Data Management Guide for Public Participation in Scientific Research

DataONE Public Participation in Scientific Research Working Group

February, 2013

Andrea Wiggins, Rick Bonney, Eric Graham, Sandra Henderson, Steve Kelling, Gretchen LeBuhn, Richard Littauer, Kelly Lotts, William Michener, Greg Newman, Eric Russell, Robert Stevenson, & Jake Weltzin

Abstract

This guide provides a step-by-step introduction to the data management life cycle. It includes examples from citizen science projects and links to best practices and tools to help project organizers optimize the quality, usability, and accessibility of their project data. Many benefits result from following best practices to manage the data life cycle for a citizen science project. Organizers spend less time on data management and more time on project goals by investing in data management plans and practices. Better data management makes data easier to find, use, analyze, share, and reuse. In the long term, following good data management practices means that citizen science data can contribute to decision-making, policies, and research beyond the project's original scope.

Introduction

Public participation in scientific research (PPSR), commonly known as citizen science, requires data stewardship like any scientific endeavor. Most scientific data management techniques apply to citizen science projects no differently from traditional research, but there are additional details to address due to the involvement of volunteers.

All citizen science projects must carefully balance trade-offs in designing protocols and systems to support participation. These choices affect data quality, utility for addressing scientific questions, and ease of participation for non-professionals. Citizen science projects can involve large collaborative teams that overwhelm the size of traditional research projects and require careful planning due to the number of individuals interacting with the project data and the wide variability of participants' training and backgrounds.

This guide was designed to meet three primary goals:

1. Introduce the data life cycle as it pertains to citizen science projects.
2. Describe and illustrate best practices related to each step of the data life cycle and provide relevant recommendations and resources.
3. Identify key opportunities and challenges in data management processes and tools.

This guide provides an introduction to data management to assist project organizers as they ensure that data collected from citizen science projects are well designed, easily acquired, readily interpreted, and effectively employed for scientific purposes. It discusses data management step-by-step, explaining the key processes, offering examples from ongoing projects, and highlighting concerns specific to citizen science projects. Each section also includes a box with links to learning modules and best practices documents from **DataONE**; other boxes offer additional resources.

The Data Life Cycle

The data life cycle shown in Figure 1 demonstrates the various processes involved in data management (Michener & Jones, 2012; Strasser, Cook, Michener, Budden, & Koskela, 2011). The data life cycle is a part of every scientific research project, whether the decisions around each aspect are made by default or with strategic intent.

Although represented as a logical cycle, these steps can occur in any number of different sequences, with some steps occurring in tandem and some repeated more often than others.

Data Life Cycle Steps

1. *Plan*: Map out the processes and resources for the entire data life cycle. Start with the project goals (desired outputs, outcomes, and impacts) and work backwards to build a data management plan, supporting data policies, and sustainability plans.

2. *Collect*: Determine the best way to get information from participants into a usable data file. The end result of this

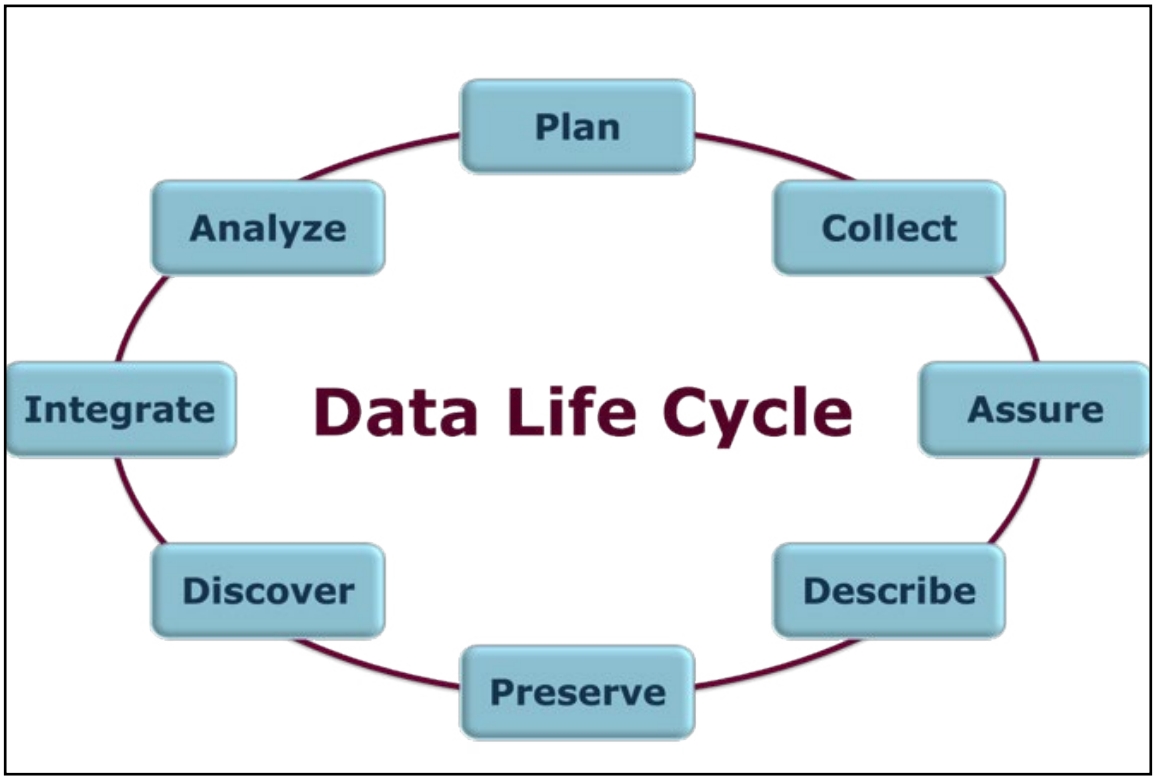

Figure 1. The data life cycle model, courtesy of DataONE.

decision process is a data model that describes the way the data are structured.

3. *Assure*: Employ quality assurance and quality control procedures that enhance the quality of data (e.g., training participants, routine instrument calibration) and identify potential errors and techniques to address them.

4. *Describe*: Document data by describing the why, who, what, when, where, and how of the data. Metadata, or data about data, are key to data sharing and reuse, and many tools such as standards and software are available to help describe data.

5. *Preserve*: Plan to preserve data in the short term to minimize potential losses (e.g., via accidents), and in the long term so that project stakeholders and others can access, interpret, and use the data in the future. Decide what data to preserve, where to preserve it, and what documentation needs to accompany the data.

6. *Discover*: Identify complementary data sets that can add value to project data. Strategies to help ensure the data have maximum impact include registering the project on a project directory site, depositing data in an open repository, and adding data descriptions to metadata clearinghouses.

7. *Integrate*: Combine citizen science project data with other sources of data to enable new analyses and investigations. Successful data integration depends on employing good data management practices throughout the data life cycle.

