4-28-2015

DataONE Personas

UA/SC WG Personas Subgroup

DataONE Personas

Developed by the DataONE Socio-cultural
and Usability & Assessment Working Groups (SC/UA)
Subgroup on Personas

Kevin Crowston <crowston@syr.edu>
Ahrash Bissell <ahrash.bissell@gmail.com>
Bruce Grant <bwgrant@widener.edu>
Maribeth Manoff <mmanoff@utk.edu>
Rebecca Davis <rebeccas@utk.edu>

Last edited: 28 April 2015

Please cite as: DataONE Socio-cultural and Usability & Assessment Working Groups Subgroup on Personas. (2015). *DataONE Personas*. Available from: http://dataone.org/persona-methods.

Acknowledgements

DataONE is a collaboration among many partner organizations. It is funded by the US National Science Foundation (NSF) under a Cooperative Agreement under Grant Numbers 0830944 and 1430508.

Disclaimers

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or of the DataONE partner organizations.

The inclusion of an image or reference to institutions and individuals within this document does not express or imply the endorsement of DataONE by those institutions or individuals, nor the endorsement by the authors or DataONE of the entity or the entities' products, services or enterprises.

All personas in this work are fictitious. Any resemblance to real persons, living or dead, is purely coincidental.

DataONE Personas

This document presents the personas developed to guide development of the DataONE cyberinfrastructure. “A persona, first introduced by Alan Cooper, defines an archetypical user of a system, an example of the kind of person who would interact with it. The idea is that if you want to design effective software, then it needs to be designed for a specific person.” [1] Personas are similar to use cases and scenarios, but with additional richness. Use cases treat all interactions as equally important; scenarios focus on tasks, rather than users [2, p. 59]. Personas add detail about user interests, emotions, settings and needs that drive system usage. “Personas are incredibly useful when you don’t have easy access to real users because they act as ‘user stand-ins’, helping to guide your decisions about functionality and design.” [1]. Having a shared set of personas help developers maintain a common vision of the user and promote agreement between different stakeholders.

There are several kinds of personas: primary (the main user or users of the system); secondary (those who will be served as long as doing so doesn't affect the primary users); negative (those who will explicitly not be served because to do so would move the project in an undesired direction); and buyer (those who make decisions about the project and whose opinions need to be understood) [3]. For DataONE, there are many primary personas, some secondary and no negative or buyer personas.

Most of the primary personas are research scientists. Research scientist personas were developed to span multiple dimensions that might affect the individual’s use of DataONE:

- Work setting: Academic (tenure and non-tenure track), government/tribal, private
- Career stage (early, mature, late)
- Subject/discipline
- Single discipline vs. use of multi-disciplinary data
- Research setting: Field, lab, modeller
- Data: Human vs. machine-collected
- Data management skills: novice to expert

DataONE personas were developed drawing from the Data Conservancy scenarios (from Anne Thessen <athessen@eol.org>), usage Scenarios developed by the DataONE Sustainability and Governance Working Group, Data Conservancy profiles from Illinois and Purdue (http://datacurationprofiles.org), the researcher survey done by the DataONE Usability and Assessment Working, interviews and the life experiences of the authors. Sources for each persona are given in the persona description.

The description of a persona for DataONE include [3]:

- Background
- Name, age, and education
- Socioeconomic class and socioeconomic desires
- Life or career goals, fears, hopes, and attitudes
- Reasons for using DataONE to share and to reuse data

- Needs and expectations of DataONE tools
- Intellectual and physical skills that can be applied
- Technical support available
- Personal biases about data sharing and reuse (and data management more generally)
- DataONE usage scenarios (see the Appendix for a generic list of functionality)

Related work

Purdue has guidelines for developing "Data Curation Profiles"; a copy is on the DataONE document site: see http://bit.ly/DCCprofiles. Cornell library developed a set of persona for library users: see http://hdl.handle.net/1813/8302. The Data Conservancy also developed personas.

Personas and the data lifecycle

For each primary persona, we show the data lifecycle (Figure 1), depicting which of the stages of the lifecycle the individual performs currently (in blue) and which might be performed using tools provided by DataONE (in mauve). See Figure 2 for an example data lifecycle figure for a persona. Stages shown shaded out are not performed by the persona; those shown in smaller or italicized font are performed but at a lesser level (i.e., less than what would be considered best practice). Solid lines from stage to stage represent workflows performed by the persona. Note that the lifecycle is only a cycle from the perspective of the data; from the perspective of a persona, there is a generally a break between preserve and discover, as the individuals preserve data for others to (potentially) use and similarly discover other people's data to use themselves. Curved 3D lines in the figure represent flows of data from one user to another (as shown in Figure 3).

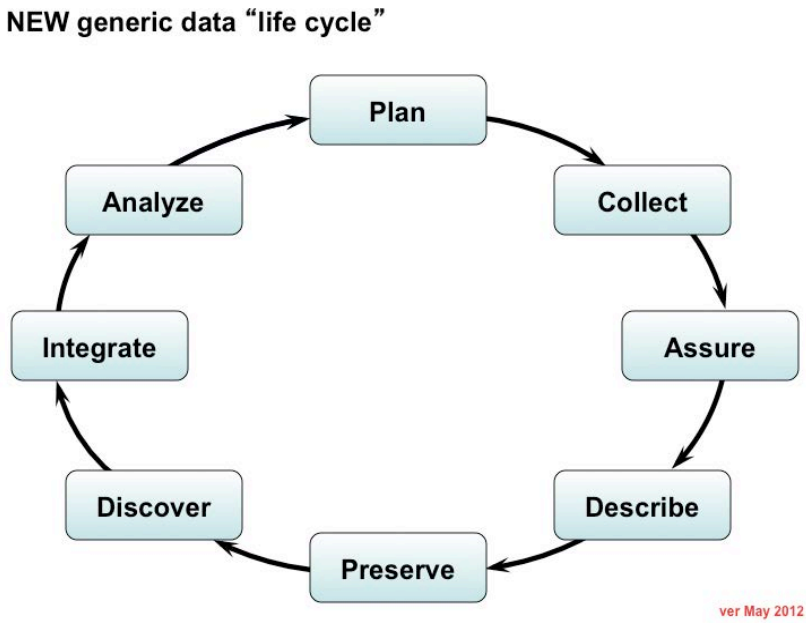


Figure 1. The DataONE data lifecycle (from https://www.dataone.org/best-practices).

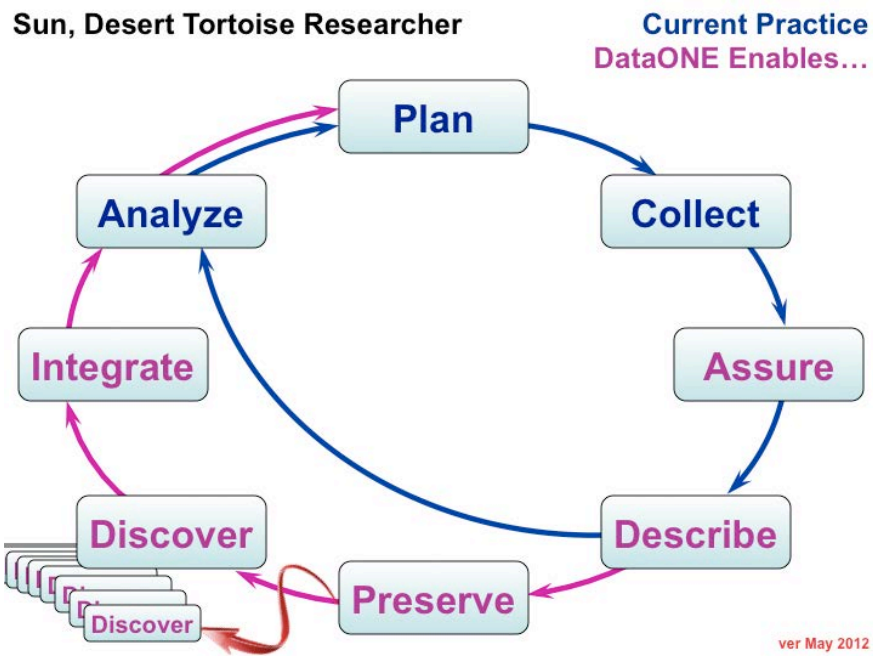


Figure 2. Example representation of workflow around data for a particular persona (Sun). Blue lines and stages represent current practice (i.e., without DataONE); mauve lines and stages represent practice enabled by use of DataONE tools; and red wavy lines represent data flows from the focus individual to other researchers.

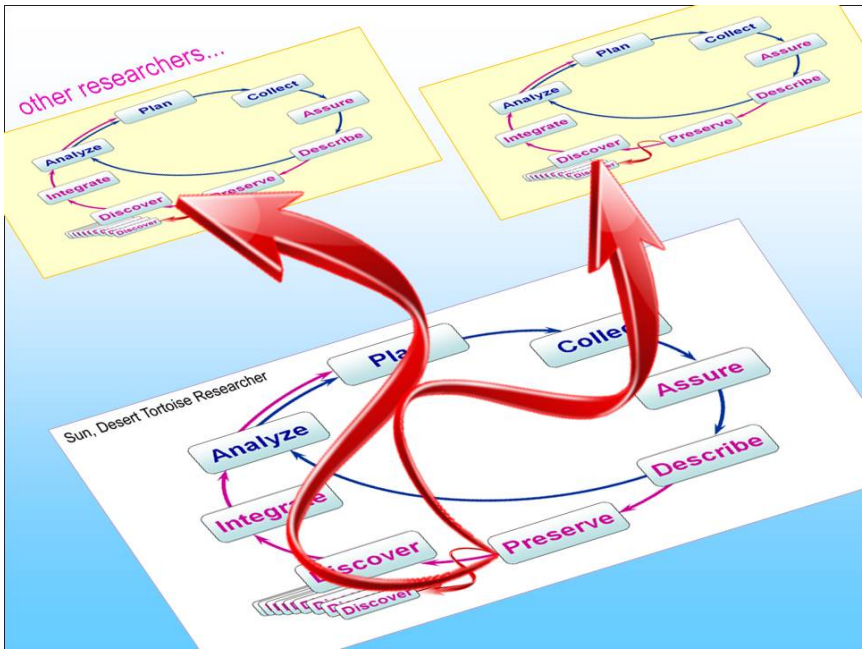


Figure 3. Red wavy lines represent data flows from the focus individual to others.

References

[1] Ambler, Scott W. (2009). *Introduction to Persona*. Available from: http://www.agilemodeling.com/artifacts/personas.htm. Accessed 28 April 2015.
[2] Madsen, Sabine and Nielsen, Lene. (2009). Exploring persona-scenarios: Using storytelling to create design ideas. In Dinesh Katre, Rikke Orngreen, Pradeep Yammiyavar, Torkil Clemmensen (Eds). *Human Work Interaction Design: Usability in Social, Cultural and Organizational Contexts: Proceedings of the Second IFIP WG 13.6 Conference*, Pune, India, 7-8 October.
[3] Rind, Bonnie. (2007). The Power of the Persona. *The Pragmatic Marketer Magazine*, 5(4), 18-22. Available from: http://www.pragmaticmarketing.com/publications/magazine/5/4/the-power-of-the-persona. Accessed 28 April 2015.

Table of Contents

Completed DataONE personas

Sun

 [1]

(Primary persona)

Source: Data Conservancy Sun persona by Anne Thessen; comments from Lynn Rogers; revised by Kevin Crowston with some details based on William I. Boarman, USGS.

