Fisheries
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ufsh20

How to Manage Data to Enhance Their Potential for Synthesis, Preservation, Sharing, and Reuse—A Great Lakes Case Study

Tracy L. Kolb [a], E. Agnes Blukacz-Richards [b], Andrew M. Muir [c], Randall M. Claramunt [a], Marten A. Koops [b], William W. Taylor [d], Trent M. Sutton [e], Michael T. Arts [f] & Ed Bissel [d]

[a] Michigan Department of Natural Resources, Charlevoix Research Station, 96 Grant Street, Charlevoix, MI, 49720

[b] Great Lakes Laboratory for Fisheries and Aquatic Sciences, Fisheries and Oceans Canada, 867 Lakeshore Rd., Burlington, ON, L7R 4A6, Canada

[c] Great Lakes Fish Commission, 2100 Commonwealth Blvd., Suite 100, Ann Arbor, MI, 48105

[d] Michigan State University, 14 Natural Resources Building, East Lansing, MI, 48824

[e] University of Alaska Fairbanks, School of Fisheries and Ocean Sciences, 905 N. Koyukuk Dr., Fairbanks, AK, 99775

[f] Environment Canada, 867 Lakeshore Rd., Burlington, ON, L7R 4A6, Canada

Published online: 07 Feb 2013.

To cite this article: Tracy L. Kolb, E. Agnes Blukacz-Richards, Andrew M. Muir, Randall M. Claramunt, Marten A. Koops, William W. Taylor, Trent M. Sutton, Michael T. Arts & Ed Bissel (2013) How to Manage Data to Enhance Their Potential for Synthesis, Preservation, Sharing, and Reuse—A Great Lakes Case Study, Fisheries, 38:2, 52-64, DOI: 10.1080/03632415.2013.757975

To link to this article: http://dx.doi.org/10.1080/03632415.2013.757975

How to Manage Data to Enhance Their Potential for Synthesis, Preservation, Sharing, and Reuse—A Great Lakes Case Study

Tracy L. Kolb

Michigan Department of Natural Resources, Charlevoix Research Station, 96 Grant Street, Charlevoix, MI 49720. E-mail: kolbt@michigan.gov

E. Agnes Blukacz-Richards

Great Lakes Laboratory for Fisheries and Aquatic Sciences, Fisheries and Oceans Canada, 867 Lakeshore Rd., Burlington, ON L7R 4A6, Canada

Andrew M. Muir

Great Lakes Fish Commission, 2100 Commonwealth Blvd., Suite 100, Ann Arbor MI, 48105, and Michigan State University, 14 Natural Resources Building, East Lansing, MI 48824

Randall M. Claramunt

Michigan Department of Natural Resources, Charlevoix Research Station, 96 Grant St., Charlevoix, MI 49720

Marten A. Koops

Great Lakes Laboratory for Fisheries and Aquatic Sciences, Fisheries and Oceans Canada, 867 Lakeshore Rd., Burlington, ON L7R 4A6, Canada

William W. Taylor

Michigan State University, 14 Natural Resources Building, East Lansing, MI 48824

Trent M. Sutton

University of Alaska Fairbanks, School of Fisheries and Ocean Sciences, 905 N. Koyukuk Dr., Fairbanks, AK 99775

Michael T. Arts

Environment Canada, 867 Lakeshore Rd., Burlington, ON L7R 4A6, Canada

Ed Bissel

Michigan State University, 14 Natural Resources Building, East Lansing, MI 48824

Cómo manejar datos para incrementar el potencial para su síntesis, preservación, intercambio y reutilización –los Grandes Lagos como caso de estudio

RESUMEN: en proyectos complejos, un manejo apropiado de datos (aplicación coordinada de estándares y estructuras a recolección, mantenimiento, recuperación y documentación) resulta esencial para asegurar la precisión y accesibilidad de los mismos. En la presente contribución se utiliza un proyecto de evaluación de los cambios en el crecimiento, condición y reclutamiento del coregono en los Grandes Lagos, como caso de estudio para ilustrar cómo un manejo adecuado de datos puede incrementar y mejorar la investigación. Las mejores prácticas en cuanto a manejo de datos incluyen: dedicar personal a la curación de datos, fijar estándares en los datos, construcción de una base de datos relacional, manejo de actualización de datos, revisión y filtro de errores en los datos, extracción de datos, documentación de bases de datos y coordinación con colaboradores del proyecto. Las acciones de manejo de datos que se tomaron resultaron en la producción de un cuerpo importante de publicaciones y en una base de datos robusta, disponible para investigaciones futuras. Los recursos invertidos en el manejo de datos permitieron que este proyecto sirviera de modelo para tomar los primeros pasos hacia el objetivo común de compartir, documentar y preservar datos que son recolectados y reportados durante el proceso de una investigación científica.

ABSTRACT: *Proper data management (applying coordinated standards and structures to data collection, maintenance, retrieval, and documentation) is essential for complex projects to ensure data accuracy and accessibility. In this article, we used a recent project evaluating changes in Lake Whitefish (*Coregonus clupeaformis*) growth, condition, and recruitment in the Great Lakes as a case study to illustrate how thoughtful data management approaches can enhance and improve research. Data management best practices described include dedicating personnel to data curation, setting data standards, building a relational database, managing data updates, checking for and trapping errors, extracting data, documenting data sets, and coordinating with project collaborators. The data management actions taken ultimately resulted in a rich body of scientific publication and a robust database available for future studies. Investing in data management allowed this project to serve as a model for taking the first steps toward a common goal of sharing, documenting, and preserving data that are collected and reported during the scientific research process.*

CONTEMPORARY FISHERIES RESEARCH NEEDS THOUGHTFUL DATA MANAGEMENT PRACTICES

Data are the infrastructure of science, and modern scientific architecture has become increasingly complex. This trajectory can be partly explained by the preference of granting agencies toward projects that address broad-scale research questions; partly by advances in computing and communications technology that allow the scientific community to work with larger data sets that transcend conventional spatial, temporal, and disciplinary boundaries (Lélé and Norgaard 2005; Wake 2008; Carpenter et al. 2009); and partly by advances in computing that have allowed data-intensive science (Newman et al. 2003) and modeling projects that rely on previously collected data to increase in frequency and magnitude (Kelling et al. 2009; Borgman 2010).

Any research project with multiple objectives or one that combines the expertise of multiple principal investigators—or even one that simply combines data from multiple institutions—will have the capacity to generate immense quantities of varied information, require the assimilation of previously acquired data, or both. This raises a variety of logistical complexities with regard to quality control, security, and accessibility of data and, as such, these projects can benefit greatly from formal data management strategies for entry, update, storage, validation, access, annotation, provenance (i.e., information regarding the origins, identification, ownership and structure of a data set), and archiving (McDonald et al. 2007; Brunt and Michener 2009; Kelling et al. 2009). Recognizing this, many funding agencies now require that all prospective grantees address data management as part of the project application (National Institutes of Health 2003; U.S. Fish and Wildlife Service 2006; National Oceanic and Atmospheric Administration [NOAA] 2010; National Science Foundation 2010).

Unfortunately, modern ecological data management practices have not evolved as quickly as their data sets (Katz et al. 2007; McDonald et al. 2007; Barnas and Katz 2010; Hook et al. 2010). Data management is an often underrecognized and underutilized tool (Michener and Jones 2011). The majority of scientists still manage data through spreadsheet entry, individualized post-entry error checking and manual grouping, or extraction of data for analysis (Porter and Ramsey 2002; Borgman et al. 2007; Nelson 2009) A recent survey of ecologists found that they felt that their own institutions lacked planning, technology, and funding for data management in the short term (during the project) and long term (post-project) and did not adequately provide training in data management (Tenopir et al. 2011). Heterogeneity in the practices and quality of data management limits data reuse, data sharing, and data integration and does not facilitate standardization of data collection methods or support economic efficiency given current fiscal climates.

A fundamental disconnect occurs between the broadly based, complex, interdisciplinary, and collaborative projects requiring data that are accessible, electronic, decipherable, error-free, and reusable and the heterogeneous and idiosyncratic data sets that are routinely being generated from the thousands of fisheries researchers collecting data in the course of their work. Fisheries managers and scientists must embrace the need to recognize data management as a critical step toward organizing their discipline and resolving this tension.