8. *Analyze*: Use data from a citizen science project for analyses that meet project goals for professional researchers, participants, and other stakeholders. Many software tools are available to support data exploration, analysis, and visualization.

In the remainder of this guide, each stage of the data life cycle is discussed as it relates to citizen science, with links to further resources throughout and examples drawn from experience. A brief, general, slightly more technical introduction is also available in DataONE's Primer on Data Management (Strasser, Cook, Michener, & Budden, 2012).

Step 1: Plan

An iterative process of project planning during which all aspects of data management are reviewed and decisions are made for documentation and implementation in later stages of data management.

Like most scientific endeavors, developing a data management plan is an important component of successfully implementing a new citizen science project or revising and updating an existing one. When compared to all the other requirements of undertaking a citizen science project, data management may seem like a small task, but in the long run, high quality data is a central metric of success for scientific outputs. The **Data Management Planning Tool** (https://dmp.cdlib.org/) provides an easy, guided checklist approach for making a plan.

Plan Learning Module

- Start with the learning module on **Data Management Planning**
- Then try out the online **Data Management Planning Tool**

Best Practices

- **Plan** data management early in the project
- **Define** roles and assign responsibilities for data management
- **Provide** budget information for data management plan
- **Recognize** stakeholders in data ownership
- **Define** expected data outcomes and types
- **Define** the data model
- **Create, manage, and document** data storage system
- **Document** the data organization strategy
- **Revisit** data management plan throughout project life cycle

Citizen science data have a few special requirements. Unlike professional data collection efforts where the data collectors and their abilities are known (e.g., a professional research team), special considerations apply to projects where participants may be anonymous or may not meet face-to-face, their familiarity or comfort with data may vary widely, or their data collection skills may be unknown.

Citizen science project organizers and coordinators often have a background in either science (interested in obtaining data for research) or education (interested in involving individuals via participatory learning). Coordinators sometimes have little experience with data management or the information systems that support a project. Involving a data manager who is familiar with these systems is ideal, when feasible, and data managers should be involved in data management planning and review. More often, however, project coordinators serve as data managers. In either case, it is important to revisit data management plans periodically (e.g., annually) and make necessary changes as projects evolve and mature.

Citizen science data management plans should address the data life cycle in the context of a specific citizen science project, with specific goals. Making decisions about data management requires understanding the project's purpose and being familiar with the data and how they are produced. Each step in the data life cycle is discussed in more detail in the rest of this guide; initial and ongoing planning must examine every step in the data life cycle.

Additional Planning Considerations for Citizen Science Projects

Citizen science projects' data management plans need to consider the implications of both working with volunteers and the operational aspects of the project or program. For instance, specific policies often need to be developed and implemented. In addition, having a sustainability plan in place at the inception of a project can help ensure the development of a sound financial foundation so that the project can continue into the future.

Policies

Policies for data access and sharing usually start with provisions for appropriate protection of privacy, confidentiality, security, and intellectual property. For legal and liability policy concerns, start by consulting organizational or institutional guidelines. Many institutions have policies, regulations and procedures that govern data access and sharing. Seeking out examples of data-related policies from other citizen science projects can also serve as a guide to crafting project policies.

Depending on the institution, different regulations and policies may apply to citizen science projects, particularly with respect to the involvement of minors. For example, participant surveys often require **Institutional Review Board** (IRB) Human Subjects research approval; a more comprehensive review may be required if the age of the contributors is unknown. Depending on the sources of project funding, field studies of wildlife in the US may also require approval from the **Institutional Animal Care and Use Committee** (IACUC) to ensure ethical treatment of animals. For some projects, liability considerations may require legal paperwork. Data security must also be addressed, which comes to the forefront when collecting or using sensitive data (e.g., identifiable information about people, threatened or endangered species, private land).

Project organizers also need to consider developing policies for data sharing and data use. Such policies can include elements such as limiting use by commercial entities, data access specifications, and forms of acceptable citation and acknowledgment (e.g., attribution). Terms of Use statements can address many data management policies in a single document, including any restrictions on access and use. For example, the **Terms of Use for *Nature's Notebook*** were reviewed and approved by the Office of Management and Budget as part of the **Paperwork Reduction Act** (which applies to most citizen science projects organized by federal organizations), providing a model that could be adapted for use by other federal organizations. These terms are more comprehensive than is needed for many citizen science projects, but offer an excellent example of a thorough and carefully crafted set of policies.

Sustainability

Data management is becoming more accessible and affordable, but for many citizen science projects, funding is not consistent or predictable. The costs of data management vary by participation levels and project design. Data management, however, is an essential part of the scientific work that supports the project. Estimated costs for data management should be included in project proposals and annual budgets. In estimating data management costs for implementation and ongoing management, consider:

- What types of personnel will be required to carry out this data management plan? Will staff training be required?
- What types of hardware, software, or other computational resources will be needed, such as domain names, web hosting, or remote backup services?
- What other expenses might be necessary, such as costs associated with depositing data in a data center or institutional repository for long-term preservation, or publication fees for open access journal publications?

To answer these questions, explore the availability of institutional resources. Some institutions have created data management plan templates and data centers; in such cases, associated personnel may be able to provide budgetary guidance as well as tools for planning the project. Consulting with other citizen science project organizers can also help identify useful strategies for sustainable project data management.

What is the difference between confidentiality and privacy?

Privacy protects access to an individual, while confidentiality protects access to information about an individual. In the context of data management, confidentiality can be thought of as information privacy.

Locking doors and drawing curtains are actions to control other's access to ourselves physically - part of what we consider privacy. Anonymizing data and reporting in aggregate are moves to keep information about individuals confidential. Maintaining data confidentiality helps maintain personal privacy because some data could be traced back to the individual, so if data are not confidential, personal privacy can be compromised.

Step 2: Collect

Determine how observations are made (e.g., by human observer, sensor, or instrument) and how the resulting data are organized and recorded.

The goal of planning data collection in citizen science is to determine the optimal mechanisms to acquire and organize observational data that are collected by the public. The end result of the decision process is a **data model** that clearly describes what the data are, their formats, and how they are to be organized and processed.

The first step in developing a data model for a citizen science project is identifying what information to collect. These details are sometimes referred to as *data components*. For

example, in **Project Budburst**, there is information about each participant (name, city, ZIP code, email address), each geographic location (latitude and longitude), and each observation (plant, date, and phenophase). **The Great Sunflower Project** collects information about each participant (name, street, city, state, postal code and country), garden location (latitude and longitude), and each observation (plant, number of flowers observed, number of bees seen, number of bumble bees, and number of honey bees).

These decisions should be influenced by anticipated uses of the data (requirements). Think broadly, as data may be most valuable in the future for unanticipated uses. For example, in Project Budburst the data are used both to answer scientific questions (shifts in phenology and changes in distributions) and to track participation goals (number of users, timing of use, which plants were tracked, geographic location of users).

