington, K., Rocca-Serra, P., Taylor, C.F., Thorley, M., Tiwari, B., Wilbanks, J., 2009. Omics data sharing. Science 326 (5950), 234–236. http://dx.doi.org/10.1126/science.1180598.

Froese, R., Lloris, D., Opitz, S., 2003. Scientific data in the public domain. ACP-EU Fish. Res. Rep. 14, 267–271.

Goni G., Trinanes J., MacFadyen A., Street D., Olascoaga M., Imhoff M., Muller-Karger F. and Roffer M., Variability of the Deepwater Horizon Surface Oil Spill Extent and Its Relationship to Varying Ocean Currents and Extreme Weather Conditions, In: Ehrhardt, M. Mathematical Modelling and Numerical Simulation of Oil Pollution Problems. Springer International Publising, 2015, 1-22 http://dx.doi.org/10.1007/978-3-319-16459-5_1.

Heidorn, P.B., 2008. Shedding Light on the Dark Data in the Long Tail of Science. Libr. Trends 57 (2), 280–299.

Joung, D., Shiller, A., 2013. Trace element distributions in the water column near the deepwater horizon well blowout. Environ. Sci. Technol. 47 (5), 2161–2168. http://dx.doi.org/10.1021/es303167p.

Joye, S., MacDonald, I., Leifer, I., Asper, V., 2011. Magnitude and oxidation potential of hydrocarbon gases released from the BP oil well blowout. Nat. Geosci. 4, 160–164. http://dx.doi.org/10.1038/ngeo1067.

Kohnke, D., Costello, M., Crease, J., Folack, J., Martinez Guingla, R., Michida, Y., 2005. Review of the International Oceanographic Data and Information Exchange (IODE). Report Submitted to the Intergovernmental Oceanographic Commission (IOC) of UNESCO, 23rd Session of the Assembly.

Lee, K., Ryan, S., 2010. Laser In Situ Scattering and Transmissometer Measurements and Supporting Data Collected in Response to the Deepwater Horizon Oil Spill Incident from April through August 2010, NODC Accession Number 0086284. Available online from the US National Oceanographic Data Center at at ⟨http://accession.nodc.noaa.gov/0086284⟩. (Last accessed 30.09.13).

Lu, X., Kataria, S., Brouwer, W.J., Wang, J.Z., Mitra, P., Giles, C.L., 2009. Automated analysis of images in documents for intelligent document search. Int. J. Doc. Anal. Recognit. 12 (2), 65–81. http://dx.doi.org/10.1007/s10032-009-0081-0.

Lunden, J., Georgian, S., Cordes, E., 2013. Aragonite saturation states at coldwater coral reefs structured by Lophelia pertusa in the northern Gulf of Mexico. Limnol. Oceanogr. 58 (1), 354–362. http://dx.doi.org/10.4319/lo.2013.58.1.0354.

Mariano, A., Kourafalou, V., Srinivasan, A., Kang, H., Halliwell, G., Ryan, E., Roffer, M., 2011. On the modeling of the 2010 Gulf of Mexico oil spill. Dyn. Atmos. Oceans 52 (1–2), 322–340. http://dx.doi.org/10.1016/j.dynatmoce.2011.06.001.

Merali, Z., Giles, J., 2005. Databases in peril. Nature 435, 1010–1011.

Miller, T., Gornish, E., 2014. Total Recoverable Petroleum Hydrocarbons from Coastal Dunes in the Northern Gulf of Mexico 2010–2011. FigShare http://dx.doi.org/10.6084/m9.figshare.830439.

Montagna, P., Baguley, J., Cooksey, C., Hartwell, I., Hyde, L., Hyland, J., Kalke, R.D., Kracker, L.M., Reuscher, M., Rhodes, A.C.E., 2013. Deep-sea benthic footprint of the deepwater horizon blowout. PLoS ONE 8, e70540. http://dx.doi.org/10.1371/journal.pone.0070540.

Moody, R., Aronson, R., 2011. Petroleum Hydrocarbon Measurements in Plant Tissue and Sediment from Coastal lagoons in the Northern Gulf of Mexico.