THE STATES OF FISHERIES DATA MANAGEMENT AND PEER-REVIEWED LITERATURE

Scientific data collection and compilation can occur at differing spatial scales, and the larger the scale, the more necessary it is to commit resources to data collection and management. Some examples of larger scale regional fisheries database efforts include FishMAP, a Great Lakes fish migration passage and knowledge database; GLATOS Web, a Great Lakes acoustic telemetry database; the Multistate Aquatic Resources Information System (MARIS); the National Fish Habitat Action Plan; the Pacific Northwest Salmon Habitat Restoration Project Tracking Database; StreamNet, which compiles and disseminates fish data from state, tribal, and federal agencies in the Pacific Northwest; and the Fisheries Information Networks (FINs), which are regional, cooperative, state, and federal data integration and management programs for the Pacific Region (PACFIN), the Atlantic Region (Atlantic Coastal Cooperative Statistics Program [ACCSP]), the Gulf of Mexico (GulfFIN), and Alaska (AKFIN; e.g., Beard et al. 1998; Katz et al. 2007; MARIS 2008; McLaughlin et al. 2010; Wang et al. 2011). In addition, there are regional fisheries databases housed at NOAA Fisheries Service Science Centers, which have long histories of managing data (NOAA 2011). At a smaller scale than these regional efforts, there are the data management endeavors of individual state agencies, coordinating groups of multi-affiliated fisheries researchers, university fisheries research teams, and the many individual projects that require the construction of databases during the course of their research (e.g., Watson and Kura 2006; Katz et al. 2007; Heidorn 2008; Frimpong and Angermeier 2009).

The current state of "how to manage fisheries-specific databases" in the peer-reviewed literature can be summarized as follows: the regional efforts listed above have multiple personnel dedicated to behind-the-scenes data management, using very sophisticated practices, but detailed descriptions of their specific efforts have not been documented in the peer-reviewed fisheries literature (e.g., K. Barnas, Pacific Northwest Salmon Habitat Restoration Project Tracking Database; D. Donaldson, GulfFIN; D. Infante, National Fish Habitat Action Plan; W. Kinney, StreamNet; C. Kruger, GLATOS Web; A. Loftus, MARIS; E. Martino, ACCSP; R. McLaughlin, FishMAP, personal communication). There are also countless textbooks on the structural mechanics of database design (Hernandez 2003; Pratt and Adamski 2007; Ling Liu and Özsu 2009), which tend to ignore the specialized needs of the scientific field. Finally, there are specific fisheries projects that required construction of a database for which the results of the findings have been published, but details of the data management plans have not (Watson and Kura 2006; Katz et al. 2007; Frimpong and Angermeier 2009; Whiteed et al. 2012). Very few generalized descriptions detailing both the technical and practical aspects of managing data generated by typical collaborative research projects are available to fisheries professionals to use as a resource (McLaughlin et al. 2001; Baker and Stocks 2007).

For individuals or teams of fisheries scientists collecting data independent from regional database efforts, formal data management guidance is not readily available as a resource, yet project-specific data management plans are increasingly required as a prerequisite for research grant applications. Therefore, the purpose of this article is to provide a synthesis of data management best practices for the typical fisheries investigator to serve as kind of a broad proxy for grant application and research plans while underscoring the added value that can be accrued by using these best practices. These best practices support data integrity throughout the project and position data to be reusable by future users by ensuring that they are accessible, electronic, decipherable, and error-free.

We used a recent collaboration among federal, state, and university fisheries researchers as a case study to highlight how data management works. Although data management practices are generally masked during the publication process, the authors feel that they are a fundamentally important part of scientific inquiry and communication and therefore should be subject to the same rigorous evaluations and discussions in the primary literature as other scientific methods. Not all fisheries projects will require all of the steps described subsequently, but we hope that this article serves as a guide for researchers asking themselves at the start of a new endeavor, "To what extent and how should data management be implemented for this project?"

LAKE WHITEFISH CASE STUDY BACKGROUND

In 2004, two teams of scientists from Purdue University and Fisheries and Oceans–Canada independently submitted pre-proposals to the Great Lakes Fishery Trust requesting grant money to study Lake Whitefish (*Coregonus clupeaformis*) recruitment dynamics in the Great Lakes (Sutton et al. 2007). In a rare occurrence, the reviewers felt that the projects were similar enough to suggest that the two teams collaborate (M. Coscarelli, personal communication). Recognizing that choosing not to collaborate meant competing as two similar projects vying for a limited pool of money, the groups merged, submitted a full proposal for one project that addressed two different sets of potential Lake Whitefish recruitment impediments, and received funding for 3 years, where yearly funding depended on the success of the collaboration. The researchers convened as a group to discuss issues associated with data management (Table 1). The result of that discussion was agreement that the expanded project and conditional nature of the funding required implementing formal data management practices and a decision was made to allocate project resources to obtain a dedicated data curator as a permanent member of the research team.

TABLE 1. Discussion items that help identify the data management needs for a collaborative research project.

Given that we want to store all project data together, does a single member of the research team have the skill set and time to manage data for the entire project? Do we know someone reliable but outside of the research team who could curate the data?
How much data will we be collecting? What is the maximum size of our data set?
Once we have collected data, will housing them require multiple tables? Can we use "flat file" (single data table) organization or do we need a relational database?
How complicated will data entry be? How many different people will be entering data, at how many different locations? The more complex data entry, the greater the probability of errors and the more dedicated error oversight required.
If multiple PIs are working on separate parts of the project, how important is it that their data be able to interact? Do the PIs need to combine data to answer research questions? If so, properly defining relationships among data is critical.
Does our grantor require data management or data sharing as part of our grant stipulations? Will our data be shared beyond the PIs?
If we need to use a relational database, how much will it change through time? How many researchers will need to access identical data simultaneously but separately? Will version control be critical to ensure that everyone is accessing the same data?
Are our data unique and can they be reproduced? Will we want to draw from these data sets for future studies? Is it worth the investment to preserve our data?

DATA MANAGEMENT BEST PRACTICES

Selecting a Data Curator

A data curator is responsible for the technical and practical aspects of data management throughout a research project—although for large, complex projects, data curation is often done by a team of individuals, which may include subject-matter experts, data users, information technology staff, computer programmers, and a metadata librarian (Lord et al. 2004; Cragin et al. 2008; Akmon et al. 2011). A curator's major responsibilities are to incorporate, organize, document, and retrieve data that they curate (Heidorn 2008; Witt 2009; Witt et al. 2009). The curator adds value to the research project by checking, verifying, and correcting data sets, as well as by providing software tools for data access, manipulation, and assimilation of any previously collected data, if required (Research Information Network [RIN] 2008; Cragin et al. 2010). Data curators apply rigorous procedures to ensure that the data sets they manage meet quality standards in relation to the structure and format of the data themselves (examples given in the following sections), ultimately contributing value by making data more discoverable and easier to access for potential reuse. A dedicated curator combines the benefits of expertise available to researchers in disciplines with centralized data repositories with the agility and advantages of localized data storage and management (RIN 2008). Though formal training in database design and management is ideal, a data curator need not be a professional database developer or computer programmer; he or she can simply be someone who has experience and is comfortable managing data. Our data curator was a postdoctoral researcher with experience managing modest-sized (<1 million records) databases obtained during past research projects.

Establishing Data Requirements

Before any Lake Whitefish project data collection occurred, our curator's job was to determine what data were going to be collected and by whom and who would be responsible for post-collection processing of data (Hernandez 2003; Pratt and Adamski 2007; Ling Liu and Özsu 2009). Three universities, two U.S. natural resources agencies, three Canadian government agencies, one tribal resource agency, and various commercial fishing operators worked together to collect data at 13 sites across lakes Michigan, Superior, and Erie from 2004 to 2006 (Figure 1). Adult, juvenile, and larval Lake Whitefish were sampled using gill and trap nets, beach seines, and plankton nets, respectively. Sampling effort parameters (e.g., date and location) and environmental data (e.g., water temperature) were collected during each sampling event. Biological data (e.g., length and weight) were collected on each life stage of the Lake Whitefish and their prey. Subsequent laboratory analyses resulted in the generation of further biological and physiological data (e.g., food habits, proximate body composition, and fatty acid composition). These collections resulted in a data set with more than 250,000 records. To ensure that data collection was standardized, our curator held initial meetings with individual project collaborators and collective meetings with the research group. During the individual meetings, the curator asked the collaborators the following questions:

- How are data defined; what formats will these data take (e.g., numbers, pictures, acoustic records, physical specimens, etc.); what are the units of measurement associated with numerical data; which data are textual? What information about data collection will be archived (e.g., sampling effort data such as weather, collection gear, sampling crew names, etc.)? How many records will be generated seasonally and over the entire project?

- How will data be captured or created (e.g., research vessel, fish tagging, moored buoy, online surveys, etc.)?

- What are the spatial and temporal coverage of data collections?

- Once data are collected, will they be postprocessed? If so, where will they be sent and what processes will occur?