Tags: non-academic, government, early career, single discipline, field, human and machine-collected data, novice data management, biology

See also: Dr Yolanda Suarez DataONE Scenario

Background

Name, age, and education

Sun is a biologist specializing in desert tortoises. She did her masters and PhD at California State University San Marcos. She has spent her career studying tortoises in their natural habitat.

Life or career goals, fears, hopes, and attitudes

Sun recently started working for the USGS Western Ecological Research Center, "one of 18 Centers of the Biological Resources Discipline of the U.S. Geological Survey" (http://www.werc.usgs.gov/who.aspx). Her broad interest is how human activity and climate change will affect tortoise populations. Her research needs to inform decisions by land managers in various state and federal agencies. She works with NGOs on conservation issues and speaks to the public on tortoises and conservation issues. For example, she collaborates with biologists at the Wildlife Research Institute (http://www.wildlife-research.org/page10.html) on a project tracking desert tortoises relocated from the expanding Fort Irwin Army Base. She writes technical reports and also publishes peer-reviewed journal articles (e.g., http://www.conservation-science.com/Products.html; http://www.werc.usgs.gov/person.aspx?personID=52).

A day in her life

Sun and other members of the research team go into the field with a notebook, camera, simple instruments and sample containers. They capture and tag tortoises before collecting data about individuals such as age, weight and sex. They also collect data about entire tortoise populations by taking a census, collecting feces and monitoring carcasses. Much of these data are recorded in a notebook and later copied onto a spreadsheet for analysis with desktop statistics software. A

[1] https://www.flickr.com/photos/armyenvironmental/2650014187 (CC BY 2.0). Picture is of Dr. Paula Khan,

number of her research subjects are radio tagged, giving her a lat/long position as often as every 10 minutes.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

Sun feels that she cannot easily share her own data for fear of disclosing sensitive information because of the work location and the fact that she works on endangered species. Even an embargoed dataset could be problematic, as tortoises keep the same home range and the lifespan of a tortoise vastly exceeds the duration of any reasonable embargo. However, she might be able to share derivative datasets, if these could be easily created, or a subset of less sensitive data, such as life history, demographic or behavioural data (e.g., home range size, daily and seasonal activity, diet, social biology or thermo-regulatory behaviour).

DataONE might also be useful in improving Sun's overall data management capabilities, e.g., educating her on best practices for data quality and metadata development. If DataONE provided tools for cataloguing and managing locally-stored data, these could be very useful. She might be willing to deposit data at a member node for limited sharing, preservation and for ensuring long-term preservation of data (e.g., migration of data formats), though only if its privacy can be assured and doing so were as easy as (or at least, not much harder than) maintaining local backups.

Sun is interested in finding additional data that correspond to the location of tortoise populations, and additional tortoise data so she can put her current study into perspective and perhaps find collaborators. For example, data on invasive species in the area she studies could help explain changes observed in the populations. She does not have much technical support, so she need the tools to be easy to use. Given that her research is motivated by both scientific interests and policy concerns, she is extremely wary of using data of unknown origin or quality, so discoverability and validation of datasets are key issues.

Intellectual and physical skills that can be applied

As a trained research scientist, there should be no overt challenges to dealing with data *per se*. However, though Sun strives to follow established data-collection protocols, the realities of field research mean that her methods are often adjusted on the fly and her data needs secondary analysis and clean-up. If DataONE provides tools to aid in the integration of similar, yet not identical, datasets, and can help her to troubleshoot data-entry and other errors in her own data, her own use and possible subsequent deposition of her data into a DataONE member node would be simple.

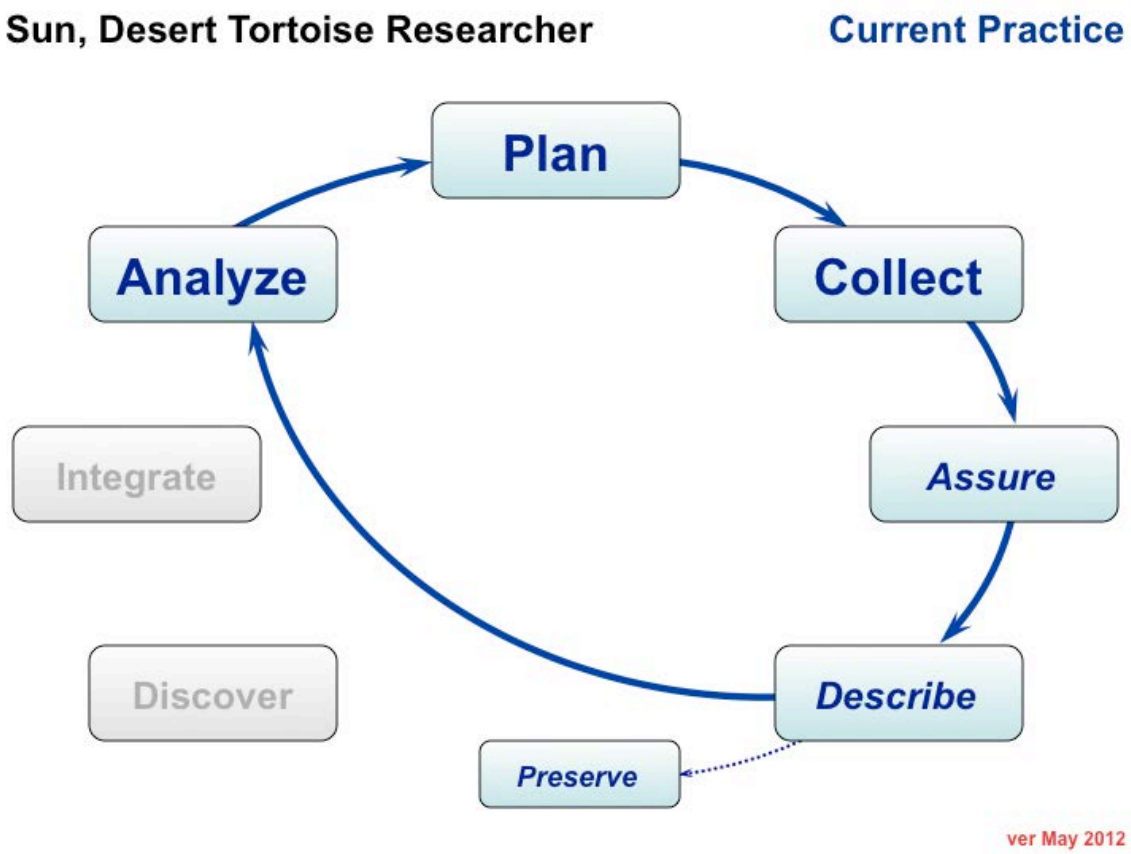
Technical support available

Sun has very little computer support within her research group and institution but she does have experience with field equipment and general computer competencies. Thus far, complex visualizations and data-handling algorithms have not been a factor in her work, so any system that

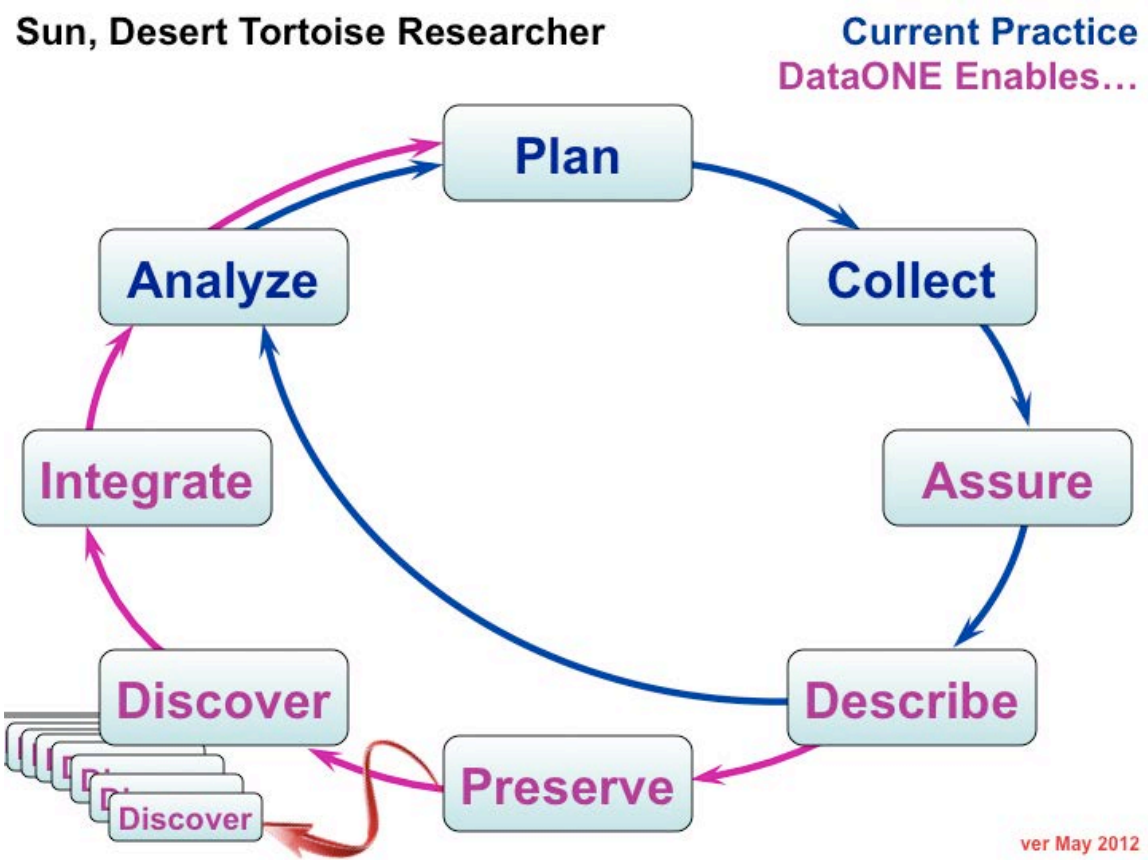
did not offer the option to work with simple datasets using easy tools would probably intimidate her.

Personal biases about data sharing and reuse (and data management more generally)

Sun is interested in reviewing data that might inform her studies, but does not depend on it and it is not yet an important part of her work. On the other hand, she does not have the technical skills to prepare her data for sharing nor does she have large quantities of data that she thinks would be of interest to others. Furthermore, she is hesitant to share her geolocated data because she works with a threatened species. So far, she has only shared raw data with close colleagues.



Sun currently collects data only for her own use. She validates her data and describes it, though not following any broadly-used data quality or metadata standards. Deposit is in the form of publications based on summaries and analyses; the raw data themselves are not shared. These data are then analyzed and used to drive further data collection.



Sun could use DataONE tools (and the training in their use) to improve her capabilities for data assurance and description. Under the right conditions, she could use DataONE tools for preparing data for deposit and preservation, and potentially even for reuse of appropriately redacted data by other researchers. The main motivation for Sun to use DataONE would be to improve her data management practices and discover potentially useful data created by other researchers to integrate into her own analyses.

Comparison of current and DataONE-enabled practices:

Project Planning

- *Management Planning:* Develops a project Data Management Plan following examples provided on the DataONE portal.

Current data collection:
Collects tortoise field data.
DataONE enabled data collection:
No change.

Current data assurance:
Validates data using own standards.
DataONE enabled assurance:
Could apply more broadly-used data quality standards and assurance tools.

Current data description:
Describes data for her own purposes, using her own data description techniques.
DataONE enabled description:

- *Training:* Learns how to use *Morpho* (a metadata management editor) based on instructional materials available in the DataONE Best Practices Database and associated downloadable video instructions.
- Creates metadata for datasets following best practices.