Data from the Great Sunflower Project are used to answer a variety of research questions across disciplines including:

- Ecology:
  - What is the average bee visitation rate per hour and how does that change across space and time?
  - How important are native and honeybees to pollinator service and how does that change across space and time?
  - In regions where honeybees have declined, are there decreases in the rate of pollination or are native bees filling in for the loss?

- Social science:
  - What factors influence recruitment and retention of participants?
  - Does participation in citizen science lead to behavioral changes that indicate changes in science identity?

- Computer science research:
  - How is technology used to support data quality in citizen science?
  - What methods from computer vision, algorithm development, and advanced statistical modeling can be used to learn more from these data?

Considering the ways others might use the project data may lead to including some new data components.

After identifying data components, map out the relationships or associations between the data components; it can help to draw a diagram. This information will help identify the appropriate technology for storing the data in the technology assessment phase of data collection planning. Different technologies, from paper data sheets, to desktop computers, to the most sophisticated handheld devices, all have different strengths and weaknesses as tools for data entry. Examples of software used for managing data include spreadsheet programs like Open Office, Google Spreadsheets, and Microsoft Excel. If data are submitted online, they are usually stored in relational databases, such as Microsoft Access or MySQL. A formal representation of the data components and relationships, called an Entity Relationship Diagram, is often used to document relational database structures, such as those supporting **eBird**, shown in Figure 2. Be aware that technologies change frequently and can become obsolete; take care to choose technologies likely to be supported for the duration of planned project activities and beyond.

Collect Learning Module

- **Data Entry and Manipulation**

Best Practices

- **Use** consistent codes
- **Store** data with appropriate date and time formats
- **Spatial** location formats
- **Units** of measurement
- **Define** parameters
- **Identify** estimated data
- **Identify** missing values and define missing value codes
- **Maintain** consistent data typing
- **Allow** the community to annotate the data
- **Separate** data values from annotations
- **Plan** for effective multimedia data

Other software can help manage participant communities, such as free, open source content management systems like Drupal, Joomla, Wordpress, and Plone. If planning to implement mobile data entry, it is ideal to accommodate the current range of technologies participants may be using – both devices and operating systems – so as not to exclude people based on technology accessibility. Offering a variety of ways to participate can help citizen science projects be more inclusive, e.g., accepting data on paper data sheets as well as through a smartphone app. An important aspect of project sustainability, however, is ensuring the maintenance of these technologies in the face of minimal funding.

As a project moves forward, organizers should also periodically revisit these steps to refine the project's data model. This is particularly important when modifying methods for data collection and changing tools to meet the project's evolving needs.

Step 3: Assure

Quality control and quality assurance procedures minimize introduction of errors, and identify and treat erroneous data.

Ensuring data quality requires knowing the criteria that data must meet for project goals and/or scientific standards: data quality is determined by fitness for its intended use. A good way to identify appropriate data quality criteria is to work through desired analyses and visualizations *before* any data are collected. Using small dummy data sets created to include known and potential problems (e.g., missing values,

suspicious time/date values, unlikely locations, etc.) is very helpful for identifying ways that data quality issues can be detected, corrected, and prevented.

There are several approaches citizen science projects can use during project development to assure quality data and make the data more useful for future research.

Assure Learning Module

- **Data Quality Control and Assurance**

Best Practices

- **Develop** quality assurance and quality control plan
- **Communicate** data quality
- **Mark** data quality control flags
- **Identify** values that are estimated
- **Identify** missing values and define missing value codes
- **Ensure** basic quality control
- **Double check** data entry

These approaches and mechanisms can be applied before, during, and after data collection. Strategies used *before and during data collection* are typically referred to as Quality Assurance (QA) and those applied *after data collection* are called Quality Control (QC).

The quality of data depends on many factors, so these procedures cannot completely guarantee usable data. However, well-documented QA/QC improves the likelihood that the data can be used and reused. It is critical to document QA/QC processes and any changes to them in as much detail as possible, as this benefits everyone who manages and uses the data. For an example, see **the QA/QC documentation for *Nature's Notebook*.**

Quality assurance refers to the processes used to ensure that the best possible data will be collected. In citizen science, QA is strongly linked to the design of participation tasks, participant training, and supporting technologies. Data entry interfaces are a particularly important tool in promoting quality assurance because they can be designed to help observers provide the most accurate data possible, minimize mistakes, and reduce missing data.

For example, ensuring that participants enter geospatial data accurately can be challenging, so using maps that allow participants to drop pins to indicate observation locations can substantially improve data quality. In the Great Sunflower Project, participants can directly enter