- What are the timelines for data collection and post-processing? Are data being collected all at once or throughout a season? Are data being generated and recorded continually or in batches?

- How soon after processing or collection will data be sent to the curator for input into a database and how soon after input will data be needed for analysis? Will data be transferred all at once or in batches?

- How do data relate to other data (e.g., will a sampling event be related to multiple fish caught during that event, or will multiple stomach contents be related back to an individual fish)?

An initial meeting with the entire research team allowed for the development and documentation of a predefined set of standards for coding categorical data, such as sampling locations and Linnaean names of fish and invertebrate species (we recommend using standard Integrated Taxonomic Information System codes) and to determine how spatial and environmental

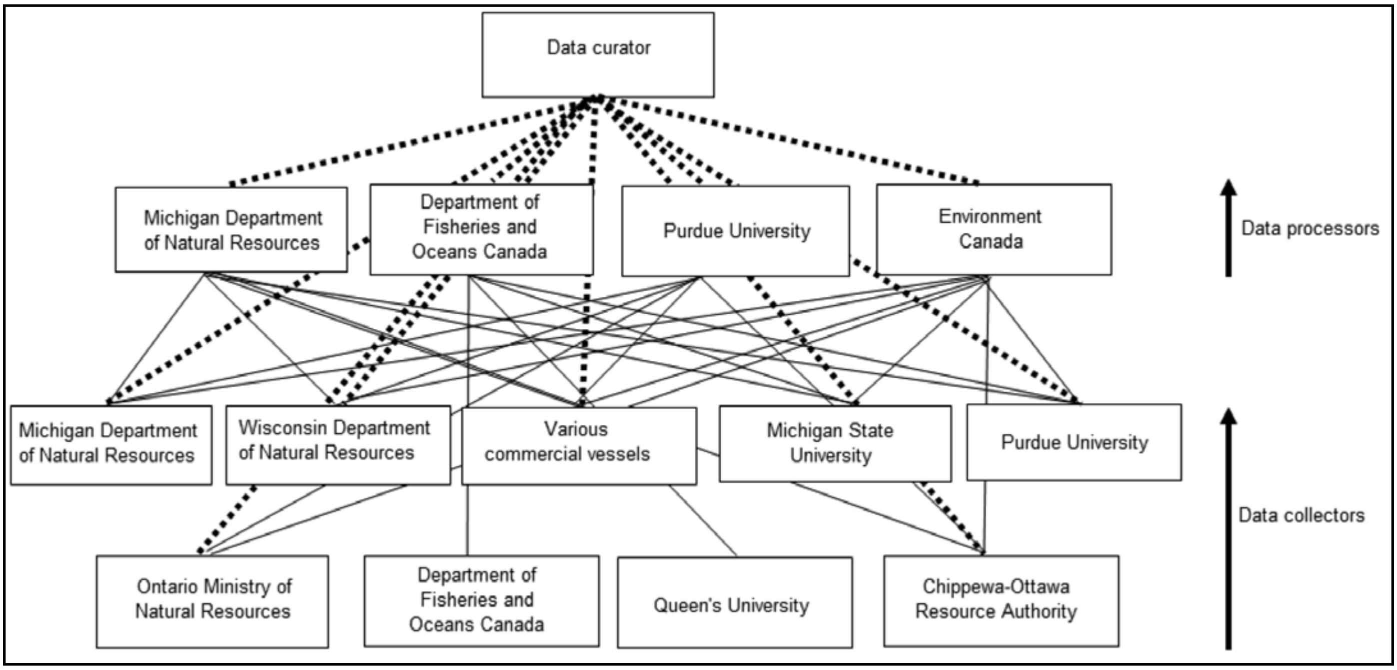
Data curator
Michigan Department of Natural Resources
Department of Fisheries and Oceans Canada
Purdue University
Environment Canada
Data processors
Michigan Department of Natural Resources
Wisconsin Department of Natural Resources
Various commercial vessels
Michigan State University
Purdue University
Data collectors
Ontario Ministry of Natural Resources
Department of Fisheries and Oceans Canada
Queen's University
Chippewa-Ottawa Resource Authority

Figure 1. Schematic of the complexity involved in collecting and processing data for a Great Lakes Fishery Trust–funded Lake Whitefish project (Sutton et al. 2007). Lines indicate data exchange between entities. Solid lines represent data exchange between collector and processor. Dashed lines represent data exchange between a collector or processor and the data curator. N.B., schematic organizes data collectors in rows for formatting reasons and does not imply any type of hierarchy.

data would be captured and classed. For example, collaborators agreed to standardize classification of all nets as trap, gill, seine, or plankton. These initial meetings helped the collaborators to improve their understanding about the scope of the project and facilitated standardizing field sampling methods. Although it seems intuitive that collaborative projects would function this way (i.e., with or without a dedicated curator), the lack of a dedicated person accountable for bringing data management issues to the group and forcing standardization at the onset of the project often results in a situation whereby issues, sometimes uncorrectable, are overlooked until much later in the research process, increasing the time required to identify and correct errors (McLaughlin et al. 2001; Wallis et al. 2008).

Creating and Populating the Database

After becoming familiar with project data and collaborators' collection procedures, a data model was created. The data model and associated entity relationship diagrams identified all attributes (i.e., data elements) in each table and defined how tables related to each other through key fields. Our model called for seven primary data tables and 11 lookup tables (Figure 2, Table 2). After the data model was created, we developed a means for data storage (i.e., relational database) using Microsoft Office Access software. (Special note on spatial data: There are two options when working with spatial data: (1) using an aspatial, tabular, relational database where records have a unique identifier, which can be linked to geocoordinates in a second, separate spatial database; or (2) working exclusively within a single spatially enabled database [supports spatial data types]. In the second case, each individual record's spatial geometry is stored as an attribute and the database is integrated with the spatial software. We used the first option and stored only the latitude and longitude of our sampling events, importing those spatial attributes into ArcGIS when we needed to make maps or wanted to do specialized spatial analyses.)

Although we used Microsoft Access to develop our database, many relational database management software programs (RDBMSs) are available to researchers (Table 3).When selecting an RDBMS, researchers should consider the advantages and disadvantages of the price, required operating system, compatibility with other software programs, user accessibility, level of technical expertise, anticipated upgrade costs (time and money), and constraints imposed by the quantity of data to be managed.

Most scientists use spreadsheet software to store their data (e.g., Excel), rather than an RDBMS (Borgman et al. 2006). Though both spreadsheets and RDBMSs organize data in rows (in data storage language these are called "records") and columns ("fields"), spreadsheets store data in individual tables as "flat files," meaning that tables are not linked, whereas RDBMSs store data across multiple, interrelated tables, with the expectation that the user will primarily want to work with data across multiple tables simultaneously. Storing data using linked tables is the foundation of the relational database.

If data can be stored in a single table, a relational database is not necessary. If multiple tables are required to store data, creating a relational database using an RDBMS is the best option, because relational databases have rules that maximize data integrity across tables (Hernandez 2003). Generally, data integrity rules include the following: (1) tables that are constructed properly and efficiently (i.e., each table represents a single entity, every column in each table is comprised of distinct fields, fields are not repeated within a table, and each record is identified with a unique value called a "primary key" used for linking data among tables); and (2) data integrity (validity) is imposed at the record, table, and relationship levels (i.e., every table has a column for the key field and keys are used to create relationships among tables).