Current data preservation:
Sun publishes summary and analysis results but does not deposit data. Data preservation is done only within her lab.
DataONE enabled preservation:
Sun might deposit data with a DataONE member node for long-term preservation, with appropriate protections for sensitive data.

- *Data Preservation:* Deposits data and metadata in the USGS data repository with appropriate protections for sensitive data and redaction to create shareable data subsets.
- *Data Preservation:* Submits a research paper to an ecological journal associated with Dryad—a DataONE Member Node. Upon acceptance, she submits the publication-relevant data, metadata, and model to Dryad where they are given a DOI (digital object identifier) and preserved in the Dryad repository.
- *Citation:* Upon publication, she adds the publication reference and the data citation (including DOIs for both; provided by Dryad and the journal) to her CV.

Current data discovery:
Does not use other researchers' data.
DataONE enabled discovery:
The possibility of discovering relevant data from other researchers is likely to be a main motivation for Sun's use of DataONE and DataONE tools.

- *Data Discovery, Access, Use and Dissemination:* Searches for tortoise food web and area meteorological data in the region at the DataONE portal. Searches for land-use histories, especially for former grazing lands. Searches for co-locality data for other animal species as possible signals for other ecological changes in the region.
- *Data Discovery, Access, Use and Dissemination:* Identifies relevant data and downloads data and metadata from previous LTER studies as well as data collected by state and Federal agency scientists (i.e., non-LTER).
- *Data Discovery, Access, Use and Dissemination:* Acquires supplemental data from another DataONE Member Node with complete citation information.
- *Citation:* Another scientist working in Mexico on a similar study discovers the new publication and data created by Sun and cites her in his work.

Current data integration:
Does not use other researchers' data.
DataONE enabled integration:

Could use DataONE tools to integrate her data with data discovered from other researchers.

Current data analyses:
Uses standard desktop data analysis tools.
DataONE enabled analysis:

- *Data Visualization:* Uses data analysis and visualization tools identified through DataONE Tools Database or available as part of the Investigator Toolkit to analyze existing data and develop initial model parameters that she will use in her own research.
- *Data Visualization:* Creates graphics using tools identified via DataONE.

Jean

 [2]

(Primary persona)

Source Data Conservancy Jean persona by Anne Thessen: Interview with Zach Lippman and comments from Sherri Simmons; revised by Kevin Crowston

Tags Academic, university, mid career, multi-discipline, experimental, human and machine-collected data, experienced data management, agriculture

See also:

Background

Name, age, and education

Jean is an agricultural scientist working at the Cornell University agricultural field station in Geneva NY. He received a PhD in horticulture from Virginia Polytechnic Institute in 1987.

Life or career goals, fears, hopes, and attitudes

Jean is a tenured associate professor in the Department of Horticulture. Jean's current project uses the tomato as a model system to study sympodial growth (a growth pattern in which the stem is a succession of growths rather than one). He is driven by the agricultural research station's program of research and the need to publish and to obtain grant funding to support his research.

A day in his life

Jean's project involves two types of data: phenotypic and genomic. The phenotypic data are collected via pencil and paper after seeds are germinated. The genomic data come off of a sequencing machine and are assembled by a computer. Jean has two types of sequence data: whole genome and transcriptome. His work produces extremely high amounts of molecular data that require significant technical support. Jean uses pedigree numbers to connect genotype, phenotype and generation and all data are stored in a pedigree book. He uses MEINS guidelines for metadata for the genomic data (Yilmaz et al., 2010, doi:10.1038/npre.2010.5252.2) as there are no metadata standards specifically for his discipline, probably because researchers are still trying to figure out how to handle and analyze the data. He knows plant ontologies exists, but doesn't use them because they do not serve his needs—they are too general.

---

[2] https://www.flickr.com/photos/cimmyt/9538063625 (CC BY-NC-SA 2.0). Picture is of Dr. Barnabas Kiula of the International Maize and Wheat Improvement Center.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

Jean is wary of going exclusively digital with his phenotype data because of the horror stories he's heard from other colleagues who have lost lots of work. However, he does transcribe data from paper to an Excel sheet. He keeps the paper copy and sometimes refers back to it to jog his memory. He uses basic visualizations within Excel to verify the accuracy of data transfer and to correct (or verify) any outliers; if these functions were easier to perform using DataONE tools, he might be convinced that digitizing his data for deposition at a DataONE member node is worthwhile.

He does have concerns about the long-term preservation of his data, as there are currently no formal process in place for long-term data management. He might be interested in depositing data at a member node for preservation (e.g., migration of data formats), though only if doing so were as easy as (or at least, not much harder than) maintaining local backups.

Intellectual and physical skills that can be applied

Jean is quite focused on his own research and has not historically involved many colleagues in collaborative work outside of his particular area of specialization. As such, he does not see the rationale for common data management protocols and believes that his data are only likely to be of interest to a very select number of researchers, most of whom he knows personally. That said, he is interested in being able to do perform longitudinal and synthetic analyses of his own work, something which is currently impossible due to the shifting standards applied to genomic data. This issue is interesting enough to Jean that he would be likely to contribute his expertise and sample data for the purpose of developing ontologies that actually meet his needs and could be supported for subsequent use in DataONE.

Technical support available

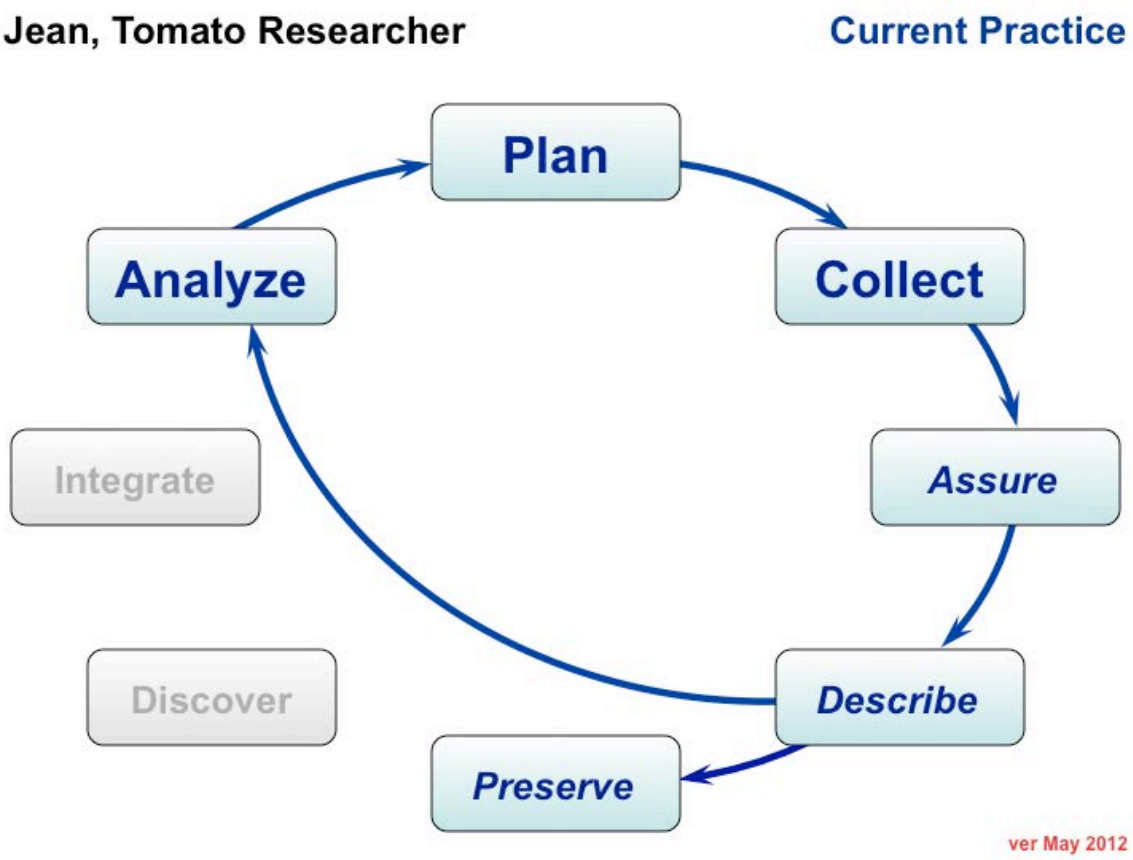
Jean funds good technical support within his research group. He knows that data management and archiving is becoming a more important issue for his field, and he is willing to devote resources to doing a better job of it, despite his concerns about the ultimate utility to his own work.

Personal biases about data sharing and reuse (and data management more generally)

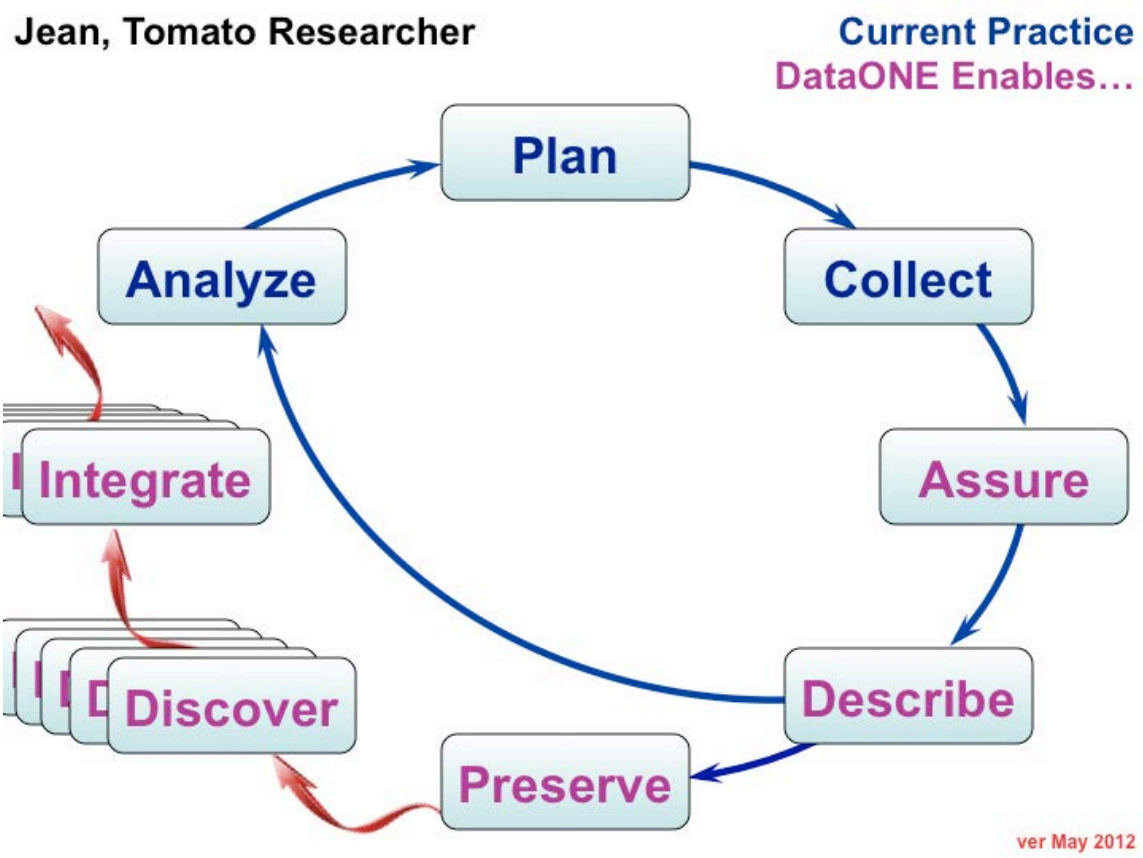
Jean does not normally share his pedigree book because it would not make sense to others, but freely distributes seeds to colleagues that ask for them. He considers these seeds to be data. When he receives seeds from others he "vets" the data by germinating the seeds and confirming the phenotype. He has hired a web developer to help visualize some of the collected data.

