Data Recovery: Aye or Nay?

The issue of whether or not to attempt to recover data that are in danger of being lost has been on mind of late for two reasons. First, I recently reviewed two proposals that focused significant resources on a set of largely unspecified data recovery efforts of questionable feasibility and benefit. Second, I read the intriguing account of Alison Specht and her colleagues who recently completed a four-year effort to recover continental-scale vegetation survey data. Whereas the proposed data recovery efforts were extremely naive in many areas, the effort by Specht and colleagues was justifiﬁed and well-organized. The project also offers many lessons that help form the basis for the data recovery ruleset that I present below.

Specht and colleagues embarked on a four-year effort from 2014 to 2017 to recover natural history and ecological data from published plant surveys dating back to the 1880’s; data that had been collected, integrated, digitized and interpreted for classiﬁcation of vegetation and conservation status in Australia initially during the 1970’s. An earlier effort to recover the data took place during the 1980’s-’90s. The most recent data recovery effort was justiﬁed on the basis of the perceived value of the historical data by at least two research organizations, the national coverage of the data, and the fear of the potential for complete data loss.

Figure 2 from their paper [reproduced here] illustrates the diversity and volume of resources available for the data recovery effort. As one might expect, the team encountered many problems during their data recovery efforts, including: (1) changes in technology; (2) changes in spatial referencing; (3) changes in data formatting and structure; (4) changes in species names; and (5) availability of documentation. Consequently, the data recovery effort was both time- (4-years) and resource-intensive (estimated at AU$ 200,000). It was assumed that the benefits will outweigh the costs, but the proof of this assertion probably lies many years out and will be based on future research and conservation and management decisions and actions that are enabled. The paper is deﬁnitely worth a read as it focuses on motivation and is critical for a ruleset that can help determine whether the effort is a complex one. Nevertheless, I propose that three questions form the basis for a ruleset that can help determine whether to proceed with a major data recovery effort:

1. Can the data be discovered and accessed? This question embodies several facets. First, one has to know what speciﬁc data are being sought. Then, it is necessary to ascertain whether the data exist and whether they can be accessed. This may not be as simple as it seems, requiring one to retrieve data from old media (e.g., tapes, keypunch cards, paper documents, etc.) as can be seen in Figure 2. But, a pilot project may help in demonstrating feasibility and determining/estimating the amount of effort involved.

2. Is there sufﬁcient metadata to interpret and [re-]use the data? Assuming the data are accessible, then it is necessary to assess the quality of the metadata to assure that it enables one to understand and potentially reuse the data and verify that it is fit for the purpose(s) intended. Often, data and metadata are stored in different locations and one or the other may not be available. Furthermore, metadata are frequently sparse and do not necessarily include all information needed to interpret and use the data (e.g., units, time collected, location and coordinate system used, species names, QA/QC procedures). If one or more of the data originators are alive, then it may be possible to reconstruct the requisite metadata.

3. Who cares and why? The ﬁrst two questions above are more technical in nature. This question may be the most important of all as it focuses on motivation and is critical...
in understanding who wants to use the recovered data and for what purpose. The answers will establish the justification or need for initiating the recovery effort as well as who might be involved, including potential funders that might be willing to support the work. As we saw from the example above, data recovery can be a long-term effort requiring a committed team and substantial funding. Conversely, the absence of individuals that are motivated to recover data and resources to support the effort virtually guarantee failure.

Once a decision is made to initiate a data recovery effort, then I would further propose that a data triage exercise may be in order especially for very large data rescue projects. In this exercise, recovery activities would be prioritized based on: (1) the likelihood that the data or subset of data would survive regardless of what actions take place; (2) the likelihood that the data or subset of the data would not survive regardless of what actions take place; and (3) the likelihood that quick action would make a difference. Of course, neither data recovery nor data triage are necessary—if data are logically structured, well-documented, and protected and preserved in the first place. DataONE’s education resources provide such training and best practices in data stewardship.

—William Michener
Principal Investigator, DataONE

Data Discovery: Aye or Nay? cont’d

Once a decision is made to initiate a data recovery effort, then I would further propose that a data triage exercise may be in order especially for very large data rescue projects. In this exercise, recovery activities would be prioritized based on: (1) the likelihood that the data or subset of data would survive regardless of what actions take place; (2) the likelihood that the data or subset of the data would not survive regardless of what actions take place; and (3) the likelihood that quick action would make a difference. Of course, neither data recovery nor data triage are necessary—if data are logically structured, well-documented, and protected and preserved in the first place. DataONE’s education resources provide such training and best practices in data stewardship.