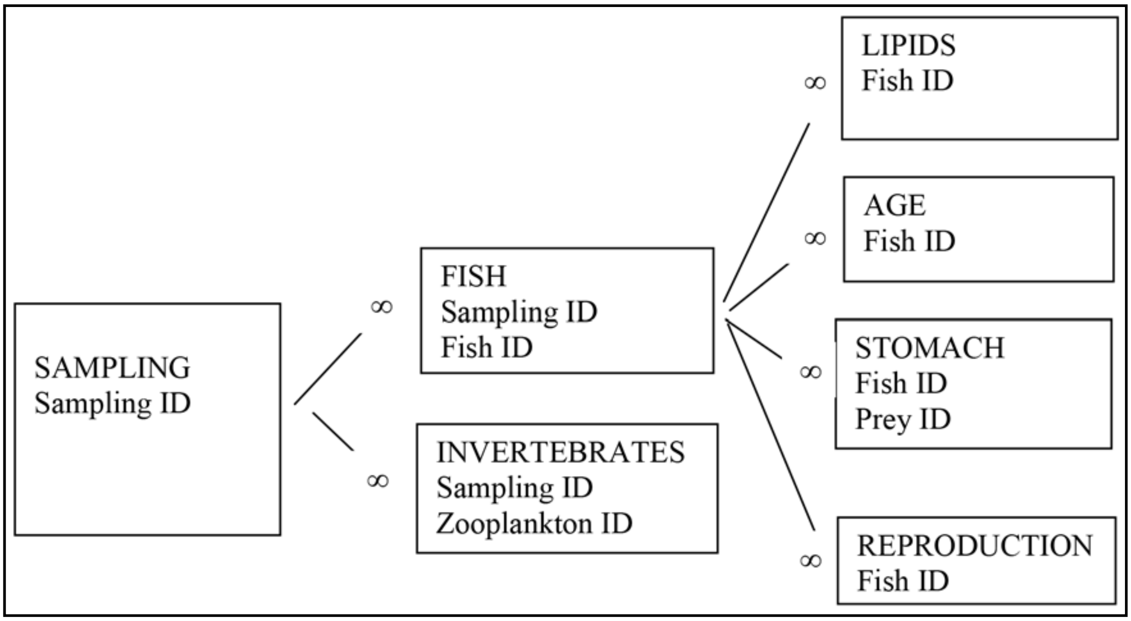

Figure 2. Data model—design of the Lake Whitefish recruitment database indicating the primary data tables, relationships between tables, and fields used as keys in the relationships. Additionally, 11 lookup tables (not included in figure) standardized the entry of location, gear, weather conditions, fish species, life stage, sex, maturity, invertebrate species, fatty acid, age structure, and prey species data (see Table 2). The ∞ symbol represents a one-to-many relationship between table IDs. For example, one sampling ID in the sampling table can relate to more than one fish or invertebrate ID in the fish and invertebrates tables, or one fish ID in the fish table can relate to more than one lipid, age, stomach, or reproduction ID in their respective tables.

An experienced curator is able to harness the strengths of the relational database model and

TABLE 2. Database table names (columns) and non-key fields (rows) for the Lake Whitefish recruitment database. Italics indicate fields linked to lookup tables.

Sampling	Fish	Invertebrates	Lipids[b,c]	Age[a,b,c]	Stomach[b, c]	Reproduction[c]
Date	*Species*	*Species*	Total lipids	Age	Stomach weight	Gonad weight
Time	*Life stage*	Length	*Fatty acid type*	*Structure used to determine age*	*Prey species*	Egg diameter
Location	Length	Weight	Fatty acid concentration		Prey frequency	Egg weight
Latitude	Weight	*Sex*			Prey weight	Sperm velocity
Longitude	*Sex*	Biomass				Sperm tail length
Depth	*Maturity*	Density				Sperm cell volume
Gear	Liver weight					Milt volume
Tow speed	Body condition					Mean spermatocrit
Tow distance	VFI[1]					
Ambient conditions	Protein[2]					
	Energy[2]					
	Moisture[2]					

[1] Visceral fat index; [2] in muscle tissue.

Alphabetic superscripts delineate data collected for life stages or groups as follows:

[a] larval Lake Whitefish.

[b] age-0 juvenile Lake Whitefish.

[c] adult Lake Whitefish.

software by taking advantage of the built-in procedural logic that relates information among tables, allowing users to focus solely on using declarative logic to extract data; for instance, when combining data from multiple spreadsheets the user has to manually relate data among different tables, whereas in the relational database the user simply has to indicate data for extraction because relationships between tables are predefined. The chances are lessened that data will become useless if knowledge of those relationships is ever lost because the relationships among data are required to be declared explicitly. Using an RDBMS, data tables can adapt to changing sampling designs and protocols without necessitating structural changes so that new data can easily be incorporated in the future. RDBMSs also offer several advantages related to data integrity and quality compared to spreadsheets. Properties of atomicity, consistency, isolation, and durability (ACID) describe the various mechanisms that the underlying RDBMS software uses to ensure data integrity between transactions (Haerder and Reuter 1983). Though spreadsheets optimize flexibility and ease of use by pairing data storage with visualization, RDBMSs optimize data integrity through ACID principles (Haerder and Reuter 1983; Hernandez 2003).

One foundation of ACID principles is the key field (previously mentioned), which is defined as a unique identifier that links data across tables. Keys provide the quickest way to retrieve data when searching or sorting and make it easy to summarize data from multiple tables. Keys can assume multiple formats as long as they are unique. The simplest format for a key is an autonumber, where each new record is assigned a sequential number starting at 1. In the Lake Whitefish recruitment database, the key for identifying each individual Lake Whitefish was a concatenation of sampling date, sampling location, life stage, and fish ID (e.g., 05_06_2007_ElkRapids_AD_001); this allowed the key itself to convey meaning and to function as more than just a serial number.

Finally, RDBMSs offer several additional advantages over spreadsheets, including the ability to store, manage, and analyze data sets of considerably larger size. RDBMSs run so efficiently because they only retrieve data required through a user-specified query, whereas spreadsheets load the entire data set into memory when the spreadsheet file is opened. In addition, the ability to partition a database into multiple files across multiple hard disks can reduce disk contention (bottlenecks caused by multiple processes accessing the same location on disk at the same time), making large and complex databases easier to work with. Additionally, RDBMSs use indexing to speed up which query results are returned for large data sets by reducing the number of records that must be scanned to return the desired result (Ling Liu and Özsu 2009).

Version Control

Sampling and postprocessing of samples collected for the Lake Whitefish project occurred over 3 years; therefore, the coordination of data submittal and updates to the database was done semiannually by the data curator. Every time new data were uploaded or existing data were corrected, a new version of the database was created.

It is critical that the curator exert control over the perpetuation of multiple versions of a single database. If version control is not implemented, different versions of files, related files at different locations, and information cross-referenced among files are all subject to the viral phenomenon of cascading repli-

TABLE 3. Comparison of available technologies to manage data.[a] Structured query language (SQL) is a programming language designed for working with relational databases. Other considerations include operating system or integration with other clients (desktop or geographic information system software).

Concerns/needs	Spreadsheet—basic	Database—intermediate	Database—advanced
Desktop or server based	Desktop	Desktop	Server based
Spatially enabled	No	No	Yes
Security	No	Low	High
Multiuser data entry	No	No	Yes
Size of data set	Limited	Limited	Unlimited
Web-based	No	No	Yes
Examples	Excel	Access, Microsoft SQL Express, SQLLite	SQL Server, Oracle, MySQL
Cost	Low	Low	High (although some open-source RDBMSs are available)
Level of programming experience needed	None	Little to none	Expert

[a] This table is meant to indicate the general advantages/disadvantages of the different tiers of technologies to manage data, and the characterizations do not hold true in all cases.

cates being distributed, utilized, and reproduced. How version control is implemented depends on whether there are single or multiple users and whether versions across space and time need to be synchronized (Van den Eynden et al. 2011). Our suggested best practices for a version control strategy are listed in Table 4.

One of the most elegant advantages of using relational databases is that they can be programmed to be as self-documenting as the users require. "Self-documenting" refers to the process in which data transactions are logged along with their identifying features, such as who ordered the transaction, a time/date stamp identifying when the transaction occurred, and the nature of the specific transaction. We designed our database to be self-documenting in the sense that all changes to data were recorded in the database itself, so users could query which data were updated, by whom, and when. Each time the database went through major updates, a description of what occurred was provided to users along with the new version. We implemented versioning control using an FTP site, because the site offered security and ease of distribution with minimal upfront programming. At the end of the project, our curator ensured that each collaborator was provided with a final version of the database that contained not only verified data but also previous versions of the database with records of any updates that occurred during the course of the project. This allowed the collaborators to audit revision history and recover deleted information if necessary.

Quality Control and Standardization

One of our curator's most valuable contributions to the Lake Whitefish recruitment project was quality control and standardization of data. Initially quality control involved coordinating the group to use standardized naming conventions and code lists during field collection and data processing. A secondary layer of quality control operated inherent in the relational design model and embedded integrity rules—eliminating duplicates, reducing redundancy, increasing consistency, and maximizing accuracy of data (Hernandez 2003; Borer et al. 2009; Hook et al. 2010). The creation of formal data definitions disallowed data of one type to be entered as another type (e.g., wind direction was stored as text so it was impossible to accidentally store it as a number). Our curator also placed checks on data ranges and coding through lookup tables for gear types, sampling sites, and fish identification codes and used input masks to ensure that data were formatted consistently.