The assembled genomes he is willing to share immediately and thinks others should do the same. The transcriptome data are used to answer a biological question and thus are more sensitive. He would be willing to share the raw transcriptome data after publication, but does not want to be scooped in publications or proposals.

Repositories exist for genome data (e.g., GenBank), but not for raw phenotypic or raw sequence reads. Jean uses standard gene nomenclature to describe mutants, but feels unqualified to handle metadata.



Jean currently collects data for his own use. He does validate his data and describes it using the MEINS guidelines for metadata. Deposit is in the form of publications based on summaries and analyses; some of the data are shared, e.g., gene sequences in Genbank.



DataONE could provide tools to help Jean maintain his data in a consistent fashion over time. The motivation to use DataONE would be for better description of his data and for long-term preservation.

D1 usage scenarios

Project Planning

- *Management Planning:* Develops a project Data Management Plan following examples provided on the DataONE portal.

Comparison of current and DataONE-enabled practices:

Current data collection:
Jean collects phenotypic and genomic data.
DataONE enabled data collection:
No change.

Current data assurance:

DataONE enabled assurance:

Current data description:
Jean uses the MEINS guidelines for metadata for the genomic data, but does not describe the phenotypic data.

DataONE enabled description:

- *Training:* Learns how to use *Morpho* (a metadata management editor) based on instructional materials available in the DataONE Best Practices Database and associated downloadable video instructions.
- Helps develop an ontology for describing data to enable longitudinal analyses.

Current data preservation:
Deposits genomic data but no other long-term data preservation plans.
DataONE enabled preservation:

- *Data Preservation:* Deposits the data and metadata in a DataONE member node data repository for long-term preservation of the data.
- *Data Preservation:* Submits a research paper to a journal associated with Dryad—a DataONE Member Node. Upon acceptance, he submits the publication-relevant data, metadata, and model to Dryad where they are given a DOI (digital object identifier) and preserved in the Dryad repository.
- *Citation:* Upon publication, he adds the publication reference and the data citation (including DOIs for both; provided by Dryad and the journal) to her CV.

Current data discovery:
None.
DataONE enabled discovery:

- *Citation:* Another scientist working in Mexico on a similar study discovers the new publication and data created by Jean and cites him in his work.

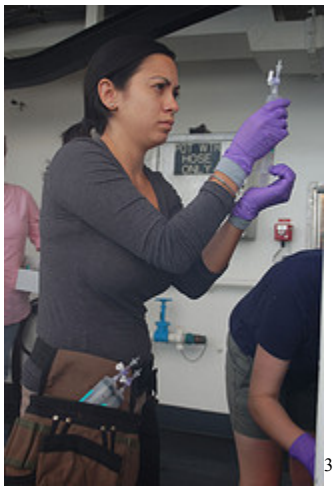
Current data integration:
None.
DataONE enabled integration:
None.

Current data analyses:
Uses standard desktop analysis tools.
DataONE enabled analysis:

- *Data Visualization:* Uses data analysis and visualization tools identified through DataONE Tools Database or available as part of the Investigator Toolkit to analyze existing data that he will use in his own research.

Data Visualization: Creates graphics using tools identified via DataONE.

- # Laura

 [3]

(Primary persona)

Source: Data Conservancy Laura persona by Anne Thessen; revised by Kevin Crowston

Tags: academic, non-university, mid career, multi-discipline, *field*, human and machine-collected data, experienced data management, *biology*

See also: Dr SA Cook D1 Scenario

Background

Name, age, and education

Laura is a tenured associate scientist at Woods Hole Oceanographic Institution. She received a PhD in marine biology from University of Rhode Island in 1990.

Life or career goals, fears, hopes, and attitudes

Laura's research investigates the effect of climate change on marine food webs, for which she needs to correlate environmental and species data. She is driven by intellectual curiosity and the need to publish and to obtain grant funding to support her research.

A day in her life

Laura has been funded by NOAA to do a series of cruises in the Gulf of Mexico, for which she goes to sea several times per year. She collects field data about the biology and chemistry of the water. During her cruises, she collects data on temperature, salinity, irradiance and fluorescence using instruments on board. Her research team collects water samples for later analysis in the laboratory (nitrate, nitrite, phosphate, ammonium, silicate and plankton counts, molecular). Some of these data are entered directly into a spreadsheet, but some are recorded onto printed data sheets. Much of her data comes off of an instrument and must be transformed to be useful. Each instrument and analysis has its own limits of detection and precision that must be taken into account.

One of the most time consuming aspects of her analysis is the plankton counts. Using a microscope she must identify and quantify plankton species. Plankton identification can be very

[3] https://www.flickr.com/photos/usoceangov/5081372308 (CC BY 2.0). Picture is of Stephanie Mendes, University of California at Santa Barbara.

tedious and there is a steep learning curve. Right now she uses a collection of books to make her identifications, but many of her taxonomic categories are used only within her lab, making it difficult to share these data broadly. These data are recorded onto a data sheet and must be transformed to be useful.

Laura's lab uses the Federal Geographic Data Committee (FGDC) metadata guidelines and the MERMAid tool for metadata management.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

NOAA requires that Laura upload her data into the NODC. She also uploads her data to the WHOI data store. She will send out data in spreadsheets to other researchers if asked and if she has the time to put them together. In some cases, such sharing has lead to interesting collaborations and she would be open to such opportunities in the future. Laura would gladly pay her data manager to upload her data into additional repositories as long as she could keep track of usage and gain citations. She would appreciate a chance to review and comment on the results of such use before publication, to receive copies of published articles and to ensure that her funding agency is acknowledged. It would be better if DataONE could streamline the process of sharing data. Finally, she is aware that other researchers have created derivative datasets based on her data (among others). While she believes this is an appropriate use of her data, she would like some way to assess the impact of this contribution and to receive credit for her work.

Laura would like to compare her field data with comparable field data and to historical data to identify trends, but she does not routinely download data from NODC because of poor usability. She would like to be able to go to one place, download all phosphate measurements made in the Gulf of Mexico (for example) and receive those data in a file, formatted to her specifications.

Intellectual and physical skills that can be applied

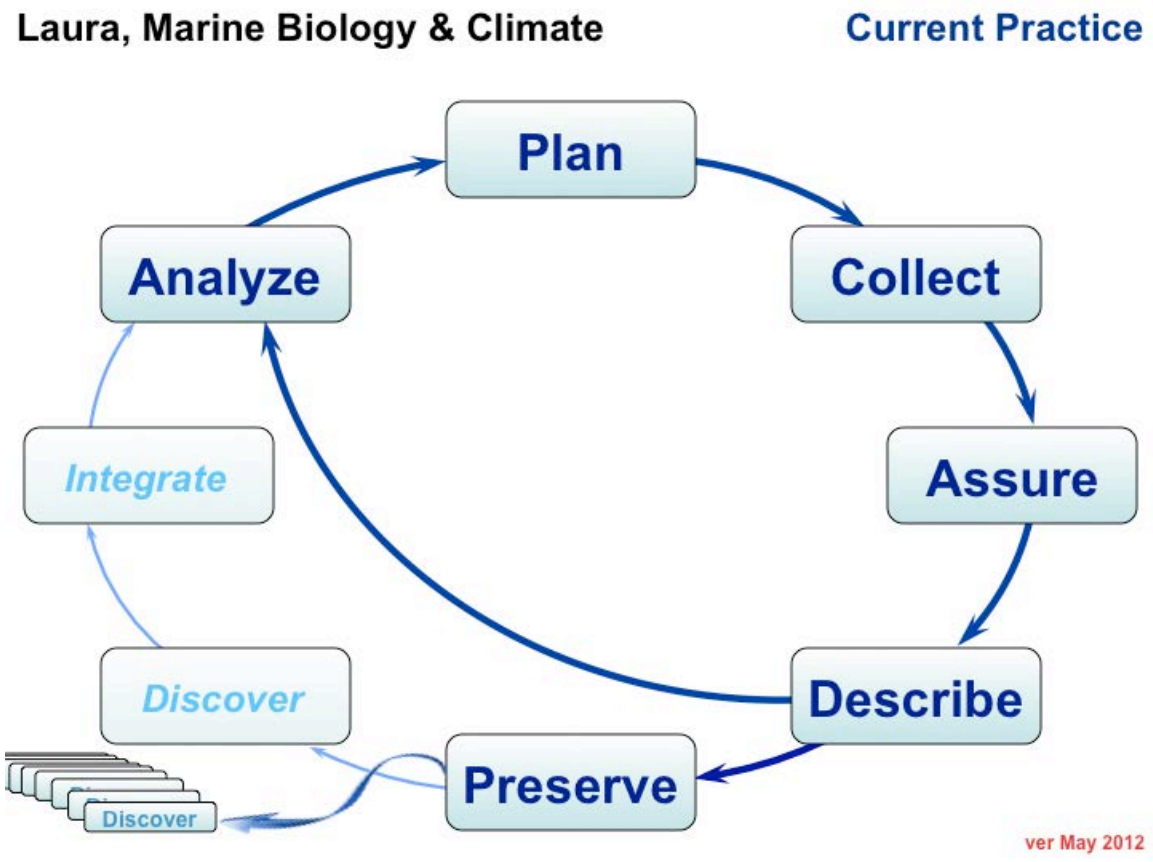
Laura is already adept at handling large datasets, including those handling abstract data fields which cannot be easily managed by hand. The process of compiling and analyzing her data can be complex; thus, she already understands that a crucial reporting requirement for sharing data is to also share the specific methods and subsequent data management steps taken to render the data amenable to analysis. She has first-hand experience with data of poor usability and she is keen to provide her data to a system and in a manner that does not suffer these same problems.

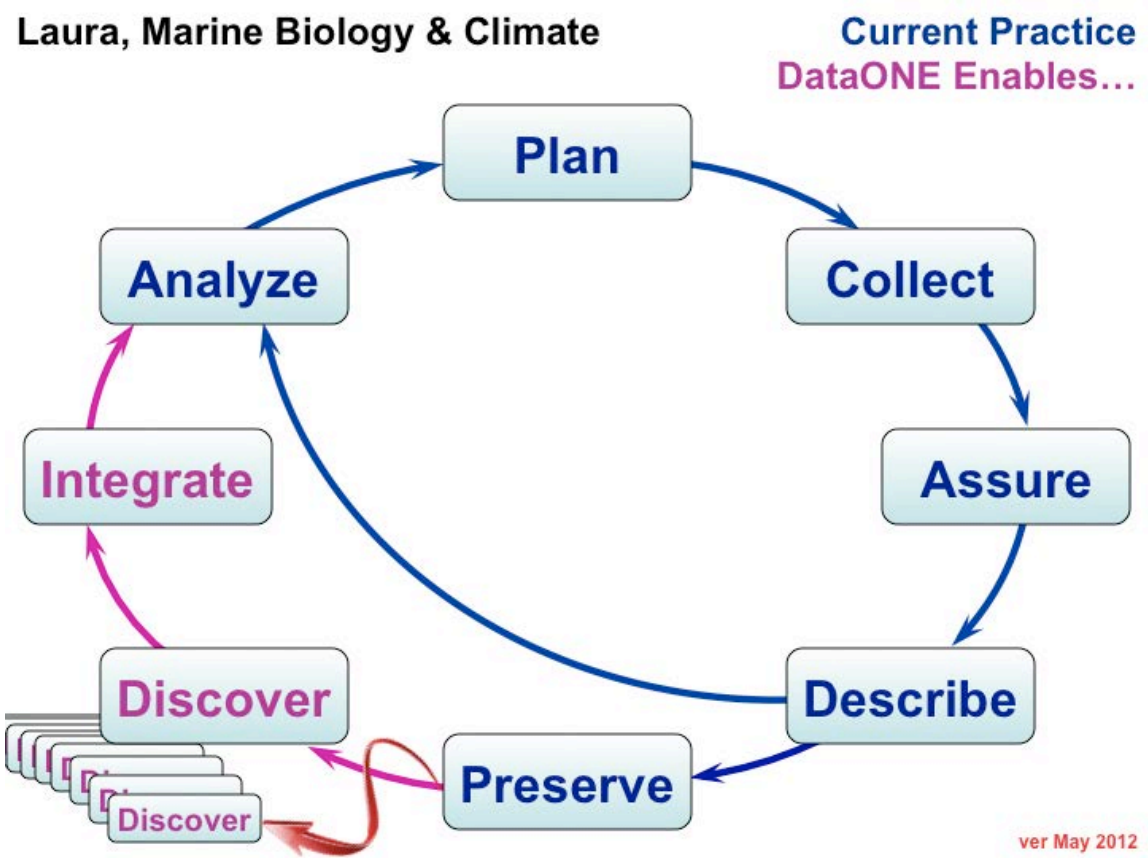
Technical support available

Laura employs a science diver who also acts as a data manager. WHOI also provides some data support. Her cruises are instrument and data heavy. Given the quantities of data involved, she must be able to utilize protocols which significantly automate the process of data deposition into DataONE affiliated repositories or else it will be too much work for her to contemplate.