Muhling, B., Roffer, M., Lamkin, J., Ingram, G., Upton, M., Gawlikowski, G., Muller-Karger, F., Habtes, S., Richerds, W.J., 2012. Overlap between Atlantic bluefin tuna spawning grounds and observed Deepwater Horizon surface oil in the northern Gulf of Mexico. Mar. Pollut. Bull. 64 (4), 679–687. http://dx.doi.org/10.1016/j.marpolbul.2012.01.034.

NERRS, National Estuarine Research Reserve System, 2012. System-wide Monitoring Program. Data accessed from the NOAA NERRS Centralized Data Management Office website: ⟨http://cdmo.baruch.sc.edu/⟩ (Last accessed 12.10.12).

Nielson, J.L., Guandique, C.F., Liu, A.W., Burke, D.A., Lash, A.T., Moseanko, R., Hawbecker, S., Strand, S.C., Zdunowski, S., Irvine, K.A., Brock, J.H., Nout-Lomas, Y.S., Gensel, J.C., Anderson, K.D., Segal, M.R., Rosenzweig, E.S., Magnuson, D.S.K., Whittemore, S.R., McTigue, D.M., Popovich, P.G., Rabchevsky, A.G., Scheff, S.W., Steward, O., Courtine, G., Edgerton, V.R., Tuszynski, M.H., Beattie, M.S., Bresnahan, J.C., Ferguson, A.R., 2014. Development of a database for translational spinal cord injury. J. Neurotrauma 31 (21), 1789–1799. http://dx.doi.org/