One data management best practice is to audit, capture, and intercept suspect values as close to the source of the collection as possible because reworking data long after the actual sampling is complete is more time consuming and can result in unusable samples. To verify that data were accurate, our curator performed simple reviews of data by periodically checking for missing values; verifying that data were in their proper columns; scanning for impossible values; and generating simple statistics such as frequencies, means, ranges, and clusters to detect errors or anomalous values (e.g., the lengths of all age-0 juveniles should fall between 20 and 70 mm; the latitude and longitude for Little Bay de Noc in northern Lake Michigan cannot map to southern Lake Michigan, etc.). Our curator also established more sophisticated error traps by communicating with project collaborators to identify what errors were most common to their data contributions. For instance, because Lake Whitefish eat small diet items, if a diet item exceeded 20% of the total length of the predator, the sample was flagged. Sampling event data and individual fish biological data were entered into the database at different times and in different places, so our curator created a query to verify that the total number of fish recorded during a sampling event was equal to the number of biological records for individual fish associated with that sampling effort.

Our curator also verified that collaborators were not violating their own sampling protocol. For instance, the collaborators agreed that when sampling adult female Lake Whitefish for fecundity, they would only collect eggs from females that were "green" (i.e., with mature eggs that were still attached to the skein) and not from females that were "running or spent" (i.e., eggs that were free flowing or had already been deposited). After data collection, egg samples were sent to one laboratory and female fish were sent to another for analysis. Upon receipt of data, the curator could use simple queries to easily match up all of the samples to ensure that fecundity was not estimated for fish that fell outside of the maturity requirements.

Data Extraction

Scientists may not have been comfortable extracting data from a relational database, yet they still needed to be able to

TABLE 4. Version control strategies—best practices.

Identify a single location for the storage of versions (in the Lake Whitefish case a secure FTP site).

Decide how many versions to simultaneously maintain (in the Lake Whitefish case, one version).

Uniquely identify versions using a meaningful naming convention, which should include the status of version (e.g., draft, working, final).

Record changes made to each version and maintain old versions for backups.

If applicable, manage any merging of entries or edits by multiple users.

Police users so that multiple working versions are not being developed in parallel.

Set permissions to read and write data to the database.

Develop formal procedures for destruction of any master files.

Properly document all version control procedures.

easily extract data for analysis. One of the greatest benefits of having a data curator was that collaborators could simply send an e-mail or make a call and ask for data to be assembled and formatted in whatever way was needed and the curator could deliver those data quickly and efficiently. For more generalized data selections, the curator set up standardized forms in the database with checkboxes that allowed collaborators to select data without assistance. As an overseer of data extractions from the database, our curator could ensure that two collaborators accessing the same data were doing so identically, decreasing the likelihood that different conclusions might be reached because different data were selected for analysis.

Because most statistical software requires data in a flat format, it might seem counterintuitive to take the time to create a database only to extract and flatten data for analysis. However, having taken the time to standardize, assemble, and properly structure data, there is no end to the various combinations that the curator can provide to the collaborators, and extracting data in any flat format takes mere seconds. Multiple statistical software packages allow selection of data within a database via structured query language (SQL); for example, users of R (R Development Core Team 2011) have several CRAN packages available that retrieve the results from relational databases as entire data frames (R Development Core Team 2011).

Initially, most collaborators felt uneasy allowing a data curator to develop and manage a database because they thought that it might limit their influence over extraction and analysis processes, thereby increasing the distance between themselves and their data. Consequently, our curator ensured that all collaborators had open access to the database and served as teacher and advisor for those who wanted to learn how to extract their own data, while allowing those who were familiar with databases the freedom to extract on their own. In fact, most of the collaborators eventually ended up extracting their own data as the project matured and they became familiar with the database structure and operation.

Data Documentation and Archival

Ultimately, the value of data are enhanced, not exhausted, by their subsequent publication and use (RIN 2008; Whitlock 2011). If data are not properly documented, no one outside of the original collectors will be able to use them properly; and because memories fade, eventually even the data originator may have trouble recalling important information relevant to a data set (Akmon et al. 2011). Broadly, “documentation” (descriptive information about data sets, also called “metadata”) includes the following components: what data are; when they were collected; how they were collected; geographic scope of the project; contact information of collectors; directions for citation; any information relevant to interpretation (e.g., processing that occurred, confounding factors, how missing data were handled, quality assessment, projection information, etc.); and individual definitions for each data field (see Table 5 for an example of data documentation for a single table of the Lake Whitefish database).

Multiple standards provide models for data documentation; the most comprehensive and broadly applicable are the Federal Geographic Data Committee Content Standard for Digital Geospatial Metadata (FGDC-STD-001-1998) and the International Organization for Standardization standards (ISO 9001:2011). For the Lake Whitefish project, our curator used Federal Geographic Data Committee standards to create a formal data dictionary, which was provided to each collaborator at the conclusion of the project. An object linking and embedding reference to the data dictionary was embedded in the data set so that it can continue to be accessed and interpretable into the future. For broadest access, it is best to archive data using open-source formats rather than proprietary formats when possible.

Scientific Value-Added Aspects of Data Management

As is becoming more common in research, field and laboratory samples for the Lake Whitefish recruitment project were obtained by multiple collectors at sites separated by substantial geographic distances. In consultation with the data curator, researchers were able to efficiently merge field and laboratory data contained in the relational database and effectively extract them to investigate complex relationships and identify mechanisms related to the effects of declines in Lake Whitefish growth and condition on recruitment potential of populations across the Great Lakes.

For example, all sampling events were recorded in the sampling table where each event had its own unique ID. Then, fish caught during a sampling event were stored in their own fish table, where every fish had its own unique ID *and* the ID of its sampling event. When lipid analyses were done on an individual fish, lipid data were stored in their own lipid table along with the ID from the fish the lipids were extracted from. Thus, even though these pieces of information were being *stored* in separate tables, the relational database, which linked the IDs among tables, allowed analyses to be performed across all fields without the onerous manual linking required if spreadsheets were

TABLE 5. Parameter documentation for sampling table in the Lake Whitefish database.

Sampling table			
Column	**Definition**	**Format**	**Example**
Date	Date of sampling	mm/dd/yyyy	06/25/2005
Time	Time (military) gear was deployed	hh:mm	13:40
Location	N, Naubinway; BBdN, Big Bay de Noc; ER, Elk Rapids; BH, Bailey's Harbor; L, Ludington; S, Saugatuck, PP, Point Peelee; WP, Whitefish Point; RB, Rowley's Bay; M, Muskegon; MC, Manitowec; FI, Fisherman's Island; BB, Brimley Bay	Text	N
Latitude	Stores the latitude of the sampling site in decimal degrees[a]	Number (double)[b]	42.64606
Longitude	Stores the longitude of the sampling site in decimal degrees	Number (double)	−86.22633
Water temp	Water temperature (°C)	Number (double)	10
Air temp	Air temperature (°C)	Number (double)	12
Depth	Stores the depth of the sampling station in meters.	Number (double)	2.3333
Gear	TN, trap net; GN, gill net; S, seine; NN, Neuston net; G, grab; SL, sled; MN, mysis net; H, hydroacoustics	Text	GN
Tow speed	The speed (m/s) at which the sampling gear was towed	Number (double)	1.5
Tow distance	The distance (m) over which the sampling gear was towed	Number (double)	822
Comments	Any comments related to the sampling event	Memo	PI was arrested by Saugatuck police

[a] Our data were projected in the World Geodetic System (WGS84)—this information belongs in the metadata or data dictionary that describes each sampling parameter in detail.

[b] Double means that the number is noninteger.

used. Using the links among the sampling, fish and lipids tables, Muir et al. (2010) found that female Lake Whitefish in poorer physiological condition had a tendency to produce age-0 juveniles with poorer body composition at some sites, but this pattern was not evident across all sites. In the same manner, using the linkages among three other tables (sampling, fish, and stomach), Claramunt et al. (2010a) were able to partially explain this spatial pattern of juvenile condition by showing that early life history survival was likely dependent on favorable growth early in development, which allows an earlier ontogenetic diet shift to emergent spring macro-invertebrates, demonstrating that the link between parental and juvenile physiological condition was influenced by early life growth rates.

The male contribution to Lake Whitefish reproduction and recruitment potential was explored by Blukacz et al. (2010) using linkages among the fish, reproduction, and lipids tables to show that male fish in better condition tended to produce higher quality sperm, suggesting that males are not irrelevant to Lake Whitefish recruitment potential. In addition, the linkages between the sampling and fish tables were used by Claramunt et al. (2010b) to relate larval fish densities to several abiotic and biotic factors, including adult stock size, abiotic conditions during incubation, and spring productivity. It was precisely the extracting and combining data using linked tables that enabled the research team to efficiently address more complex and related questions and provide a more thorough understanding of Lake Whitefish recruitment potential.