Personal biases about data sharing and reuse (and data management more generally)

Laura accepts and takes seriously the obligation to upload her data. She needs a wide variety of data and believes that sharing data is important to be able to create a holistic view of the ecosystems she studies. She also believes that many of the data she needs for her own research likely already exist, albeit in inaccessible or unhelpful form. Part of her motivation to share data is to illustrate best practices and to encourage more people to develop tools and processes which might help in her own research.





Comparison of current and DataONE-enabled practices:

Project Planning

- *Management Planning:* Develops a project Data Management Plan following examples provided on the DataONE portal.

Current data collection:

- Field data.
- Digital and analog.

DataONE enabled data collection:

- No change.

Current data assurance:

- Undertakes basic validation steps as preparation for using her own statistical software.
- Must transform much of her data to analyze and interpret it.
- Her data manager handles most of these workflows. If they encounter errors in the batch sampling files, they have to discard all of those data since it might be reflective of instrument error.

DataONE enabled assurance:

- If DataONE tools can perform some of these validation steps more easily than is possible through the programs Laura currently uses, she would be willing to switch her workflow to managing the process using these tools. Absent that improvement, she might also be willing to use a workflow-management tool from DataONE, using a DataONE template for assuring that she has taken care of all the necessary steps for data deposition (including description and deposit), even though most of the actual work at each step is taking place outside of DataONE. Note that this possibility is especially attractive if adherence to the DataONE template guarantees that she will be able to seamlessly deposit her data into multiple member repositories with no additional requirements.

Current data description:

- Laura uses some field-supported standardized metadata schemas already.
- Laura also adds some custom fields for internal purposes, though she thinks these fields are likely to be useful to others as well.
- Other data description is peculiar to her lab, which hampers sharing data.

DataONE enabled description:

- As with data assurance, if DataONE provides tools for facilitating data description, including template support for standardized and customized metadata schemas, then she might migrate this part of the workflow to DataONE. Otherwise, this part of her workflow is not likely to be significantly impacted by DataONE.

Current data preservation:

- Laura and her data manager currently have to exert some effort to upload her data into existing repositories for her discipline. She feels strongly that this is an important activity and will continue to exert this effort regardless.
- *Data Preservation:* Collects data during summer research season and deposits the data in a data repository (a DataONE Member Node).
- She is interested in giving greater access to her data but does not know where else to put the data and cannot devote even more time to further data deposition.
- Laura presumes that her current activities are sufficient for long-term preservation, perhaps even ideal under current options.

DataONE enabled preservation:

- If DataONE can provide a single point of deposit for subsequent republication of the data into multiple data repositories, that would be a huge attraction for Laura. Note that this can occur via deposition directly into a member node, but then the data might flow through DataONE from that member node to other nodes. Alternatively, DataONE might provide a template or suite of tools for assurance/description which is guaranteed to streamline deposition of the data into any member nodes, which would also be very attractive for Laura. At a minimum, DataONE can provide guidance for choosing an appropriate repository.
- *Data Preservation:* Deposits the data and metadata in the LTER data repository.
- Laura wants to see evidence that her data is being used and is having an impact. If DataONE provides that evidence, Laura would be willing to spend even more time and

effort on data deposition.

- *Citation:* Another scientist working on a similar study discovers the new publication and data created by Laura and cites her in his work.
- Laura currently feels that her data management processes are sufficient for long-term preservation, but if DataONE enhances this offering in some way, that is of interest and could be a motivating factor.
- *Data Preservation:* Submits a research paper to an ecological journal associated with Dryad—a DataONE Member Node. Upon acceptance, she submits the publication-relevant data, metadata, and model to Dryad where they are given a DOI (digital object identifier) and preserved in the Dryad repository.

Current data discovery:

- Laura is already interested in other datasets, both to enhance her own research (i.e., put it into context) as well as to identify possible collaborative opportunities. But she has had limited to no success to date identifying collaborators via data of shared interest, nor in identifying other useful datasets. It is her sense that new tools are needed, and that existing repositories are not user-friendly enough to be worth her time.
- Laura will occasionally identify possible collaborators based on paper she reads or talks she attends, but she has not found any good way to find their existing data to ascertain the potential for collaboration, and she is reluctant to spend the time contacting these people and building trust when she is not sure if there is any real collaborative opportunity.

DataONE enabled discovery:

- If DataONE makes it easy to discover datasets of interest, Laura is likely to become a power user
- Laura is likely to search for datasets by authorship, paper-association, etc. Easy faceted search capability is going to be a key feature of DataONE.
- If the system can use Laura's own data as a point of reference for making automated recommendations of possibly related datasets, that would complete the data cycle for Laura and drive her interest in getting her data into the system.
- *Data Discovery, Access, Use and Dissemination:* Acquires supplemental data from another DataONE Member Node with complete citation information.
- *Data Discovery, Access, Use and Dissemination:* Identifies relevant data and downloads data and metadata from previous LTER studies as well as data collected by state and Federal agency scientists) (i.e., non-LTER).

Current data integration:

- Laura understands the logic of data integration, but she does not currently do any integration with datasets that she did not collect herself.

DataONE enabled integration:

- If DataONE provides a toolset for integrating disparate datasets, Laura is likely to become a power user.
- Another design option for DataONE is to enable data transforms and annotations during the process of discovery so that downloaded data already exhibit some of the necessary characteristics for integration (using some third-party tool).

- Laura will want to be able to (re)publish the integrated dataset so that others can build on her work and insights. It's not yet clear what attribution expectations she might have for these derivative datasets, nor how much work she would be willing to do to attribute the original data contributors.

Current data analyses:

- Laura analyzes all of her existing data using standard statistical packages, using data stored on her own hard drive.
- Upon completing her analyses, she typically only captures and publishes the statistical results (and relevant summary variables) due to the lack of any appropriate place to publish more information or visuals.
- In papers and publications, Laura typically provides a link to the entire dataset, if it has been deposited into a repository at that point. She does not provide (and does not know how to provide) direct links to the specific subsets of the data relevant to any analysis.

DataONE enabled analysis:

- If DataONE provides robust analytical tools, there is a good chance that Laura will leverage those offerings. However, Laura is more likely to export any data of interest to perform the analyses using her current methods and tools.
- She would be even more interested in analytics within DataONE if the DataONE system can track and record the specific methods of analysis, enabling direct reference to the subset of the data used for each analysis and also enabling anyone to reconstitute the analysis (step by step, ideally) to examine the assumptions, transformations, and other aspects of her thinking. This functionality could be provided as a link to an executable script which also serves as a sort of surrogate for attribution.
- *Data Visualization:* Creates graphics using tools identified via DataONE.

Data Visualization: Uses data analysis and visualization tools identified through the DataONE Tools Database or available as part of the Investigator Toolkit to analyze existing data and develop initial model parameters that she will use in her own research.

Andreas

 [4]

(Primary persona)

Source: Data Conservancy Zoe persona by Anne Thessen, interview with David Keller and comments from Raleigh Hood and Sam (DataONE scenario). Revised by Kevin Crowston

Tags: Academic, university, mid career, multi-discipline, experimental, human and machine-collected data, experienced data management, agriculture

See also: Dr SA Cook D1 Scenario

Background

Name, age, and education

Andreas is a biogeochemical modeler at Michigan State University with a PhD from the Max Planck Research School for Global Biogeochemical Cycles, received 16 years ago.

Life or career goals, fears, hopes, and attitudes

Andreas has been a modeler throughout his career. For Andreas, the coming flood of data and the growing numbers of analytical and visualization tools are extremely exciting and he seeks ways to stay at the forefront of this rapidly moving field.

A day in his life

Andreas has written several models of various complexity. Right now he is in the final stages of developing a model that can predict plankton dynamics in Tolo Harbour, Hong Kong using nitrogen units. He stores his model runs on a server at his institution. Each run is saved in a folder (named with the date and a runID) as a NetCDF file and sometimes with a text file of notes on the run. However, some of his older output files are csv. Visualizing the model output requires sending the NetCDF or csv file to MatLab. Andreas assesses his model by comparing model output and real-world monitoring data on two axes: time (day of the year) and location. He performs a Model Skill Assessment to assess the accuracy of his model statistically in addition to graphical examination. It is the MSA that others will use to judge the quality of his model.

[4] https://www.flickr.com/photos/lowercolumbiacollege/4464497223 (CC BY-NC-ND 2.0)

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

To assess his latest model, Andreas wants to be able to compare his prediction of maximum spring bloom biomass with what was actually observed. Monitoring data are gathered by the government and are available online, but significant work must be put in to the dataset before it can be used. If DataONE could provide the data in a more usable format, it would save considerable work.

Like most modelers, Andreas hopes that his models can elucidate both the specific biogeochemical dynamics for the area in question as well as be applicable to comparable systems elsewhere. Currently, too much of his time is focused on acquiring and managing the data in the specific context of his research, and he cannot afford to test the applicability of his model to other systems. DataONE could potentially solve that problem, expanding Andreas' research capabilities and revealing his work to a broader base of researchers.

Intellectual and physical skills that can be applied

Andreas has significant programming knowledge in MatLab and Fortran. He is likely to be able to overcome functional deficiencies in DataONE tools as long as he doesn't have to spend too much time cleaning up the datasets themselves. Andreas' work is likely to illustrate some of the more powerful analytical capabilities gained from sharing datasets via DataONE, but only if he uses the data referencing protocols so that users can track those links to his models.