Data Discovery: Aye or Nay? cont’d

Once a decision is made to initiate a data recovery effort, then I would further propose that a data triage exercise may be in order especially for very large data rescue projects. In this exercise, recovery activities would be prioritized based on: (1) the likelihood that the data or subset of data would survive regardless of what actions take place; (2) the likelihood that the data or subset of the data would not survive regardless of what actions take place; and (3) the likelihood that quick action would make a difference. Of course, neither data recovery nor data triage are necessary—if data are logically structured, well-documented, and protected and preserved in the first place. DataONE’s education resources provide such training and best practices in data stewardship.

How-to Make your Data Count

The Make Data Count project is excited to share the release of the project’s Version 1 of standardized data usage and citation metrics. The team has been working to implement comparable, standardized data usage and citation metrics with both the California Digital Library’s Dash and DataONE repositories, with the DataONE usage and citation metrics user interface coming in July 2018.

Learn more about implementing Make Data Count at your repository and additional information about the Make Data Count project by visiting their webpage and for a full overview you can watch one of their introductory webinars at: https://makedatacount.org/presentations/

In each newsletter issue we highlight one of our current Member Nodes. The full list of Member Nodes and summary metrics can be found on the DataONE.org site at bit.ly/D1CMNs.

Chinese Ecosystem Research Network (CERN)
http://www.cern.ac.cn/

Meeting the needs of both the national and international ecological research communities

Chinese Ecosystem Research Network (CERN) was established in 1988 to conduct research on China’s major ecosystems on a long-term basis by integrating ground-based networking observation and experiments with simulation and modeling. CERN is an important part of the knowledge innovation project of the Chinese Academy of Sciences and has become a national-level ecological network with significant international influence, including the Long-Term Ecological Research Network (LTER Network) and the UK Environmental Change Network.

CERN’s monitoring and experimental activities are distributed across one synthesis research center, five disciplinary sub-centers, and 44 ecological research stations representing diverse ecosystems, including 15 stations for agriculture, 11 for forest, 2 for grassland, 6 for desert, 2 for marsh, 3 for lake, 1 for urban, 1 for karst, and 3 for marine ecosystems. This network makes it possible to conduct ecological comparative research in China, which can provide more comprehensive and systematic scientific data for national macro decision-making.

Recently, CERN made the decision to develop their own data repository using a DataONE Metacat software implementation and become a DataONE Member Node. As their data contributions slowly ramp up, CERN’s participation in the DataONE network will help advance their mission to offer data and information to scientists and policy-makers and contribute to ecosystem management, wise use of natural resources, and sustainable socio-economic development.

How-to Make your Data Count

The Make Data Count project is excited to share the release of the project’s Version 1 of standardized data usage and citation metrics. The team has been working to implement comparable, standardized data usage and citation metrics with both the California Digital Library’s Dash and DataONE repositories, with the DataONE usage and citation metrics user interface coming in July 2018.

Learn more about implementing Make Data Count at your repository and additional information about the Make Data Count project by visiting their webpage and for a full overview you can watch one of their introductory webinars at: https://makedatacount.org/presentations/

In each newsletter issue we highlight one of our current Member Nodes. The full list of Member Nodes and summary metrics can be found on the DataONE.org site at bit.ly/D1CMNs.

Chinese Ecosystem Research Network (CERN)
http://www.cern.ac.cn/

Meeting the needs of both the national and international ecological research communities

Chinese Ecosystem Research Network (CERN) was established in 1988 to conduct research on China’s major ecosystems on a long-term basis by integrating ground-based networking observation and experiments with simulation and modeling. CERN is an important part of the knowledge innovation project of the Chinese Academy of Sciences and has become a national-level ecological network with significant international influence, including the Long-Term Ecological Research Network (LTER Network) and the UK Environmental Change Network.

CERN’s monitoring and experimental activities are distributed across one synthesis research center, five disciplinary sub-centers, and 44 ecological research stations representing diverse ecosystems, including 15 stations for agriculture, 11 for forest, 2 for grassland, 6 for desert, 2 for marsh, 3 for lake, 1 for urban, 1 for karst, and 3 for marine ecosystems. This network makes it possible to conduct ecological comparative research in China, which can provide more comprehensive and systematic scientific data for national macro decision-making.

Recently, CERN made the decision to develop their own data repository using a DataONE Metacat software implementation and become a DataONE Member Node. As their data contributions slowly ramp up, CERN’s participation in the DataONE network will help advance their mission to offer data and information to scientists and policy-makers and contribute to ecosystem management, wise use of natural resources, and sustainable socio-economic development.