Follow-on components of the Lake Whitefish recruitment project were also able to benefit from efforts to create and curate the Lake Whitefish relational database. For example, some members of the Lake Whitefish recruitment project team secured additional funding to analyze stable isotopes from tissue samples archived during the original project sampling. The project database made it straightforward to match the stable isotope data to the original project data through the addition of a stable isotope table. Information queried from linkages among the fish, lipids, and new stable isotope tables was used to address questions about the connection between Lake Whitefish condition and prey quality (Fagan et al. 2012) and the use of C:N ratios to predict lipid content (Fagan et al. 2011).

DATA CURATION BEYOND THE SINGLE RESEARCH TEAM

It is the sum of all of the actions our curator took to ensure proper data management throughout the life cycle of the project (setting initial standards, coordinating data transfer, building a relational database, managing data updates, error checking/trapping, data extraction, data documentation, teaching, coordinating and communicating with project collaborators) that ultimately resulted in a rich body of scientific publication and a robust database available for future study (Brenden et al. 2010). In our experience, the efforts toward effective data management more than justified the time and expense.

But does data management for a single project have benefits beyond that project? Does it behoove a granting agency to impose proper data management if the collaborators are already going through the formal publication process? We suggest that it does, because a well-designed and defined database is the equivalent of formal documentation of scientific methods. In a time of restricted financial resources, grantors who want to maximize the scope of their investments would be well served by minimizing the redundancy in data collection that occurs when data sets are lost through lack of proper management and, by extension, archiving.

Devoting resources to data management has benefits even beyond increasing the consistency and accuracy of individual project data. Relational database approaches facilitate integration of information from multiple sources, affording more robust, scientifically defensible decision-making capabilities (McLaughlin et al. 2001; Baker et al. 2005; Baker and Stocks 2007). Effectively documented and structured data sets encourage data sharing and communication among collaborators by motivating them to make explicit all of the nuances of their data (McLaughlin et al. 2001; Porter and Ramsey 2002; Birnholtz and Bietz 2003). Databases can serve as storage for unique or irreplaceable records that can only be properly preserved for reuse though effective documentation and management (Brunt 1994; Borgman et al. 2007; Heidorn 2008). Only well-managed and documented data allow for reproduction of research where checks and balances operate at the most fundamental level (Parr 2007; Heidorn 2008; Borgman 2010). Others believe that because most research is publicly funded, data belong to society at large, and best practices should be used when managing those data for preservation and reuse (Costello 2009; Guttmacher et al. 2009; Borgman 2010). Effectively managed data allow for repurposing, thereby saving money that might otherwise be used for redundant collections (Hale et al. 2003; Carlson and Anderson 2007; Heidorn 2008). Finally, properly documented and organized data have unlimited potential for reuse by providing archival material to address future problems, thereby advancing science in ways possibly unforeseen by the original collectors (Postel et al. 2002; Nelson 2009; Borgman 2010).

One contentious issue surrounding data reuse is reluctance by researchers to share data beyond the original collaborators or close colleagues. Secrecy in guarding research has been part of scientific culture throughout history, and recent articles exploring the data sharing attitudes find scientists overwhelmingly unwilling to freely share data within and among their own community (Blumenthal et al. 1997; Campbell et al. 2002; Blumenthal et al. 2006; Vogeli et al. 2006; Haas 2011; Tenopir et al. 2011), where willingness to share data is positively correlated with the ease of extraction and relationship to requestor (Witt et al. 2009; Cragin et al. 2010). In some sense, curators negate certain issues surrounding resistance to sharing that have to do with expending time and energy to prepare data, but addressing the underlying scientific-professional reward structure that does not reward sharing remains outside their scope of influence (McDade et al. 2011).

Issues surrounding ownership and security also determine the extent to which data are shared (Beard et al. 1998). When research projects are funded by federal government agencies, philanthropic organizations, or private industries, grantor-specific stipulations often influence how data will be retained and disseminated (Fishbein 1991) as well as being subject to the Freedom of Information Act (5 U.S.C. 552). One simple solution to data sharing and ownership issues is a data sharing agreement. Data sharing agreements should be specific to each project and should include intended level of exposure (e.g., within the group only, within the field only, publicly accessible), level of control applied to data outflow, whether an embargo period will be applied to data availability, and how data will be recognized when being used by others. In our case, our data sharing agreement stipulated that data would flow freely among principal investigators (PIs) and that each PI could decide to share or not share their portion of the data beyond the original collaborators at their discretion.

Building and managing databases can be challenging, especially if long-term data management is underfunded. Granting institutions may recognize the benefits of requiring data sets as deliverables but may also be loath to become their ultimate resting place. One field that is taking on the challenge of long-term digital curation is library science. University libraries are creating institutional repositories as part of a larger technology and service structure that can contribute resources and expertise in data curation (Cragin et al. 2010). Data centers (open-standard, interoperable, nonproprietary web services) are also becoming widely established (Baker and Bowker 2007; Costello 2009). The lure of data centers is that by providing open or semi-open access to data, they act as a dual facilitator for finding and storing data and, as of yet, no one repository has been established as the mainstay for fisheries data. Examples of established open-access ecological data repositories are the Long Term Ecological Research Network, DataONE, and MARIS (Baker et al. 2000; MARIS 2008; Michener et al. 2012). All three provide a framework for assimilation and management of disparate data sets with tools for data discovery and guidance on data management. The NOAA also has its own sophisticated internal data centers, whose services, as far as we were able to ascertain, are not available to non-NOAA researchers.

Not everyone is sold on the idea of depositing their data in open-access repositories though. Tenopir et al. (2011) found that only 15% of ecologists expressed a willingness to place their data into an open-access repository, and the majority expressed different conditions for doing so, including the following: opportunities to collaborate (80%), mandatory reprints provided (75%), coauthorship (65%), results of analyses not disseminated without data providers' approval (46%), legal permissions obtained (40%), and monetary reimbursement (28%). Not included in the survey was an embargo period allowing PIs the first chance to publish on data, but we assume that would also be a consideration of data providers. To mitigate issues specifically related to recognition, formal and consistent citation of databases will need to become more common in our field (NOAA 2012).

The concept of depositing data in an open-access repository was so foreign to the Lake Whitefish project team that we never seriously considered using one for our data set. This decision resulted in data that can now only be obtained through communication with a PI personally. We realize that our decision not to use a repository undermines a key message of this article, which is that data will not remain accessible without a plan for their preservation (Uhlir 2010), but the decision also accurately depicts the state of existing data preservation practices of most scientists in our position and field. We believe that whatever the future of institutional repositories and open-access data centers, they will continue to stay underutilized if they cannot support existing data practices specific to each scientific field and adequately mitigate the cultural issues associated with data sharing and recognition.

Given the shift toward large collaborative projects, we predict that formalized data management will become a more integral part of research and will require explicit allocation of funds and recognition of professional productivity (McDade et al. 2011). We recommend that resource commitment associated with data management be estimated at the outset of the project so that dedicated resources can be requested from funding agencies for proper data curation. We also exhort funding agencies to support these requests for additional resources by realizing the benefits they provide in ensuring data availability for future research. Furthermore, an actual data set offered as a product of externally funded research is perhaps one of the most concrete and useful deliverables that can be produced as return on investment.

Organization is an emergent property for any complex system, and efforts like the Lake Whitefish database are necessary as first steps in developing greater information organization within the fisheries research community. Looking beyond the development of a single database ultimately probes at a number of underlying systemic issues relating to large-scale information leveraging, in particular, resistance to sharing data, how to preserve and use historical data sets, the general lack of methodological standardization, and assessing whether the creation of these large-scale data endeavors yields returns enough to justify their investment in resources. Ultimately, the fisheries community should continue to examine ways to improve efficiency (reduce fragmentation) in research, reduce the duplication of effort in data collection, and spearhead efforts to coordinate data standards at a national level in order to adequately transfer scientific information. This can only be accomplished if we take the first steps toward a common goal of sharing, documenting, and preserving data that are collected and reported.

ACKNOWLEDGMENTS

We thank E. Volkman, C. Benoit, A. Charlton, R. Cripe, G. Fodor, J. Hoffmeister, V. Lee, A. McAlexander, R. Mollenhauer, S. Shaw, D. Rajchel, B. Williston, and W. Zak for their assistance in the field and the laboratory. Thanks also to M. Ebener and D. Tagerson for assistance with the Lake Superior samples, C. Krause on Lake Erie, and L. Barbeau, D. Frazier, D. Hickey, K. King, T. King, P. Jensen, P. Peeters, B. Peterson, and J. Peterson for Lake Whitefish collections in Lake Michigan. We thank D. Clapp, E. Grove, K. Porath, T. Pattison, D. Wiefreich, J. Witzig, and two anonymous reviewers for their thoughtful comments that helped improve this article. Funding for this project was provided by the Great Lakes Fishery Trust, project number 2004.570, Department of Forestry and Natural Resources at Purdue University, Department of Fisheries and Oceans (to M. Koops) and Environment Canada (to M. Arts).