10.1089/neu.2014.3399.
NOAA/ESRL, National Oceanic and Atmospheric Administration Earth System Research Laboratory 2013. Chemical Sciences Division and university collaborators collected and analyzed these data in response to the Deepwater Horizon oil spill incident in the Gulf of Mexico. Available online at ⟨http://esrl.noaa.gov/csd/groups/csd7/measurements/2010gulf/⟩. (Last accessed 30.09.13).
NOAA/NOS/ORR, National Oceanic and Atmospheric Administration National Ocean Service Office of Response and Restoration, 2013. Collection of scribe databases compiled in response to the Deepwater Horizon oil spill incident in the Gulf of Mexico from 04/23/2010 to 11/08/2011 (NODC Accession 0086261). Available online from the US National Oceanographic Data Center at ⟨http://accession.nodc.noaa.gov/0086261⟩ (Last accessed 30.09.13).
Overton, E., Scott Miles, M., Meyers, B., Gao, H., 2013. Fall 2011 Coastal LA sampling Coastal Waters Consortium (CWC). Accessible in GRIIDC ⟨https://data.gulfresearchinitiative.org/data/R1.x139.142:0005/⟩.
Parsons, M., Godøy, Ø., LeDrew, E., de Bruin, T.F., Danis, B., Tomlinson, S., Carlson, D., 2011. A conceptual framework for managing very diverse data for complex, interdisciplinary science. J. Inf. Sci. 37 (6), 555–569. http://dx.doi.org/10.1177/0165551511412705.
Patterson, W., 2010. SeaBird CTD casts on the Florida Panhandle and Alabama Shelf. Available at University of West Florida Center for Environmental Diagnostics and Bioremediation ⟨http://uwf.edu/cedb/Pattersonctd.cfm⟩.
Reddy, C., Arey, J., Seewald, J., Sylva, S., Lemkau, K., Nelson, R., Carmichael, C.A., McIntyre, C.P., Fenwick, J., Ventura, G.T., Van Mooy, B.A.S., Camilli, R., 2012. Composition and fate of gas and oil released to the water column during the Deepwater Horizon oil spill. Proc. Natl. Acad. Sci. USA 109 (50), 20229–20234. http://dx.doi.org/10.1073/pnas.1101242108.
Rohatgi, A., 2015. WebPlotDigitizerv. 3.8. ⟨http://arohatgi.info/WebPlotDigitizer⟩.
Roman, M.R., Boicourt, W.C., Pierson, J., 2011. CTD Profiles from deployment PE11-06. Part NGOMEX-Living Marine Resources Northern Gulf of Mexico (GoMX-NGOMEX) Project. Accessible at BCO-DMO ⟨http://www.bco-dmo.org/dataset-deployment/454875⟩.
Ryerson, T., Camilli, R., Kessler, J., Kujawinski, E., Reddy, C., Valentine, D., Atlas, E., Blake, D.R., de Gouw, J., Meinardi, S., Parrish, D.D., Peischi, J., Seewald, J.S., Warneke, C., 2012. Chemical data quantify Deepwater Horizon hydrocarbon flow rate and environmental distribution. Proc. Natl. Acad. Sci. USA 109 (50), 20246–20253. http://dx.doi.org/10.1073/pnas.1110564109.
Savage, C.J., Vickers, A.J., 2009. Empirical Study of Data Sharing by Authors Publishing in PLoS Journals. PLoS One 4 (9), e7078. http://dx.doi.org/10.1371/Journal.Pone.0007078.
Shiller, A., Joung, D., 2012. Nutrient depletion as a proxy for microbial growth in Deepwater Horizon subsurface oil/gas plumes. Environ. Res. Lett. 7 (4), 045301. http://dx.doi.org/10.1088/1748-9326/7/4/045301.
Silliman, B., van de Koppel, J., McCoy, M., Diller, J., Kasozi, G., Earl, K., Adams, P.N., Zimmerman, A., 2012. Degradation and resilience in Louisiana salt marshes after the BP–Deepwater Horizon oil spill. Proc. Natl. Acad. Sci. USA 109 (28), 11234–11239. http://dx.doi.org/10.1073/pnas.1204922109.
Snyder, R., Jeffrey, W., 2011. Hydrographic data for the Florida Panhandle Shelf. University of West Florida, Center for Environmental Diagnostics and Bioremediation, United States ⟨http://uwf.edu/cedb/gom-parameters-cruise-data.cfm⟩.
Soniat, T., King, S., Tarr, M., Thorne, M., 2011. Chemical and physiological measures on oysters (Crassostrea virginica) from oil-exposed sites in Louisiana. J. Shellfish Res. 30 (3), 713–717. http://dx.doi.org/10.2983/035.030.0311.
Tenopir, C., Allard, S., Douglass, K., 2011. Data sharing by scientists: practices and perceptions. PLoS One 6 (6), e21101. http://dx.doi.org/10.1371/journal.pone.0021101.
USM, The University of Southern Mississippi Department of Marine Science, 2013. Central Gulf of Mexico Ocean Observing System (CenGOOS). ⟨http://www.cengoos.org⟩ (Last accessed 28.08.13).
Thessen, A.E., Patterson, D.J., 2011. Data issues in life sciences. ZooKeys 150, 15–51. http://dx.doi.org/10.3897/zookeys.150.1766.