How-to Make your Data Count

The Make Data Count project is excited to share the release of the project’s Version 1 of standardized data usage and citation metrics. The team has been working to implement comparable, standardized data usage and citation metrics with both the California Digital Library’s Dash and DataONE repositories, with the DataONE usage and citation metrics user interface coming in July 2018.

Learn more about implementing Make Data Count at your repository and additional information about the Make Data Count project by visiting their webpage and for a full overview you can watch one of their introductory webinars at: https://makedatacount.org/presentations/
Participating in DataONE with the Generic Member Node

Any preservation-oriented data repository may join the DataONE federation as a Member Node and so enjoy the many benefits of participation. Such benefits include comprehensive metadata indexing across all repositories ensuring data are more findable [1]; more accessible data and metadata through enforcement of persistent, unique IDs and an identifier resolution service for all content [2]; client access to data resources is simplified by a consistent set of APIs to support interoperability across all Member Nodes [2]; support for W3C-PROV through the PROV-ONE model [3] helps to ensure that datasets are more reusable. The DataONE infrastructure also provides replication of content to help ensure long term access, usage metrics following the COUNTER code of practice for research data [4] and new emerging features such as metadata quality evaluation [5].

Data repositories participate in the DataONE federation by implementing the necessary APIs or more readily by deploying a compatible data repository system such as Metacat [6] (a web application implemented in Java) or the Generic Member Node [7] (GMN, a web application implemented in Python), both of which fully implement the DataONE APIs at all tiers of functionality. Metacat supports the MetacatUI, the same search interface used in all tiers of functionality. Metacat also provides an EML metadata editor which is enabled when running on a Member Node (see for example, https://knb.ecoinformatics.org/submit). The GMN offers flexible deployment options and can act as a proxy for existing repositories to facilitate translation of existing services such as OAI-PMH [8], OGC-CSW [9], or schema.org publishing patterns [10] to the necessary DataONE Member Node services. In this manner, GMN offers the ability for existing repositories to join DataONE through hosting of an appropriately configured instance of GMN.

In response to the increasing popularity of the GMN repository proxy approach, a new multi-site hosting GMN feature has been released. This new capability enables a single instance of GMN to appear as one or more DataONE Member Nodes with each acting as a repository or as a proxy for an existing repository. In this manner, an individual institution or organization may operate a single GMN instance and provide access to all their data repositories as distinct DataONE Member Nodes and so ensure full recognition through function and branding of the distinct data repositories. Figshare for Institutions [11] is one example of an organization where multiple independently branded portals are provided as a service for Institutions. Previously, it would be necessary to install individual instances of GMN to preserve an independent view of each participating institution as presented by their Figshare portal. The new GMN capability significantly simplifies the process for additional institutions using Figshare to also participate as a Member Node in DataONE.

This new capability is being leveraged by the Cary Institute of Ecosystem Studies [12] of Millbrook, New York which is a leading independent environmental research organization that relies on services such as Figshare for Institutions to provide a secure repository for their research output. Similarly, the Interdisciplinary Earth Data Alliance [13] (IEDA) is currently in the testing phase of leveraging GMN to share content through the schema.org pattern of dataset publishing for the IEDA hosted EarthChem [14], USAP-DC [15], and Marine-Geo Digital Library [16] (MGDL) repositories. In each case, an individual instance of GMN is associated with Figshare and IEDA respectively and may be easily extended to support additional repositories without the need for additional service installations.

GMN is available from GitHub at https://github.com/DataONEorg/d1_python

Metacat is available from GitHub at https://github.com/NCEAS/metacat

For more information on participating in the DataONE federation, please contact us at https://www.dataone.org/contact

— David Vieglais
Director for Development and Operations, DataONE
A Researcher’s Guide to DataONE

At DataONE we support the search and discovery of Earth and environmental data across repositories around the world through a continuously growing network. We also provide open access to educational resources, webinars and comprehensive materials on data management best practices. To help navigate these materials within the dataone.org landscape we have created a handy roadmap that directed users to relevant content as they work through their research process. Walk through the interactive roadmap and click on each of the icons to discover DataONE guidance, tools and webinars in support of your research data management questions.

Visit https://www.dataone.org/researcher-guide to interact with the live Guide.

Now Playing at the AGU!