REFERENCES

Akmon, D., A. Zimmerman, M. Daniels, and M. Hedstrom. 2011. The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservation needs. Archival Science 11:1–20.

Baker, K. S., B. J. Benson, D. L. Henshaw, D. Blodgett, J. H. Porter, and S. G. Stafford. 2000. Evolution of a multisite network information system: the LTER information management paradigm. Bioscience 50(11):963–978.

Baker, K. S., and G. C. Bowker. 2007. Information ecology: open system environment for data, memories and knowing. Journal of Intelligent Information Systems 29:127–144.

Baker, K. S., S. J. Jackson, and J. R. Wanetick. 2005. Strategies supporting heterogeneous data and interdisciplinary collaboration: towards an ocean informatics environment. Pages 1–10 *in* Proceedings of the 38th Hawaii International Conference on System Sciences. (HICSS) 2005, 3–6 January, Big Island, Hawaii. IEEE, New Brunswick, New Jersey.

Baker, K. S., and K. I. Stocks. 2007. Building environmental information systems: myths and interdisciplinary lessons. Pages 253b–253b *in* System Sciences, 2007. Hawaii International Conference on System Sciences (HICSS) 2007. IEEE, New Brunswick, New Jersey.

Barnas, K., and S. L. Katz. 2010. The challenges of tracking habitat at various spatial scales. Fisheries 35(5):232–241.

Beard, D. T., D. Austen, S. J. Brady, M. E. Costello, H. G. Drewes, C. H. Young-Dubovsky, C. H. Flather, C. L. Gengerke, A. J. Loftus, and M. J. Mac. 1998. The multi-state aquatic resources information system. Fisheries 18(2):14–18.

Birnholtz, J. P., and M. J. Bietz. 2003. Data at work: supporting sharing in science and engineering. *in* Proceedings of the 2003 International ACM SIGGROUP Conference on Supporting Group Work, Sanibel Island, Florida. 9(12):339–348.

Blukacz, E. A., M. A. Koops, T. M. Sutton, M. T. Arts, J. D. Fitzsimmons, A. M. Muir, R. M. Claramunt, T. B. Johnson, R. E. Kinnunen, M. P. Ebner, C. Suski, and G. Burness. 2010. Linking Lake Whitefish (*Coregonus clupeaformis*) condition with male gamete quality and quantity. Journal of Great Lakes Research 36(1):78–83.

Blumenthal, D., E. G. Campbell, M. S. Anderson, N. Causino, and K. S. Seashore. 1997. Withholding research results in academic life science: evidence from a national survey. Journal of American Medical Association 277(15):1224–1228.

Blumenthal, D., E. G. Campbell, M. Gokhale, R. Yucel, B. Clarridge, S. Hilgartner, and N. A. Holtzman. 2006. Data withholding in genetics and other life sciences: prevalences and predictors. Academic Medicine 81(2):137–145.

Borer, E. T., E. W. Seabloom, M. B. Jones, and M. Schildhauer. 2009. Some simple guidelines for effective data management. Bulletin of the Ecological Society of America 90(2):205–214.

Borgman, C. L. 2010. Research data: who will share what, with whom, when, and why? Paper presented at the China–North America Library Conference, Beijing, China.

Borgman, C. L., J. C. Wallis, and N. Enyedy. 2006. Building digital libraries for scientific data: an exploratory study of data practices in habitat ecology. Paper presented at the 10th European Conference on Research and Advanced Technology for Digital Libraries, September 17–22, Alicante, Spain.

Borgman, C. L., J. C. Wallis, and N. Enyedy. 2007. Little science confronts the data deluge: habitat ecology, embedded sensor networks and digital libraries. International Journal of Digital Libraries 7(1–2):17–30.

Brenden, T. O., M. P. Ebner, and T. M. Sutton. 2010. Special Issue on Assessing the Health of Lake Whitefish Populations in the Laurentian Great Lakes. Journal of Great Lakes Research 36(1):1–142.

Brunt, J. W. 1994. Research data management in ecology: a practical approach for long-term projects. Pages 272–275 *in* J. C. French and H. Hinterberger, editors. Seventh International Working Conference Scientific and Statistical Database Management. IEEE Computer Society Press, Washington, D.C.

Brunt, J. W., and W. K. Michener. 2009. The resource discovery initiative for field stations: enhancing data management at North American biological field stations. Bioscience 59(6):482–487.

Campbell, E. G., B. R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgartner, N. A. Holtzman, and D. Blumenthal. 2002. Data withholding in academic genetics: evidence from a national survey. Journal of the American Medical Association 287(4):473–479.

Carlson, S., and B. Anderson. 2007. What are data? The many kinds of data and their implications for re-use. Journal of Computer-Mediated Communication 12(2):635–650.

Carpenter, S. R., H. A. Mooney, J. Agard, D. Capistrano, R. S. DeFries, S. Díaz, T. Dietz, A. K. Duraiappah, A. Oteng-Yeboah, H. M. Pereira, C. Perrings, W. V. Reid, J. Sarukhan, R. J. Scholes, and A. Whyte. 2009. Science for managing ecosystem services: beyond the Millennium Ecosystem Assessment. Proceedings from the National Academy of Sciences 106(5):1305–1312.

Claramunt, R. M., A. M. Muir, J. Johnson, and T. M. Sutton. 2010a. Spatio-temporal trends in the food habits of age-0 Lake Whitefish. Journal of Great Lakes Research 36(1):66–72.

Claramunt, R. M., A. M. Muir, T. M. Sutton, P. J. Peeters, M. P. Ebner, J. D. Fizsimons, and M. A. Koops. 2010b. Measures of larval Lake Whitefish length and abundance as early predictors of year-class strength in Lake Michigan. Journal of Great Lakes Research 36(1):84–91.

Costello, M. J. 2009. Motivating online publication of data. Bioscience 59(5):418–427.

Cragin, M. H., C. L. Palmer, J. R. Carlson, and M. Witt. 2010. Data sharing, small science and institutional repositories. Philosophical Transactions of the Royal Society 368:4023–4038.

Cragin, M. H., C. L. Palmer, and T. C. Chao. 2008. Relating data practices, types, and curation functions: an empirically derived framework. Proceedings of the American Society for Information Science and Technology 47(1):1–2.

Fagan, K. A., M. A. Koops, M. T. Arts, and M. Power. 2011. Assessing the utility of C:N ratios for predicting lipid content in fishes. Canadian Journal of Fisheries and Aquatic Sciences 68:378–385.

Fagan, K. A., M. A. Koops, M. T. Arts, T. M. Sutton, and M. Power. 2012. Lake Whitefish feeding habits and condition in Lake Michigan. Advances in Limnology 63:399–415.

Federal Geographic Data Committee. 1998. FGDC-STD-001-1998. Content standard for digital geospatial metadata (revised June 1998). Federal Geographic Data Committee, Washington, D.C. Available: http://gdc@usgs.gov. (January, 2013).

Fishbein, E. A. 1991. Ownership of research data. Journal of Academic Medicine 66(3):129–133.

Freedom of Information Act (FOIA). 5 United States Code, Section 552. Department of Defense Publication 5400.7, Part 286 of Chapter 32. Code of Federal Regulations.

Frimpong, E. A., and P. L. Angermeir. 2009. Fish traits: a database of ecological and life history traits of freshwater fishes of the United States. Fisheries 34(10):487–495.

Guttmacher, A. E., E. G. Nabel, and F. S. Collins. 2009. Why data-sharing policies matter. Proceedings of the National Academy of Sciences 106(40):168–194.

Haas, M. R. 2011. To share or not to share? Professional norms, reference groups, and information withholding among life scientists. Organization Science 21(4):873–891.

Haerder T., and A. Reuter. 1983. Principles of transaction-oriented database recovery. ACM Computing Surveys 15(4):287–317.

Hale, S. S., A. H. Miglarese, M. P. Bradley, T. J. Belton, L. D. Cooper, M. T. Frame, C. A. Friel, L. M. Harwell, R. E. King, W. K. Michner, D. T. Nicolson, and B. G. Peterjohn. 2003. Managing troubled data: coastal data partnerships smooth data integration. Environmental Monitoring and Assessment 81:133–148.

Heidorn, P. B. 2008. Shedding light on the dark data in the long tail of science. Library Trends 57:280–299.

Hernandez, M. J. 2003. Database design for mere mortals: a hand's on guide to relational database design. Addison-Wesley, New York.

Hook, L. A., S. K. S. Vannan, T. W. Beaty, R. B. Cook, and B. E. Wilson. 2010. Best practices for preparing environmental datasets to share and archive. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee.