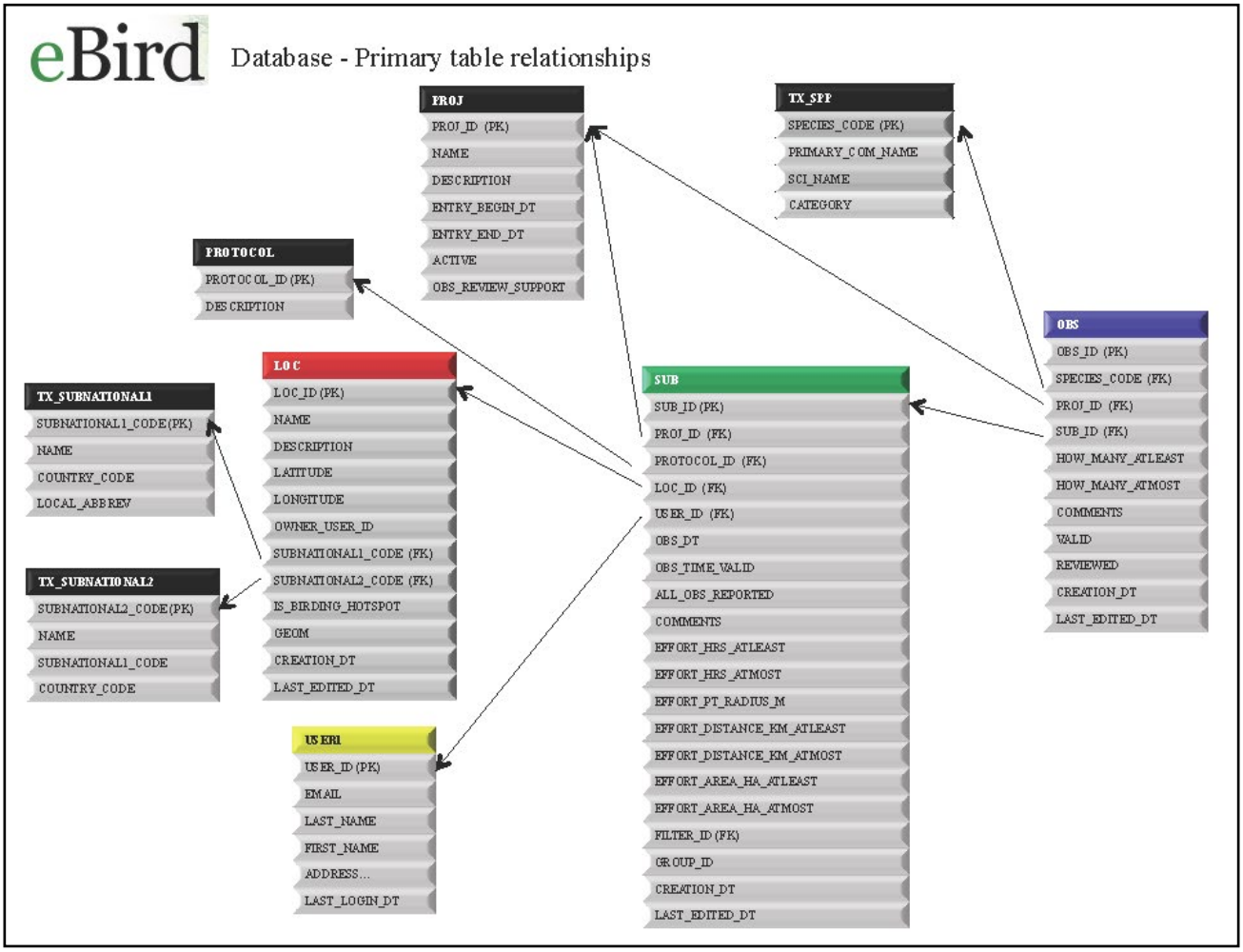

Figure 2. Entity relationship diagram, used with permission from Tom Fredericks and eBird.

GPS locations but this information is not always accurate. Including street addresses and using Google Maps for location selection on the project's online garden description form helps ensure that the correct location is recorded.

Quality control is a set of processes to evaluate the quality of the data after they are collected. This involves "data cleaning" and making decisions about issues like how to handle missing data and estimated values. QA/QC processes vary substantially in their cost and effectiveness; QC is considered more difficult and resource-intensive than QA. It is easier to prevent than repair problems, and much cheaper in the long run.

Questionable data are often "flagged" as potentially invalid. This is an important technique because the data displayed to the public, to participants, and to project organizers and researchers may differ. For example, contributors may wish to see all of their data, but flagging a data point allows researchers to include or exclude the data for analysis.

Because citizen science projects have potential to become very large, it is also important to plan for scalability. For example, it can be difficult or impossible to have professionals perform expert review of data for a project that expands to involve thousands of participants. Instead, developing a volunteer expert review network, like those used by eBird and **Butterflies and Moths of North America** (BAMONA) is a more scalable strategy (Wiggins, 2013). In projects dealing with visual processing tasks like transcription and image classification, having data entered multiple times by different volunteers is feasible, easy to implement, and a robust solution for addressing data quality. Many projects also collect paper data sheets in addition to online data entry in case questions arise about data entry accuracy, although maintaining paper records introduces additional complexity and costs.

During planning, consider not only which data quality mechanisms are appropriate for different life stages of the project, but also the associated costs. The resources needed for managing a reviewer network or developing a system with automated data checking processes are different, and new costs are introduced when transitioning to a new QA/QC strategy.

QA and QC
(Wiggins, Newman, Stevenson, & Crowston, 2011)

Sample QA Processes

- Participant training
- "Beta" testing protocols
- Controlled values and formats for data entry
- Uniform measurement tools

Sample QC Processes

- Expert review
- Automatically flagging unusual records for review
- Using photos for ad hoc validation

Step 4: Describe

Information about data (metadata) are recorded so that others may discover, acquire, interpret, and effectively use data.

Describing or documenting data is essential to future data use and reuse. The information that describes data is called metadata. It is useful to think of metadata in terms of the why, who, what, when, where, and how of the project (Michener et al. 1997). Although the structure of data documentation may take another form, answering these basic questions will help ensure a thorough description.

- Why were the data collected (i.e., project goals and intended outcomes)?
- Who collected the data?
- What does the data include?
- When were the data collected?
- Where were the data collected?
- How were the data collected and how was data quality ensured?

Answers to these questions provide important context for the data, particularly as time passes. The metadata can be used both by humans and software programs to help discover, integrate, and analyze data. There are a number of metadata standards that make this information more useful; for example, **Darwin Core** is one of the more common metadata standards used in ecological sciences.

Producing detailed metadata also protects the project's investment in data collection. Changes in technology, personnel, or simply the effects of time on human memory can cause the loss of information about data over time. Keeping detailed records about data will protect against loss of important details, ensuring that the data remain usable.

Why

Understanding data requires understanding the scientific context of its creation. These details should include the scientific questions or purposes for data collection, reasons for involving volunteer contributors, and a brief overview of the data set.

Who

Metadata should describe both who collected the data and who sponsored the collection of the data. It should also specify whom to contact with questions about the data and how to cite the data set to ensure that everyone involved gets proper credit for the work. Data citation provides evidence of the value of citizen science projects, which may be useful for future funding requests.

What

To fully describe the data, several categories of details are needed. They are listed below along with examples of what to include in each category; consult the best practices listed above for more information. Several of these points overlap with those from the Assure step, so documented data quality procedures will make this step easier.

1. *The digital context* – names of files, file formats, modification dates, example files, related data sets, and data processing procedures

2. *Parameter/variable details* – measurement units, data formats, precision, and accuracy

3. *Information about data* – taxonomies, coding, QA/QC procedures, known problems like sampling bias

4. *Contents of data files* – definitions of parameters and explanation of their formats, quality review notes (e.g., flags on untrusted data points), missing values

5. *Additional taxonomic information* – using standard taxonomies when possible

6. *Organization of the data* – relationships among data entities, files, directories and/or database tables; when possible, including a database structure diagram (see Figure 2) is ideal.

When

The temporal extent and resolution of the data should be as specific as possible, noting year, month, day, and time as appropriate.

Three facets of temporality should be included in data descriptions, along with the data format.

1. *Temporal boundaries* – the entire time range for observations included in the data set, e.g., eBird data include historical data as early as 1900, with ongoing data collection through the current day.

2. *Temporal extent of data collection* – the entire time range for data acquisition, e.g., eBird data were acquired starting in 2002 through the current day, for both contemporary and historical data.

3. *Temporal resolution* – the frequency at which data are collected or acquired, e.g., eBird data are collected daily and even hourly, but not necessarily at regular intervals.

Where

Like temporal aspects of the data, location information is important for most citizen science projects, and formats can be very important for data use and reuse. The three characteristics to document for geographic information are the spatial extent and resolution of the data, and data formats.

1. *Spatial extent*: boundaries of the data set; at minimum, north, south, east, west boundaries. Where available, **polygons** can be a good way to document these details. Projects with a global contributor base may need to describe spatial extent textually, as it would be clearer to note that the data include no observations from the McDonald Islands and Equatorial Guinea, for example, instead of listing every other country where data were collected.