Technical support available

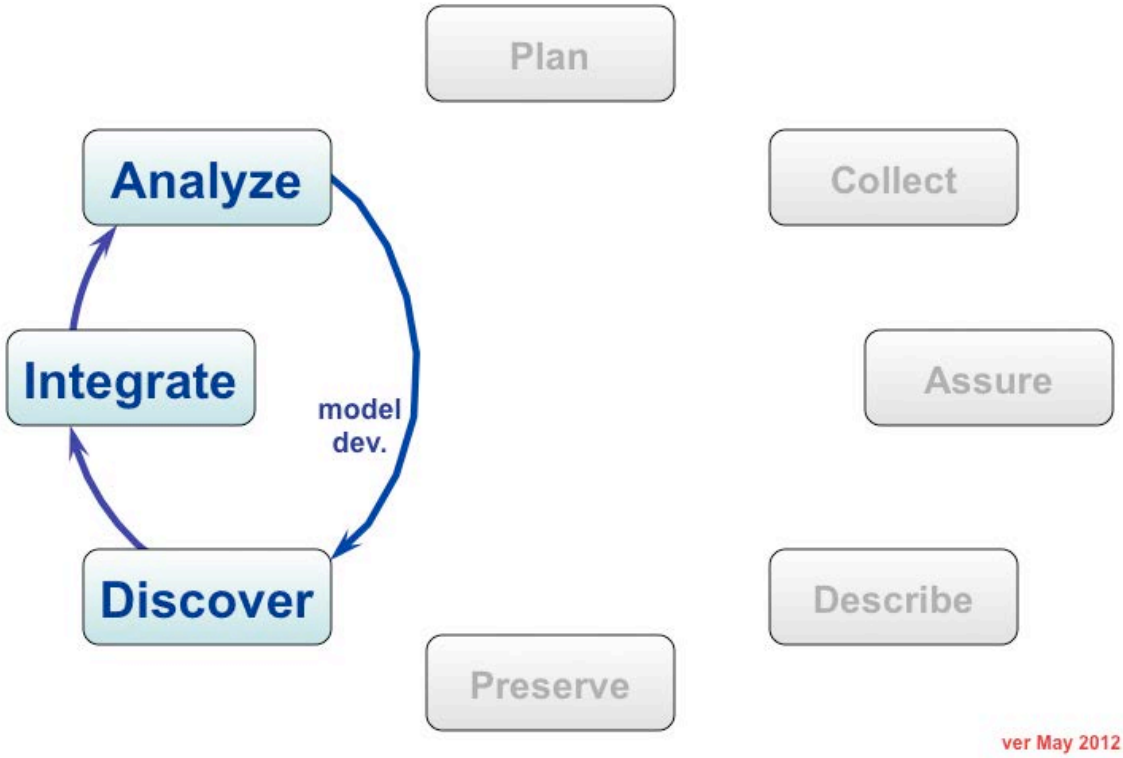
Andreas is part of a highly sophisticated technical community, with whom he can work both formally and informally. However, he has no additional technical support for his own work beyond himself.

Personal biases about data sharing and reuse (and data management more generally)

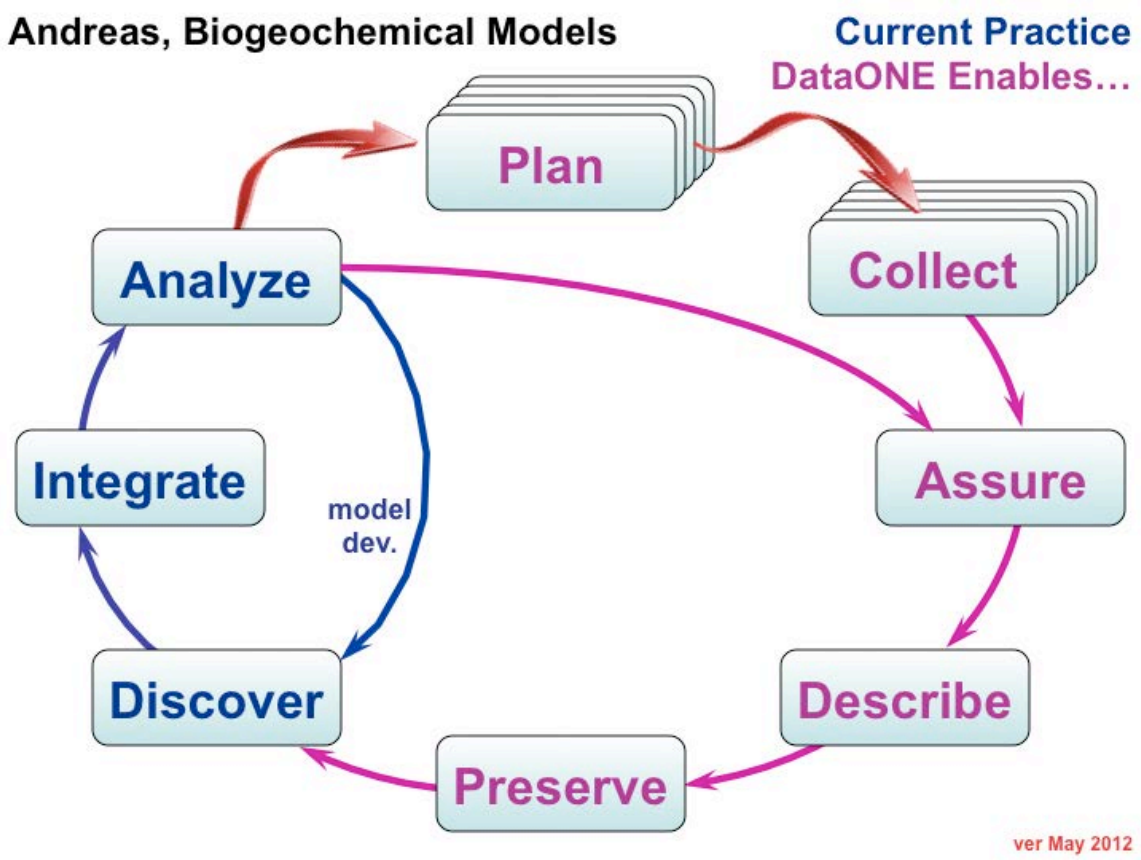
Andreas needs real world data about his area of interest, both to calibrate and to assess his model.

Some of Andreas’s model code is publicly available and open source and he has already published on most of it. However, he is confused as to what he should do with his model output. Normally it just sits on his servers and is used for a couple years by him alone. Other researchers have asked to look at his code, but no one has asked about his model output. Andreas doesn’t think the older model output is very useful, in part because it is difficult for anyone other than him to understand, and would like to delete it to free up server space for new output, but worries that he might be losing.

Andreas, Biogeochemical Models **Current Practice**



ver May 2012



Comparison of current and DataONE-enabled practices:

Current data collection:

- None

DataONE enabled data collection:

- No change.

Current data assurance:

- Andreas does not deposit data, so he has no assurance steps for his own data. However, he does worry about how much he can trust the data he uses to validate his model, though he does not have many tools to perform these validation steps himself, instead trusting the checks on data uploads to the repositories he uses.

DataONE enabled assurance:

- Andreas may find it helpful to use DataONE tools to double-check the validity of interesting data in the system.
- He may also use templates provided by DataONE as guides for assuring the quality of any aggregated datasets he generates and republishes, and also perhaps for any model outputs he wants to contribute.

Current data description:

- Andreas does not deposit or retain any data once he has completed a model run. The only

information he needs is a reference back to the publication of the original data.

DataONE enabled description:

- To the extent that Andreas publishes aggregated datasets he creates, he will likely benefit from some guidelines or tools to manage the process of describing his data. This will be especially important for Andreas because it is otherwise unclear how to apply useful metadata to a dataset which is itself composed of heterogeneous subsets of other datasets.
- Andreas already understands how important it is to describe data well since his work relies on being able to discover data relevant to his research questions. This implies that he may be likely to devote more time than most scientists to data description activities as part of his reciprocity to the DataONE community.

Current data preservation:

- Andreas does not deposit data currently. He wishes everyone else would though.
- Andreas does not preserve his data. However, his work depends on finding and analyzing existing data which remain accessible and identifiable with a stable identified (URI or DOI) so that peers can evaluate his models correctly.

DataONE enabled preservation:

- To the extent that Andreas can and wants to deposit his aggregated datasets, DataONE may be a key facilitator. Andreas' greatest motivation is likely to be a sense of obligation to "give back" to the community of data depositors upon which he depends for his work.
- Andreas is not so interested in data preservation except to the extent that others preserve their data. Nonetheless, he would be happier contributing his data to a repository if he knew that preservation services were sound and the data useful.

Current data discovery:

- Andreas spends an enormous amount of time searching and discovering data that are relevant for his models. This process is so difficult that he usually limits himself to government data repositories where at least the data are likely to achieve a known standard, are presented in some standard form, and have no restrictions on use
- Because Andreas must rely on relatively "raw" data due to limited availability of data in other forms, he also spends a lot of time transforming the data to meet his analytical needs and the limitations of his models. These processing steps are captured in his methods but are not otherwise recorded or automated in any manner.

DataONE enabled discovery:

- If DataONE can save Andreas time in discovering useful data, as well as reveal additional datasets of interest, he will be a power user very quickly.
- Andreas believes that the processing steps he takes to render raw data suitable for subsequent analysis are valuable, but he has never had any place to deposit such datasets, and there are no guidelines regarding norms of attribution and other variables. If DataONE can accept these datasets and automate the provenance and attribution aspects, this is likely to convert Andreas from a strict consumer of data to a data contributor (or, more precisely, an enhancer).
- Andreas is also curious to know who might find his datasets and analyses to be of interest, as he is interested in collaboration but has lacked clear pathways to identifying

possible collaborators outside of his tight circle of peers. DataONE may open up those possibilities for him.

Current data integration:

- As with data discovery, Andreas currently spends huge quantities of time integrating datasets, where possible, for his analyses.
- Andreas also limits the potential of his work by intentionally steering clear of problems that require complicated integrations. It is simply too difficult to manage the process and takes too much of his time for the payoff.

DataONE enabled integration:

- If DataONE provides a toolset for integrating disparate datasets, Andreas is likely to become a power user.
- Andreas sees integrated (aggregate) datasets as potentially valuable contributions of his work. If DataONE can accommodate deposition of such datasets, with as little fuss as possible, Andreas will probably become a major contributor.

Current data analyses:

- Andreas' analyses consist of model runs, using the data he discovered and integrated. Different models produce different outputs, and it is not clear that the outputs are of interest to anyone outside of their utility in validating the model.
- Andreas labors under the presumption that his models may be useful for answering empirical questions, but he isn't personally involved in such efforts, though he often wishes he could be.

DataONE enabled analysis:

- It is likely that Andreas would still export his integrated datasets in order to test his models outside of DataONE. However, DataONE may provide the means for him to refer to different model runs and the associated data more easily, and also to tweak the variables and rerun models more easily, tracking those changes each time.
- For some models, the outputs might actually be of interest, and reference to the original data, as well as the model itself, would make such outputs easier for other people to interpret and build on. DataONE would need to provide functionality along these lines in order for Andreas to consider depositing the model runs in this manner.

William

 [5]

(Primary persona)

Source Based on Data Conservancy Pedro persona by Anne Thessen: Comments from David Patterson; revised by Kevin Crowston

Tags Academic, university, late career, single discipline, field, human-collected data, novice data management, plant taxonomist, international

Background

Name, age, and education

William is a plant taxonomist working in the University Herbarium at University of Michigan. He is 68 and is looking forward to retiring after a long and productive career.

Life or career goals, fears, hopes, and attitudes

William is nearing retirement. He has an office full of paper, photographs, and pressed plant specimens that represents his life's work. If he doesn't move the contents of his office after he retires to the limited space in his basement at home, the university will throw it out. He does not want this to happen, but doesn't know how to stop it. Some of his data have been published in monographs that are accessible in a few libraries around the world, but much of it is not associated with a publication. William would like to make the contents of his office, his life's work, available to early-career taxonomists who could (potentially) put it to good use.