International Organization for Standardization. 2011. Certification for high standards of quality management, ISO 9001:2011. International Organization for Standardization, Geneva, Switzerland.

Katz, S. L., K. Barnas, R. Hicks, J. Cowan, and R. Jenkinson. 2007. Freshwater habitat restoration actions in the Pacific Northwest: a decade's investment in habitat improvement. Restoration Ecology 15(3):494–505.

Kelling, S., W. Hochachka, D. Fink, M. Riedewald, and R. Caruana. 2009. Data-intensive science: a new paradigm for biodiversity studies. Bioscience 59(7):613–620.

Lélé, S., and R. B. Norgaard. 2005. Practicing interdisciplinarity. Bioscience 55(11):967–975.

Ling Liu, M. and T. Özsu, editors. 2009. Encyclopedia of database systems. Springer, New York.

Lord, P., A. Macdonald, L. Lyon, and D. Giarretta. 2004. From data deluge to data curation. Pages 371–357 *in* Proceedings of the UK e-Science All Hands Meeting.

(MARIS) The Multistate Aquatic Resource Information System. 2008. About MARIS. Available: http://www.marisdata.org. (January, 2013).

McDade, L. A., D. R. Maddison, R. Guralick, H. A. Piowar, M. L. Jameson, K. M. Helgen, P. S. Herendeen, A. Hill, and M. L. Vis. 2011. Biology needs a modern assessment system for professional productivity. Bioscience 61:619–625.

McDonald, L. L., R. Bilby, P. A. Bisson, C. C. Coutant, J. M. Epifanio, D. Goodman, S. Hanna, N. Huntly, E. Merrill, B. Riddell, W. Liss, E. J. Loudenslager, D. P Phillipp, W. Smoker, R. R. Whitney, and R. N. Williams. 2007. Research, monitoring, and evaluation of fish and wildlife restoration projects in the Columbia River Basin. Fisheries 32(12):582–590.

McLaughlin, R. L., L. M. Carl, T. Middel, M. Ross, D. L. G. Noakes, D. B. Hayes, and J. R. Baylis. 2001. Potentials and pitfalls of integrating data from diverse sources: lessons from a historical database for Great Lakes stream fisheries. Fisheries 26(7):14–23.

McLaughlin, R. L., M. L. Jones, N. E. Mandrak, and D. Stacey. 2010. FishMAP online: a web application supporting science-based decisions concerning fish movement and passage. Great Lakes Fishery Commission, Ann Arbor, Michigan.

Michener, W. K., S. Allard, A. Budden, R. B. Cook, K. Douglass, M. Frame, S. Kelling, R. Koskela, C. Tenopir, and D. A. Vieglai. 2012. Participatory design of DataONE—enabling cyberinfrastructure for the biological and environmental sciences. Ecological Informatics 11:5–15.

Michener, W. K., and M. B. Jones. 2011. Ecoinformatics: supporting ecology as a data intensive science. Trends in Ecology and Evolution 27(2):85–93.

Muir, A. M., T. M. Sutton, M. T. Arts, R. M. Claramunt, M. P. Ebner, J. D. Fitzsimons, T. B. Johnson, R. E. Kinnunen, M. A Koops, and M. M. Sepúlveda. 2010. Does condition of Lake Whitefish spawners affect physiological condition of juveniles? Journal of Great Lakes Research 36(1):92–99.

National Institutes of Health. 2003. Data sharing policy and implementation guidance. Available: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm. (January, 2013).

National Science Foundation. 2010. Dissemination and Sharing of

Research Results. Available: http://www.nsf.gov/bfa/dias/policy/dmp.jsp. (January, 2013).

Nelson, B. 2009. Data sharing: empty archives. Nature 461(7261):160–163.

Newman, H. B., M. H. Ellisman, and J. A. Orcutt. 2003. Data intensive e-science frontier research. Communications of the ACM 46(11):68–77.

(NOAA) National Oceanic and Atmospheric Administration Fisheries Service. 2010. Data and information management policy directive. Available: http://www.nmfs.noaa.gov/op/pds/documents/04/04-111.pdf. (January, 2013).

———. 2011. Information and data administration policy. Available: http://ias.pifsc.noaa.gov/lds/docs/DMPolicyDraft1008.doc. (January, 2013).

———. 2012. Data citation guidelines. Available: https://www.nosc.noaa.gov/EDMC/documents/edmcon/2012_breakout_sessions/Duerr-ESIP_Data_Citation_Guidelines.pdf. (January, 2013).

Parr, C. S. 2007. Open sourcing ecological data. Bioscience 57(4):309–310.

Porter, J. H., and K. W. Ramsey. 2002. Integrating ecological data: tools and techniques. Pages 396–401 *in* N. Callaos, J. Porter, and N. Rishe, editors. Proceedings of the 6th World Multi-Conference on Systematics, Cybernetics and Informatics, Orlando, Florida.

Postel, B. R., L. A. Shapiro, and J. C. Biesanz. 2002. On having one's data shared. Journal of Cognitive Neuroscience 14(6):838–840.

Pratt, P. J., and J. J. Adamski. 2007. Concepts of database management, 4th edition. Boston, Massachusetts.

R Development Core Team. 2011. R: a language environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available: http://www.R-project.org. (January, 2013).

(RIN) Research Information Network. 2008. To share or not to share: publication and quality assurance or research data outputs. Available: http://eprints.soton.ac.uk/266742/1/Published_report_-_main_-_final.pdf. (January, 2013).

Sutton, T. M., M. A., Koops, M. T. Arts, A. M. Muir, A. E. Blukacz, R. M. Claramunt, M. P. Ebener, J. D. Fitzsimons, T. B. Johnson, and R. E. Kinnunen. 2007. Project completion report: does adult body condition affect recruitment potential in Lake Whitefish (*Coregonus clupeaformis*)? Great Lakes Fishery Trust, Lansing, Michigan.

Tenopir, C., S. Allard, K. Douglas, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame. 2011. Data sharing by scientists: practices and perceptions. PLoS ONE 6(6):1–21.

Uhlir, P. 2010. Information gulags, intellectual straightjackets, and memory holes: three principles to guide the preservation of scientific data. Data Science Journal 7:ES1–ES5.

U.S. Fish and Wildlife Service. 2006. Department of Interior—departmental manual, part 378: data resource management. Available: http://www.fws.gov/policy/274fw1.html. (January, 2013).

Van den Eynden, V., L. Corti, M. Woollard, L. Bishop, and L. Horton. 2011. Managing and sharing data—best practice for researchers. UK Data Archive, University of Essex, Colchester, UK.

Vogeli, C., Y. Recai, E. Bendavid, L. M. Jones, M. S. Anderson, K. S. Louis, and E. G. Campbell. 2006. Data withholding and the next generation of scientists: results of a national survey. Academic Medicine 81(2):128–136.

Wake, M. H. 2008. Integrative biology: science for the 21st century. Bioscience 58(4):224–232.

Wallis, J. C., C. L. Borgman, M. Mayernik, and A. Pepe. 2008. Moving archival practices upstream: an exploration of the life cycle of ecological sensing data in collaborative field research. International Journal of Digital Curation 3(1):114–126.

Wang, L., D. Infante, P. Esselman, A. Cooper, D. Wu, W. W. Taylor, D. Beard, G. Whelan, and A. Ostroff. 2011. A hierarchical spatial framework and database for the National River Fish Habitat Condition Assessment. Fisheries 36(9):436–449.

Watson, R., and R. Y. Kura. 2006. Fishing gear associated with global marine catches: database development. Fisheries Research 79:97–102.

Whiteed, D. C., J. S. Kimball, J. A. Lucotch, N. K. Manmenee, H. Wu, S. D. Chilcote, and J. A. Stanford. 2012. A riverscape analysis tool developed to assist wild salmon conservation across the North Pacific Rim. Fisheries 37(7):305–314.

Whitlock, M. C. 2011. Data archiving in ecology and evolution: best practices. Trends in Ecology and Evolution 26(2):61–65.

Witt, M. 2009. Institutional repositories and research data curation in a distributed environment. Library Trends 57:191–201.

Witt, M., J. Carlson, and S. Brandt. 2009. Constructing data curation profiles. International Journal of Digital Curation 3(4):93–103.

From the Archives

Our great opponents in this have been the net-fishermen at the mouth of the river. Above that, every man wants a closed time; but, time says, "Everyone is going in, and I will go in too;" and they do, and catch all they can.

Mr. Eugene G. Blackford (1878): "Peculiar Features of the FishMarket", Transactions of the American Fisheries Society, 7:1, 77-87.