Take some time at AGU to watch the new DataONE video “Meeting the Data Needs of the Environmental Sciences.” DataONE is one of the universities, institutions, and organizations being highlighted in five-minute documentary style films during the AGU conference. These will be shown around the convention center and on the Fall Meeting website. Learn more at: https://fallmeeting.agu.org/2018/exhibitors/agu-tw/. Please take a few minutes to watch the video, now available at the DataONE website and on our Vimeo channel, https://www.vimeo.com/dataoneorg/.

The film summarizes DataONE’s ongoing community data goals and features members of the DataONE team. We would like to especially thank Ben Halpern (NCEAS) and Jeanette Clark (NCEAS) for volunteering to participate in the film. Read the summary below to learn more about the film.

The DataONE video summary:

Data are diverse - spanning time, space, and disciplines - and are being generated at unprecedented rates. DataONE works to engage the community in creating solutions that enable data managers, librarians, researchers and others to leverage data in support of pressing ecological challenges. DataONE helps researchers address these challenges by providing a single search interface that allows discovery of content from an ever-growing collection of data repositories. In addition to simplifying data discovery, DataONE offers high quality resources for data management, including teaching materials, webinars and a database of best-practices, which help educators and librarians with training and improve methods for data sharing and management. By joining our community of repositories, data managers can increase visibility and exposure of their data, and have the opportunity to replicate their content across DataONE’s geographically distributed network. Visit DataONE.org and you’ll spend less time searching and more time finding the data you need.
Ten Principles to Improve EBV Informatics

A new paper published in the journal Ecological Informatics outlines an interoperability framework for Essential Biodiversity Variable (EBV) data products.

Since EBVs were first proposed (Pereira et al. 2013, Science), a key question has been how to prepare data products for EBVs on a global scale. How can comprehensive EBV data products be compiled for any geographical area, over any required time period, for any species, assemblage, ecosystem, or biome of interest, and with data that may be held by any (or across multiple) data repositories?

A key step to answer those questions includes the improvement of cooperation and interoperability among multiple stakeholders, including data and research infrastructure providers around the world such as GBIF, Atlas of Living Australia, DataONE, NEON, SAEDON, SANBI, CRIA, the Biodiversity Committee of the Chinese Academy of Sciences, and many others. Building on discussions of informatics experts and representatives of such infrastructures during four workshops of the Horizon 2020 project GLOBIS-B funded by the European Commission (http://www.globis-b.eu/), the published paper provides implementation guidelines of the emerging EBV operational framework. The publication suggests ten areas where data and informatics interoperability among infrastructures can be improved in support of EBVs. The ten areas cover data management planning, data structure, metadata, services, data quality, scientific workflows, provenance ontologies/vocabularies, data preservation and accessibility.

For each area, a core interoperability principle is described and desired outcomes as well as short- and long-term goals are provided. Collectively, these ten principles aim to improve trans-national and cross-infrastructure workflows for EBV production. The implementation guidelines are presented as the ‘Bari Manifesto’, named after the town in southern Italy where they were specified. The Bari Manifesto provides a strong basis for understanding the operationalization of EBVs, we are still lacking sufficient technical, semantic and legal interoperability to actually build multiple EBV data products at a global scale’, explains W. Daniel Kissling, the scientific coordinator of the GLOBIS-B project. ‘With the Bari Manifesto we have been able to bring many informatics experts and infrastructure operators together and we hope that our suggested implementation guidelines will allow to build scientific workflows for making reproducible and transparent EBV data products’, says Kissling.

Publication:
DATA DISCOVERABLE THROUGH DATAONE

48 TB of content
804 K metadata
1.17 M data

2,430 Visitors to our page*
1,160 Searches conducted*
99.99% Uptime of Coordinating Nodes

*metrics are running monthly averages; symbol denotes change since last quarter

OUR COMMUNITY

42 Contributing Members

New Member this Quarter
ENVIRONMENTAL SYSTEM SCIENCE DATA INFRASTRUCTURE FOR A VIRTUAL ECOSYSTEM

500 DataONE User Group members

5,300+ Users trained

EDUCATION AND OUTREACH

Webinar Series
34 Webinars
92 Average number of attendees
1,722 Unique webinar attendees

Education Resources
19,628 Visits to the public webpage*
217 Education Module downloads*

*metrics are running monthly averages; symbol denotes change since last quarter

Most Downloaded Resources
1 Data Management Plan Example from NSF
2 Best Practices Primer
3 Data Management Plan Example from Manua Loa

Most Visited Pages
1 Education Module homepage
2 Data Management Planning
3 Best Practices: Create and document data backup policy

Repositories in the DataONE Federated Network

SOURCE: CN.DATAONE.ORG
Only the first version of each file is counted

Metrics inclusive of November 2018