A day in his life

William no longer does much field work himself, but he has a wealth of data from his career: collections of species occurrences, measurements and images (20–30,000 35mm slides). Most of the data are in the many field notebooks he has amassed over his lifetime, include daily notes about where he was and when, indications of pictures taken, collector's numbers of the specimens he has collected, and descriptions of habitats visited, including comments on soils, local distribution, species abundance, and phenology of species not collected. His locality notes for each collected specimen are recorded on the label for each specimen. The rest of the comments in

[5] https://www.flickr.com/photos/sevendipity/4487712789 (CC BY-NC 2.0). Picture is of Charles Stirton.

his notebooks are habitat descriptions and organism occurrence observations that are not tied to collected specimens. Little of the data has been digitized: the species data are in an Excel spreadsheet and some of the images are stored on a portable HD (500GB); he has no system for annotating these other than folder and file names. He collaborates regularly with colleagues in Spain who are interested in the same species.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

William would like to have help digitizing his data and a place to put it where it will be used and where he could get credit for it and could see how it is being used. There is a plant taxonomy web site where some of his colleagues have uploaded data, but functionality is limited and it is hard to find. He would prefer to register his data, so people would know about it, but he wants people to ask for permission to use it. This way he could prevent misuse or misinterpretation of his data.

Intellectual and physical skills that can be applied

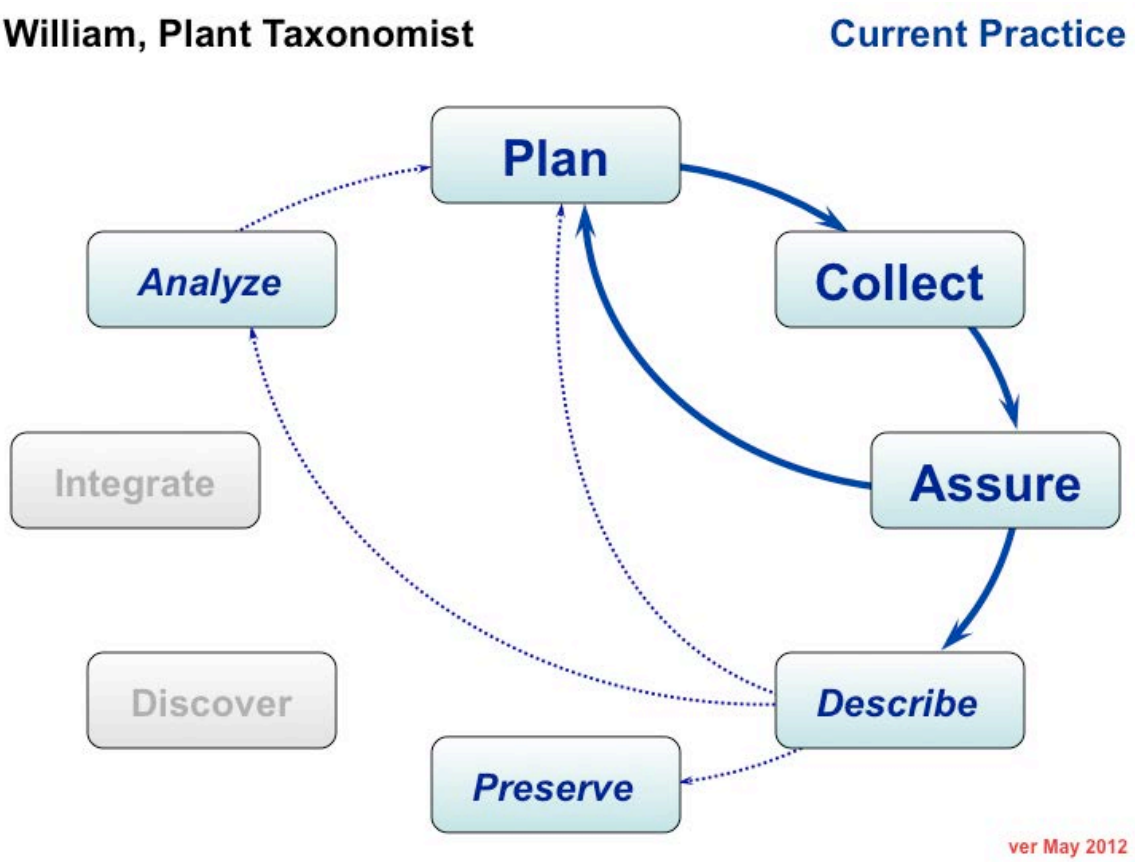
William knows how to use a scanner, but is discouraged by the amount of time it would take to scan the contents of his office. On the other hand, if he only needed to invest a little time beyond the actual act of using the scanner itself in order to deposit his data at a DataONE member node, he would probably start chipping away at his collection.

Technical support available

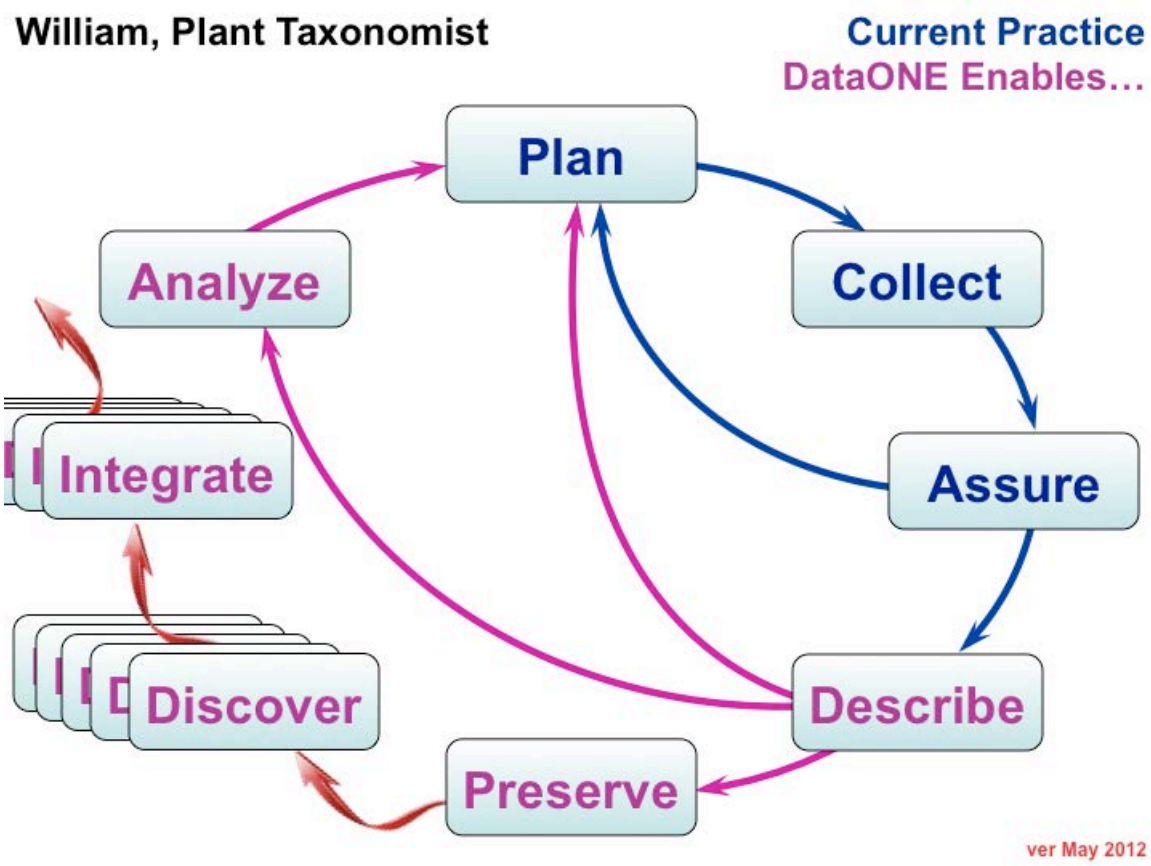
William has little to no technical knowledge and little technical support. He is also unsure how to go about identifying possible sources of support for such activities.

Personal biases about data sharing and reuse (and data management more generally)

Williams is generally suspicious of raw data sharing. He is more accustomed to publishing his findings in monographs. He is convinced that there is enormous value locked up in the contents of his office, but he is not sure who is best positioned to realize that value or how to do it.



William has obviously been a prolific collector; he has his own processes that have worked over the years for data assurance. Description has been dependent on whether there have been other demands for the data, e.g., for publication. He has a desire for his data to be preserved for future use, but limited motivation due to some suspicion about digitized data and no real knowledge of how to go about the large amount of work it will be to deposit the data in a repository.



DataONE has the potential to inspire William to begin the arduous task of processing the mountain of paper that is his office, making parts of it available through a data repository. When the graduate students who are working with him in his current data collection become aware of DataONE and the plant taxonomy data that has already been deposited there, they will start to understand the gaps that could be filled in current knowledge by digitizing, describing, and depositing some of William's specimens and notes. With the help of the data librarian at the University of Michigan they can help William map out a plan for evaluating and describing his collection of slides and field notes.

Comparison of current and DataONE-enabled practices:

Current data collection:
Collects plant taxonomy data (field notes, photographs, specimen slides).
DataONE enabled data collection:
No change.

Current data assurance:
Validates data using own standards.
DataONE enabled assurance:
Data could be assured using standard tools as part of a digitization project.

Current data description:
Data has been described where published.
DataONE enabled description:

- *Training:* Graduate students working with William learn *Specify v.6* to describe William's specimens using Darwin Core.

Current data preservation:
None.
DataONE enabled preservation:

- *Data Preservation:* Graduate students deposit data and metadata in the University of Michigan (UM) data repository
- *Data Preservation:* Preservation functions of the UM repository are enhanced by acceptance as a DataONE Member Node

Current data discovery:

- Does not use other researchers' data.

DataONE enabled discovery:

- *Data Discovery, Access, Use and Dissemination:* Other researchers discover and use William's data through DataONE.

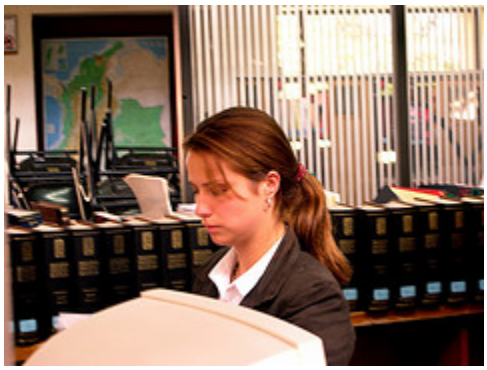
Current data integration:

- Does not use other researchers' data.

DataONE enabled integration:

- *Data Discovery, Access, Use and Dissemination:* Other researchers discover combine William's data with their own for more complete coverage of the region.
- *Citation:* Combined datasets are published and William's data is cited.

Abby


[6]

(Primary persona)

Source: Written by Kevin Crowston based on interviews with Gail Steinhart (Cornell), Lynn Yarmey (Stanford) and Jacob Carlson (Purdue)

Tags: data librarian

See also Tenzig Norgay D1 Scenario

Background

Name, age, and education

Abby is a data librarian at University of California, Berkeley. She studied earth sciences as an undergraduate at Purdue University, then entered a graduate program at Berkeley. She earned a master's degree but ended up moving to a staff research associate position in the Berkeley Seismological Laboratory when the funding ran out on the grant that was supporting her. In that position, she found herself taking on more and more responsibility for data management for various projects. After interacting with librarians developing the campus digital repository, and with their encouragement, she decided to pursue a masters in library and information science at San José State University to further develop her skills and knowledge in data management. Upon graduating, she was hired as the University library's first science data librarian.