2. *Spatial resolution*: the specificity of spacing for location data. For example, data points at a meter versus a kilometer square area may be used differently in analysis. This information is not often recorded when participants choose and report locations, but if a mapping tool is used for selecting observation locations, recording the zoom level for each data point can help. For most citizen science projects, this information will need to be described textually due to volunteers' autonomy in selecting locations.

3. *Spatial data formats*: describing spatial data formats means specifying a coordinate system from several options for formatting spatial data. If there is a standard format for the most closely related research field, it should be the first choice. Regardless of format, use the most specific values possible and a single coordinate system for the entire data set, as mixing coordinate systems is problematic for data use.

Describe Learning Modules

- **Metadata**
- **How to Write Good Quality Metadata**

Best Practices

- **Describe** overall organization of data
- **Identify** and use relevant metadata standards
- **Describe** contents of data files
- **Describe** derived data products
- **Document** steps in data processing
- **Describe** measurement techniques
- **Describe** date and time formats
- **Assign** descriptive file names
- **Document** taxonomic information
- **Describe** temporal extent and resolution of data set
- **Store** data with appropriate precision

Metadata Tools

- **Metavist**
- **Mercury Metadata Editor**
- **Morpho**
- **Numerous other options**

How

How were the data created? These details are key to data reuse and interpretation. At a minimum, data collection protocols, measurement techniques, and QA/QC methods need to be included in documentation for all data sets. Including relevant diagrams, schematics, and copies of data sheets and/or screenshots of electronic data entry forms is also helpful. If participants may be altering the protocol in a specific way, such as collecting data outside of a specified transect, it is best to include that information as well as known ways to detect or correct it in the data.

Data documentation also needs to include details such as measurement instruments, like the standardized rain gauges used by the **Community Collaboratory for Rain, Hail and Snow** (CoCoRaHS) observers. This guideline also applies to GPS devices, digital cameras, and smartphones used to capture images and sound. Unlike typical remote sensing and conventional scientific data collection, in most cases participants will be using different tools—whatever they already own—which affects the data integration and analysis. For image data, digital photos can be submitted with **EXIF** data about the exposure that the camera generates automatically (e.g., time, date, camera make and model). Smartphone applications can also automatically report the operating system and app versions. Another option is to ask volunteers to provide this information directly on a one-time basis, e.g., when signing up to participate. Including multimedia in a scientific data set brings new challenges both for analysis and data management, so **plan carefully for effective multimedia management.**

Most data collected or generated in citizen science projects are considered "derived" because they have been processed to some degree, if only to amend known errors, regularize measurements, and/or correct spelling errors. Documenting this kind of data processing usually involves describing the primary input ("raw" data, as it is received, prior to any manipulation) and the "cleaned" derived data, along with the rationale, assumptions, steps for data processing, and how potential anomalous values were identified and treated.

Depending on the nature of the data, additional details may be needed, particularly if data processing techniques changed over time. Documentation of data processing and analysis enables tracing the use of data sets (which supports attribution) and identifying effects of errors in the original data sets or derived data sets, all of which can benefit the project as a whole.

Special Considerations: Data About Participants

A key question for citizen science projects is how to describe who collected the data when hundreds or thousands of people contributed. In addition, privacy concerns for data about volunteers are a very important consideration, especially when participants provide addresses or other personally identifiable information. It is impractical and generally inadvisable to include a list of all contributors for the purpose of assigning credit (unless it is part of the participation agreement).

For citizen science projects operated or supported by federal agencies and organizations, certain policies must be considered, and certain permissions may be required (e.g., **Privacy Act of 1974**, Paperwork Reduction Act). Federal employees must contact the appropriate agency's Privacy Officer and Information Collections Clearance Officer (or similar) for more information.

When describing what the data set contains, data about volunteers can form an ancillary data set for interpretation and analysis of the main data set. Indicating which individuals made specific observations can be useful in evaluating data quality, but privacy is again a matter of substantial concern. Making the identity of individuals anonymous may or may not provide adequate protection of participant privacy depending on the data structure and many other details.

These issues further extend to geospatial data that can be resolved to observers' homes or neighborhoods. Again, there are no specific rules to follow, but use good sense and consider whether it is appropriate to use strategies like reducing spatial resolution of geographic data as a compromise to protect participant privacy. The US Department of Health and Human Services' **Office for Human Research Protections** is a good starting place for learning about these issues. IRBs are also a good resource for guidance on procedures to anonymize personally identifiable information, as it is a common criterion for human subjects research regardless of whether the data are shared. For projects that are not subject to an institutional IRB, there are accredited independent companies who provide this service (e.g., **IRB Services**, for US and Canadian researchers).

Step 5: Preserve

Preservation involves submitting data to an appropriate long-term archive, such as a data center or repository.

Preserving data isn't just a step that should happen after the data collection is completed. It is an important, ongoing data management task for all projects, including those with indefinite end dates. Further, data preservation occurs on both short-term and long-term scales, and each involves different approaches and decisions.

Short-term storage is the most familiar form of data preservation, and refers to backup files that are created manually or with an automated storage system like an external hard drive. Backups are copies of the original file as a snapshot of the data in time; they are required for restoration if the file is corrupted, lost, irreversibly altered, or destroyed. These backups should be periodically tested to verify their integrity. In addition to keeping backups of data that have been modified for QA/QC purposes, it is important to keep a copy of the original unprocessed data, which is easily accomplished with short-term backups.

Both software-assisted and manual backups can take advantage of online data storage to create remote backups as well as local copies. This ensures that if facilities were destroyed, the data would still be retrievable. Additional options include version control (revision control, source control) systems. They permit changes on the same data by multiple users, storing all changes, and at any point in time,

Preserve Learning Module

- **Data Protection and Backups**

Best Practices

- **Document** and store data using stable file formats
- **Ensure** integrity and accessibility of data backups
- **Backup** project data
- **Create** and document a data backup policy
- **Identify** suitable data repositories
- **Ensure** reliability of storage media
- **Decide** what data to preserve
- **Identify** data with long-term value
- **Ensure** flexible data services for virtual data sets
- **Preserve** information: keep raw data raw
- **Provide** version information for use and discovery
- **Create**, manage, and document the data storage system
- **Identify** and manage sensitive data

prior versions of the data can be restored.

Long-term storage generally has different requirements. An archival data set is a set of preserved records, typically a historical snapshot of data that are no longer actively changing. An archival data set preserves data for future needs, so the data need to be retrievable in a stable form. The ideal situation is data archiving that is freely accessible and uses a broadly adopted data format.

Long-term monitoring and other ongoing projects also require long-term storage, but data from these projects must be handled slightly differently from data collected during projects of limited duration. Archiving a quarterly or annual data export, either of the full data set or changed records and files since the last export, is an approach that requires very little effort from the depositor once an initial backup plan is in place.

Long-term storage depends on enduring technology infrastructure that is usually beyond the scope of most organizations, so shared repositories are often the best solution. Identifying an appropriate repository for the project's data is a key early step, as it may affect the collection and description of the data.