Thessen, A.E., McGinnis, S., North, E., NOAA/NOS Office of Response and Restoration, NOAA/NOS Environmental Response Management Application, NOAA/ESRL, CSIRO, NERRS, EPCHC, British Petroleum, Shiller, A., Roffer, M.A., Asper, V., Camilli, R., Clement, T.P., Hayworth, J., Joung, D., Kessler, J., Muller-Karger, F., Pilley, C., Reddy, C., Seewald, J.S., Valentine, D.L., Walker, N., Adams, P.N., Allan, S., Ameen, A., Anderson, K., Arey, J., Aronson, R., Atlas, E., Baguley, J., Bianchi, T., Biber, P., Blake, D.R., Bodinier, C., Boicourt, W.C., Bordelon, A., Briggs, K.B., Brunner, C., Carcia Tigreros, F., Carmichael, C.A., Cebrian, J., Chan, E.W. Cook, R., Cooksey, C., Cordes, E., Daley, M., Daly, K.L., de Gouw, J., Diller, J., Du, M., Dubansky, B., Dzwonkowski, B., Earl, K., Ervin, G., Farwell, C., Fenwick, J., Finnegan, C., Galvez, F., Gao, H., Garcia, T., Gawlikowski, G., Georgian, S., Goni, G., Gornish, E., Green, N., Guo, L., Habtes, S., Halliwell, G., Hartwell, I., Hatch, R., Heintz, M., Hetland, R., Hu, L., Hyde, L., Hyland, J.L., Imhoff, M.L., Ingram, G., Jeffrey, W., Joye, S., Kalke, R.D., Kang, H., Kasozi, G., Keim, J., King, G., King, S., Kinnaman, F.S., Kolic, P., Kolker, A.S., Kourafalou, V., Kracker, L.M., Kujawinski, E., Lamkin, J., Lee, K., Leifer, I., Lemkau, K., Liu, Z., Lohrenz, S., Louchouran, P., Lunden, J., MacDonald, I., MacFadyen, A., Mariano, A., McCoy, M., McIntyre, C.P., Meinardi, S., Mendes, S., Meyers, B., Miles, S., Miller, T., Montagna, P., Moody, R., Muhling, B., Mulabagal, V., Nelson, R.K., Normand, A.E., Ojwang, L.M., Olascoaga, M., Osburn, C.L., Overton, E., Park, K., Parrish, D.D., Patterson, W., Peischi, J., Perdue, E., Peterson, M., Pham, L., Pierson, J., Pino, J., Raghunathan, V., Redmond, M., Reuscher, M., Rhodes, A.C.E., Rice, C.D., Richards, W.J., Roach, J.L., Roman, M.R., Ryan, E.H., Ryan, S., Ryerson, T., Schneider, C.L., Scott Miles, M., Silliman, B., Simpson, S., Smith, B., Smith, R.W., Snyder, R., Soniat, T., Srinivasan, A., Street, D., Sylva, S., Tarr, M., Thorne, M., Trinanes, J., Upton, M., van de Koppel, J., Van Mooy, B.A.S., Ventura, G.T., Villanueva, C.J., Walter, R.B., Warneke, C., Whitehead, A., Williams, M., Wu, W., Yeager, K., Yin, F., Yvon-Lewis, S., Zhang, Y., Zhou, Z., Zimmerman, R., 2014a. Deepwater Horizon Database, Version 1.0. ⟨http://gisr.hpl.umces.edu⟩.
Thessen, A.E., McGinnis, S., North, E., 2014b. Table of Hydrocarbon Terms. Figshare (Retrieved 15:11, May 05, 2015 (GMT))http://dx.doi.org/10.6084/m9.figshare.942516.
Valentine, D., Kessler, J., Redmond, M., Mendes, S., Heintz, M., Farwell, C., Hu, L., Kinnaman, F.S., Yvon-Lewis, S., Du, M., Chan, E.W., Garcia Tigreros, F., Villanueva, C., 2010. Propane respiration jump-starts microbial response to a deep oil spill. Science 330 (6001), 208–211. http://dx.doi.org/10.1126/science.1196830.
Vines, T.H., Andrew, R.L., Bock, D.G., Franklin, M.T., Gilbert, K.J., Kane, N.C., Moore, J.S., Moyers, B.T., Renaut, S., Rennison, D.J., Veen, T., Yeaman, S., 2014. Mandated data archiving greatly improves access to reasearch data. FASEB J. 27, 1304–1308.
Walker, N., Pilley, C., Bordelon, A., Pino, J., 2010. Louisiana State University (LSU) Earth Scan Laboratory (ESL) (⟨http://www.esl.lsu.edu⟩) DWH SAR Oil Spill Contour Database. ⟨http://www.esl.lsu.edu/research/deepwater-horizon-surface-oil-research/⟩.
Whitehead, A., Dubansky, B., Bodinier, C., Garcia, T., Miles, S., Pilley, C., Raghunathan, V., Roach, J.L., Walker, N., Walter, R.B., Rice, C.D., Galvez, F., 2012. Genomic and physiological footprint of the Deepwater Horizon oil spill on resident marsh fishes. Proc. Natl. Acad. Sci. USA 109 (50), 20298–20302. http://dx.doi.org/10.1073/pnas.1109545108.
Williams, M., Cebrian, J., Ervin, G., 2010. Analysis of Petroleum-associated Hydrocarbons in Plants and Sediment of the Northern Gulf Coast following the BP-Deepwater Horizon Oil Spill. Virginia Polytechnic Institute and State University, Dauphin Island Sea Lab, and Mississippi State University, United States.
Zhou, Z., Guo, L., Shiller, A., Lohrenz, S., Asper, V., Osburn, C., 2013. Characterization of oil components from the Deepwater Horizon oil spill in the Gulf of Mexico using fluorescence EEM and PARAFAC techniques. Mar. Chem. 148, 10–21. http://dx.doi.org/10.1016/j.marchem.2012.10.003.