The primary types of data centers and repositories include:

- Institutional repositories, which may be specific to single academic and other research institutions, such as the **University of New Mexico's LoboVault** (for publications by UNM employees) and Oak Ridge National Research Laboratory's Distributed Active Archive Center (**DAAC**, for data from NASA's terrestrial ecology research projects).Institutional repositories are quickly growing in popularity and often focus solely on archiving materials produced by their organizational members.
- Topical or field-based repositories, such as the **Avian Knowledge Network** for bird observation data, include data from the eBird and Project FeederWatch projects. These can be further distinguished by geographic coverage; for example, the **Global Biodiversity Information Facility** accepts data from around the world.
- Journal publication-based repositories, such as **Dryad** store data used in peer-reviewed bioscience articles in over 150 journals. Such repositories are relatively new and are only available to authors of journal publications.
- Public sector data repositories, such as **data.gov** support federal and state agencies.

Despite the variety of repository types, there are relatively few that are appropriate for any given data set, and the level of appropriateness for citizen science data varies. In addition, some repositories "mirror" or duplicate data resources, aggregating data sets deposited in other (usually smaller) repositories, ensuring both data preservation and discovery.

To choose a data repository, look for matches between the project's focus (e.g., birds) and a related research field, or for those working in an organizational or institutional environment, ask whether an institutional repository exists. Data repositories also have requirements for file formats, metadata standards, and data documentation. This information can be used to plan ahead for long-term data storage and reduce future work to support data preservation. For example, the Avian Knowledge Network uses a metadata schema that is an extension of Darwin Core and therefore interoperable with other major repositories.

Additional points for comparison between repositories include:

- How concerns about privacy and sensitive data can be addressed;
- Access control options related to project data use, privacy, and sharing policies;
- Attribution policies and enforcement;
- Availability of information about data use that can demonstrate project impact (e.g., download figures);
- Whether the repository has a backup policy in place; and
- Whether there are costs associated with using the repository.

After identifying a suitable repository for long-term data storage, the next set of decisions relate to data selection. Not all data need to be preserved and/or shared, so identifying the data of greatest value can help streamline data documentation and deposit. On the other hand, there are many unanticipated uses for data, as discussed in the Collect section, which complicates these decisions.

Archiving multiple data sets may be the best choice depending on the project's data products. Usually data that

are deposited in repositories are derivative data products with some minimal processing, e.g., data cleaning, but in some cases raw data and/or analyzed data may be most appropriate to preserve. For ongoing projects, the best course is working with repository administrators to determine which data sets to preserve and how often to update them.

Ancillary data sets also play into data selection decisions. In citizen science projects, these are most likely data about volunteers, like contact information, performance metrics such as reliability or number of data points contributed, duration of engagement, and other personal information or anecdotes. These data are common in citizen science, and there are several reasons to preserve them along with primary data sets:

1. Ancillary data—both about volunteers and other related data (e.g., data loggers at observation sites)—can be used to contextualize the primary data set.

2. Other researchers may be able to use the data for research in other fields, even when the primary data are of relatively little use for project goals and interests.

3. Ancillary data may include QA/QC details that can both enhance trust of the data set and make it more useful for a wider variety of purposes.

The privacy issues previously mentioned come to the forefront when archiving ancillary data sets. At a minimum, decisions must be made about handling these concerns, e.g., by anonymizing identities or fuzzying observer locations. Access control for these data is particularly important because it may be appropriate to restrict which groups of people are able to view data and at what levels of detail. As with the other elements of data sets, ancillary data like these also require documentation. It is best to include the rationale for the way privacy issues were handled, plus relevant information about compliance with institutional policies and IRB approvals, where appropriate.

Step 6: Discover

Approaches for locating and acquiring potentially useful data and metadata, and for making data discoverable.

Data discovery has two faces; the first is finding existing data for analysis in conjunction with other sources of information. Generic web searches are usually not specific enough to find usable data, which are more readily found through one of the directories listed in Table 1 below. A new alternative for finding environmental data is the **ONEMercury** search engine.

The second aspect of data discovery is making information about the data available so others can discover and access it. Increasing the visibility of the project and its data improves its potential for broader use and benefit to scientific research, decision support, and policy development.

Master Directory	Description of Contents	Web Contents
Citizen Science Central	Directory of projects involving public participation in scientific research	citizenscience.org
Data.gov	Official U.S. government site providing public access to federal government data sets	data.gov
Databib Research Data Repositories	A collaborative, annotated bibliography of primary research data repositories	databib.org
Data Observation Network for Earth	A federation of data repositories that provides access to biological, ecological, environmental, and Earth science data as well as researcher tools	dataone.org
Global Biodiversity Information Facility	An interoperable network of biodiversity databases and information technology tools	gbif.org
Dryad	An international repository of data underlying peer-reviewed articles in the basic and applied bio-sciences	datadryad.org
Global Change Master Directory	A comprehensive directory of information about Earth science data, including the oceans, atmosphere, hydrosphere, and terrestrial environment	gcmd.nasa.gov
Knowledge Network for Biocomplexity	A national network that supports use of complex ecological data from distributed field stations, laboratories, research sites, and individual researchers	knb.ecoinformatics.org

Table 1. Master directories potentially relevant to citizen science projects.

Many citizen science project organizers are interested in sharing data, but there are a number of variables that influence how data are shared and with whom, such as data use policies and potential for malfeasance, misuse, or misinterpretation. If project leaders are not comfortable providing the data in an Open Access repository where anyone can download it from the Internet, there are other ways to increase the visibility of the data for potential reuse. All of the strategies discussed below can be used in isolation or in combination.

A natural place for people to search for information about citizen science data is at one of the sites that register citizen science projects, such as **citizenscience.org**. Listing the project in a directory will help both potential participants and researchers discover it. Directory listings link back to project websites, where project organizers can make it easy for others to find data products. These basic items are a minimum first step, but are generally inadequate for making data truly accessible, as most researchers are unlikely to begin a search for data by looking specifically for citizen science data.

***Discover* Learning Module**

- **Data Sharing**

Best Practices

- **Choose** and use standard terminology to enable discovery
- **Check** data and other outputs for print and web accessibility
- **Advertise** the data using datacasting tools

A second way to support data discovery is by depositing data in a repository, as discussed in the Preserve section. The contents of repositories are often summarized and indexed in Master Directories, which are specifically designed to help people find data (see Table 1). Repositories help promote data reuse by advertising data resources and enabling user communities to tag and annotate data in ways that further improve search results.

Another option is registering the data set with a "metadata clearinghouse," such as the **USA National Phenology Network's data registry system** for phenological data or the **Knowledge Network for Biocomplexity** (KNB). This approach offers some of the same advantages of depositing data in a repository but leaves data depositors in control of the data. Registering only metadata, so that interested researchers can contact data set authors directly to request the data, may be more appropriate for projects with sensitive data and for ancillary data sets with information about participants. The combination of depositing a primary data set in a repository along with descriptive information about an ancillary data set, both of which can also be registered with a metadata clearinghouse, may be a suitable solution to complex issues around sharing data related to volunteers.

Step 7: Integrate

Data from multiple sources are combined into a form that can be readily analyzed.

There are several scenarios for integrating citizen science data:

- Integrated data from multiple projects are needed to address complex issues. When project organizers invest effort to produce interoperable data that can be easily integrated, new opportunities arise for investigating complex issues or questions.
- Sparse data needs to be augmented by existing data to support analysis. Combining citizen science data with other data sources, such as loggers, professional observer records, historical records, and/or statistical model estimates can make a sparse citizen science data set more valuable for analysis.
- Additional data are needed to contextualize citizen science observations. When participants are not suitable sources for the data or do not provide data at a useful scale, or other reliable data sets are already available, integrating data can sometimes aid in verification and interpretation.

The final scenario is likely most common, with the data integration process following identification of data needed for analysis and the data discovery process. For example, The Great Sunflower Project integrated housing density data with bee observations to compare the pollinator service performance of rural, suburban, and urban habitats. Participants' categorizations of the urban-ness of their garden locations turned out to be a judgment that was too subjective to be useful in analysis.

Another example is the **eBird Reference Data Set** (ERD). The ERD is a carefully curated data product that integrates numerous geographical covariates with observations. Along with extensive data documentation, the ERD provides all the relevant information that most researchers need in one convenient package. Although it required substantial funding and technical expertise to create, the ERD is an example of the type of valuable multipurpose data products that citizen science projects have potential to generate.

Successful data integration—for the purposes of the project's stakeholders and for others using citizen science data—is highly dependent on good data documentation and careful description. When thorough data description is available, potential data users can more easily assess their ability to reuse data produced by others. For example, undescribed measurement techniques or units, undefined field labels, varying spatial scales, and conflicting data collection protocols are just a few of the potential challenges for data integration. Ensuring data interoperability is one of the main reasons to adhere to widely used standards for measurements, file and data formats, variables, and metadata. The actual processes of integrating data vary widely depending on the characteristics of the data sets that are merged.

For example, the map shown in Figure 3 includes data

points from 23 different citizen science projects. This example of collaborative data integration and visualization was enabled by use of standard geographic coordinates, freely available tools, a few hours of work, and willingness to share data.

Another way that citizen science projects can create interoperable, and increasingly high value data sets, is by using shared protocols. This practice is particularly advantageous for new citizen science projects, which can save substantial time (often several years) and funding by adopting an identical or slightly modified protocol from an existing project. Reuse or extension of existing protocols also opens up new avenues for collaboration and cooperation that can improve outcomes for all parties involved. Though standardized protocols are optimal, it is not always possible to adopt them prior to the start of a project, and existing protocols may not be appropriate for some projects' goals. With adequate data documentation, however, post hoc translation into a new format to enable data integration is often feasible.

***Integrate* Learning Module**

- **Analysis and Workflows**

Best Practices

- **Consider** the compatibility of the data you are integrating
- **Document** the integration of multiple datasets
- **Document** steps used in data processing
- **Understand** the geospatial parameters of multiple data sources

Step 8: Analysis

Create analyses and visualizations to identify patterns, test hypotheses, and illustrate findings.

This step concerns analysis of data collected or generated by citizen science participants, as opposed to data processing tasks in online citizen science projects. Planning analyses from the outset, prior to data collection, will help create a better data collection protocol.

Analysis should be guided by the project's goals and related data needs, and the expectations of intended audiences for project results (e.g., policy makers, scientific community). Plan to seek advice from experts when appropriate and develop interdisciplinary collaborations to address challenging questions or work with challenging data, such as statisticians who can develop models that accommodate incomplete data, computer scientists interested in data mining and artificial intelligence, social scientists with experience in qualitative text analysis (Kelling, 2012).

Current challenges in citizen science data analysis include needs for:

- Strategies for creating useful and usable data sets;
- A toolbox of techniques for analytically addressing common problems with the data;
- Guidelines for accommodating wide variability in participant interest and skills when they are involved in data analysis; and
- Affordable, flexible tools that can enable data exploration, visualization, and analysis, as well as easy

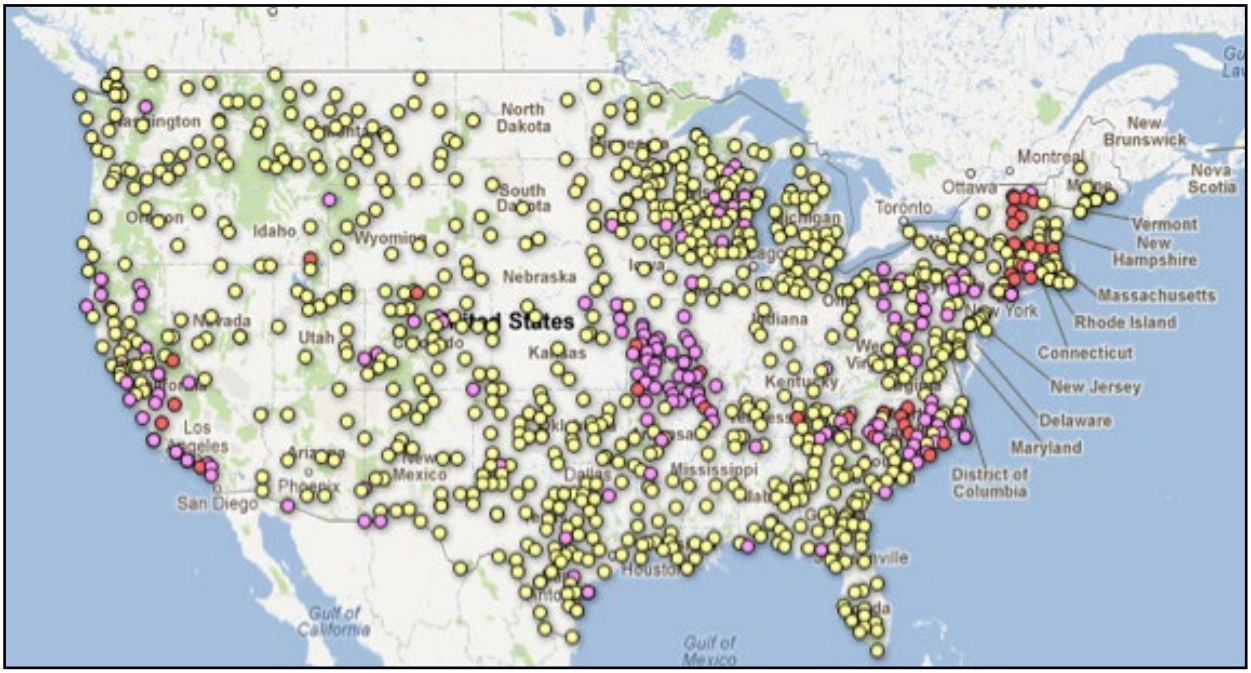

Figure 3. Observation locations for 23 citizen science projects, used with permission of Gretchen LeBuhn, Kelly Lotts, and Greg Newman.

participant access to these features.

Who participates in data analysis depends on research goals as well as project resources, interest, and capacity for engaging volunteers in data analysis. Each approach also has varying advantages and disadvantages. Where feasible, engaging both professionals and participants in analysis seems to offer the best results, as each group has unique knowledge of the data.

Notably, there are multiple levels of participant engagement in analysis processes. The simplest may be enabling access to data and means for communicating about it, e.g., with downloadable Excel files and website forums. Additional tools that offer more ways to explore the data, especially with data visualizations, can significantly increase the ability of participants to engage in analysis independently. Creating mechanisms for communication between participants and researchers about analysis requires more human attention but can be supported by relatively simple, free technologies like forums and blogs. In-person analysis workshops are another option for localized projects.

A variety of software tools can support data analysis and visualization. An increasing variety are freely available online, or at low cost. A few examples are listed in this guide, with the caveat that the fast growing array of tools and services means that any such list quickly becomes obsolete. Significant advances have also been made in software supporting the creation and management of complex scientific workflows. Scientific workflow management software serves to integrate, analyze, and visualize data as well as document the exact steps used in those processes (Hull et al., 2006; Ludäscher et al., 2006). Although these tools offer impressive power, scalability, flexibility, and the ability to retain thorough analysis records, they are used primarily for computationally intensive analyses.

Conclusion

Scientific data management is a new skill set for many researchers who rarely have formal training in data management. Managing data for citizen science projects must also take into consideration the impact of the involvement of volunteers in the scientific work and related data management decisions. These are most notable in the areas of policies, privacy protection, data collection, and data quality. By implementing and upgrading data management practices to take the full data life cycle into account, citizen science projects have the potential to further extend outcomes supporting policy, decision-making, and scientific knowledge production.

Good data management requires adopting an evolving set of best practices, and establishing useful data management practices can take time. While it is relatively easy to set out data management plans for a brand-new citizen science project, even ongoing projects can benefit from revisiting each step in the data management life cycle and making gradual changes: add a terms-of-use policy if the project does not yet have one; document the steps involved in QA/QC; or locate a few repositories to look at more closely for long-term data storage. At the end of the day, data stewardship is the key to unlocking the tremendous potential of citizen science data.

Analysis Checklist
(Pilz, Ballard, & Jones, 2005)

- How will data be analyzed? Who will conduct the analyses?
- Will analyses be planned and documented during the sampling design phase and in advance of data collection?
- How will appropriate statistical expertise be retained to conduct the analyses? Will a professional statistician review them?
- How will all participants be included in the review and interpretation of results?
- What advance criteria will be used to interpret results? How will these criteria be collaboratively developed and applied?
- If any of the participants have concerns about the means of analysis, potential results, interpretation of the results, or use of the information, how will these concerns be addressed?
- If consensus cannot be reached about interpretation of the results, how will results be reported? Can alternate interpretations be included for comparison?

Example Free Analysis Tools

- **Many Eyes**
- **GeoServer**
- **Patuxent Wildlife Research Center's Software Archive**
- **The R Project for Statistical Computing**
- **Kepler, Taverna**, & **Pegasus**
- **OpenOffice**
- **Google Fusion Tables**

Acknowledgments

The development of this publication was supported by DataONE and NSF Grant #OCI-0830944.

The authors thank Caren Cooper and Heather Henkel for editorial feedback, and Marion Ferguson and Grace Lerner for assistance in preparing the guide for distribution.

This publication elaborates upon the DataONE Primer on Data Management (Strasser et al., 2012) and draws upon DataONE education modules (Henkel et al., 2012), linked in each section.

The guide should be cited as follows:

Wiggins, A., Bonney, R., Graham, E., Henderson, S., Kelling, S., Littauer, R., LeBuhn, G., Lotts, K., Michener, W., Newman, G., Russell, E., Stevenson, R. & Weltzin, J. (2013). Data Management Guide for Public Participation in Scientific Research. DataONE: Albuquerque, NM.

Contributorship

AW & WM wrote the introduction. SH & GL wrote the section on planning. GL, SH, KL & EG wrote the section on collecting. GN & RS wrote the section on assurance. ER, RS, & AW wrote the section on describing. KL, RL, AW & JW wrote the section on preserving. RS, WM & AW wrote the section on discovering. ER, SK & AW wrote the section on integrating. EG, GL, GN, ER, SK, & AW wrote the section on analyzing. AW wrote the conclusion, assured unified tone, and coordinated.

References

Henkel, H., Hutchison, V., Strasser, C., Rebich Hespanha, S., Vanderbilt, K., & Wayne, L. (Producer). (2012). 1/4/2013 Lessons 1-10. DataONE Education Modules. [Powerpoint files] Retrieved from http://www.dataone.org/education-modules

Hull, D., Wolstencroft, K., Stevens, R., Goble, C., Pocock, M. R., Li, P., & Oinn, T. (2006). Taverna: a tool for building and running workflows of services. Nucleic Acids Research, 34(suppl 2), 729-732.

Kelling, S. (2012). Using bioinformatics in citizen science. In J. L. Dickinson & R. Bonney (Eds.), Citizen Science: Public Participation in Environmental Research (pp. 58-68): Cornell University Press.

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., . . . Zhao, Y. (2006). Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience, 18(10), 1039-1065.

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. Trends in Ecology & Evolution (27), 85-93.

Michener, W.K., J.W. Brunt, J. Helly, T.B. Kirchner, and S.G. Stafford. 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7:330-342.

Pilz, D., Ballard, H. L., & Jones, E. T. (2005). Broadening participation in biological monitoring: guidelines for scientists and managers. Institute for Culture and Ecology, Corvalis, OR.

Strasser, C., Cook, R., Michener, W., & Budden, A. (2012). Primer on Data Management: What you always wanted to know (but were afraid to ask). Electronic: DataONE.

Strasser, C., Cook, R., Michener, W. K., Budden, A., & Koskela, R. (2011). Promoting data stewardship through best practices. Paper presented at the Environmental Information Management Conference, University of California, Santa Barbara.

Wiggins, A. (2013). Free As In Puppies: Compensating for ICT Constraints in Citizen Science. In Proceedings of the 16th ACM Conference on Computer Supported Cooperative Work and Social Computing. San Antonio, TX, 23–27 February, 2013.

Wiggins, A., Newman, G., Stevenson, R. D., & Crowston, K. (2011). Mechanisms for Data Quality and Validation in Citizen Science. Paper presented at "Computing for Citizen Science" workshop, IEEE eScience Conference. Stockholm, SE, 5 December, 2011.