Life or career goals, fears, hopes, and attitudes

Abby finds librarianship and the library a good fit for her skills and interests. She is hopeful that academic libraries will build and maintain a role in data management across the campus, enabling her to make a career in this position. She does sees her role shifting over time from helping individual researchers to developing programs and services that are generally useful and perhaps expanding her responsibilities as more people move into data management ("someday there may be an assistant dean of the library for data"). But to get to that point requires learning more about researchers' needs and expectations, which is her current focus.

A day in her life

Abby spends most of her day interacting with researchers. She finds their research and the data management issues inherently interesting, though she is sometimes challenged by unfamiliar kinds of research and data. In this way, her current job is an extension of her interest in science,

[6] https://www.flickr.com/photos/stephen_downes/206250405 (CC BY-NC 2.0).

just by other means. Abby provides support that is similar to traditional library reference, such as guiding researchers to useful data resources and providing instruction in use of data tools. She also helps researchers develop data management plans by helping them think through the kind of data they have and the issues in managing it. Funding is often written into grants to support her work in carrying out those plans. She is also partnering on larger grants with researchers who are developing data resources, e.g., a data portal.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

Abby is hopeful that DataONE will provide tools that will help the researchers with whom she works, most of whom are not data managers nor interested in becoming one. She maintains a list of repositories that are relevant to her researchers and will add DataONE and guide researchers to use it when it is operational. She would like resources to support instruction in use of data and data tools as well as guidance about data sharing standards. For example, groups struggle with issues such as deciding what data are worth depositing and developing consistent metadata—if DataONE can provide help in these areas, it will be welcomed.

Intellectual and physical skills that can be applied

Abby brings numerous diverse skills to her position. She brings traditional librarian skills such as knowledge of the subject area data resources, and an ability to assess researchers' needs and match them to the resources. Through a data interview (parallel to a reference interview), she can help researchers think through their data and its management issues. She has skills in information organization and knowledge of good ways to format data and of controlled vocabulary. Perhaps her most important skills are in personal interaction, being able to listen and her flexibility, such as the capability to interact with multiple granularities of data at the same time, from DataONE and the institutional repository, and then with a lab with single laptop.

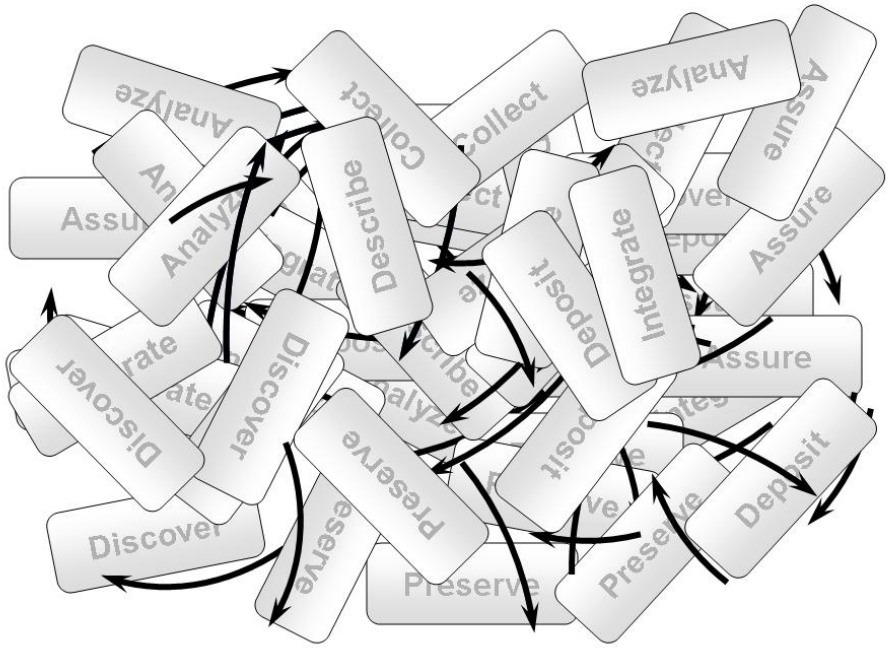
Technical support available

Abby has some technical skills, but relies on the library and campus information technology departments for system support.

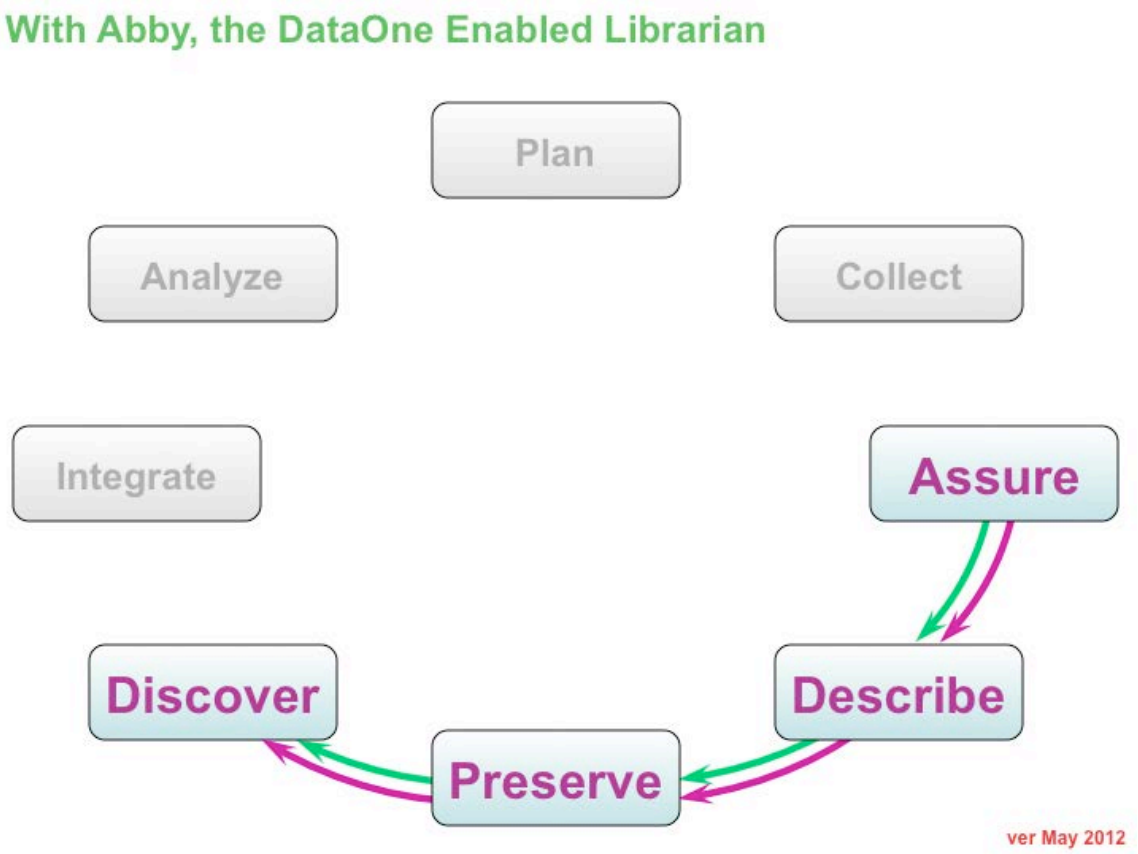
Personal biases about data sharing and reuse (and data management more generally)

Abby is a strong believer in the importance and value of data sharing and reuse and has always worked with scientists who share that belief. However, she is aware that researchers need to receive recognition and credit for their work to be successful in their careers and that sharing data is only a viable option if it provides that. She is interested in learning about other ways to recognize and assess the impact of data sharing so she can provide that information to her patrons. Furthermore, she knows enough of the science to realize that the value of data is tied to the scientific understanding behind the data, so it is critical to share the later to make sharing the former sensible.

without Abby, the DataOne enabled Librarian



Abby's current work with researchers is on an *ad hoc* basis as needs for data management arise in the progress of individual research projects. She is developing her own expertise in a number of steps in the data lifecycle, especially in the methods of data description and discovery of relevant and useful data. She is helping the researchers she works with to discover relevant metadata schemas and tools for creating that metadata. She can also assist them in finding appropriate repositories for their datasets, with some education on the importance of data preservation added to the mix. Her interventions on the data discovery side are making researchers aware of a multitude of data resources that can be valuable to their individual research efforts. What Abby often finds lacking in her work is a way to connect one step to another, to create a comprehensive program of educational modules that could help researchers proceed through the steps of managing and discovering data in a more methodical and seamless manner.



DataONE enables Abby to much more easily help the researchers she works with "connect the dots" in their data management. Best practices documents for Data Management Plans and for data assurance methods—available through DataONE—enable her to intervene in the project planning stages so that the describe, deposit, and preserve steps can be planned for in advance and can flow throughout the project. Tools offered through DataONE for metadata creation and for data deposit into Member Nodes make it easier for Abby's researchers to become proficient on their own in these steps. In the data discovery area, DataONE offers Abby a powerful, "one-stop shop" resource to offer researchers for data discovery in a broad range of earth and environmental science disciplines. Abby can turn to DataONE for help in keeping her list of data repositories current. DataONE helps Abby to demonstrate to researchers the importance of data preservation as they see the link between their discovery and reuse of other data and the possibilities for the reuse of the data that they deposit. The citation and tracking tools provided by DataONE are also powerful motivations factors for Abby's researchers to engage in the data deposit and preservation steps.

Comparison of current and DataONE-enabled practices:

NB - It doesn't make much sense to discuss prior practices regarding the data lifecycle since they are not likely to have involved a librarian in any meaningful manner. Therefore, we only describe the DataONE enabled practices here.

D1 enabled Project Planning

- Data curation and metadata management: Develop data management plan for grant submission -- Use DataONE's resources for preparation of project data management plans.

DataONE enabled assurance:

Abby offers training on DataONE-recommended data quality tools

DataONE enabled description:

- Data interoperability, standards and integration: Ensures compliance with DataONE's standards and best practices.

DataONE enabled preservation:

- Data discovery, access, use and dissemination: Data Librarian works with faculty to expose and publish data.
- Data protection: Data Librarian will advise on intellectual property issues and other use rights considerations associated with DataONE.
- Data deposition/acquisition/ingest: Works with faculty to ensure proper deposit, proper description, appropriate versions of data.
- Data curation and metadata management: Implement guidelines for selection and/or sampling of longitudinal data.
- Data curation and metadata management: Data Librarian provides guidance on usage of DataONE infrastructure.

DataONE enabled discovery:

- Data curation and metadata management: Once data are deposited with a member node, the data Librarian delivers usage statistics to faculty on the use of their dataset.
- Data discovery, access, use and dissemination: Data Librarian identifies relevant resources in DataONE portal.
- Data curation and metadata management: Collection development (selection, deselection, collection rescue).
- Data interoperability, standards and integration: Optimize collections for ease of use in the virtual research environment (for use by NEON, LTER, etc. (designated communities)).
- Data deposition/acquisition/ingest: Build reference collections to support data use, discovery, integration, etc.

DataONE Community Activity

- Data interoperability, standards and integration: Coordinates member node activities for the campus.

Tina: Citizen Science Project Leader

 [7]

(Secondary persona)

Source: DataONE PPSR Working Group: Andrea Wiggins & Sandra Henderson

Tags: citizen science project leader

Background

Name, age, and education

Tina is a 42-year old science educator who runs a citizen science project in Portland, OR. She works for Multnomah Nature Center (MNC), a 36-year old nonprofit organization focused on natural history education and conservation research. MNC is funded by private donations, development activities such as fundraising events, and an allocation from the Oregon Lottery. Tina has a dual MS in Science Communication and Natural Resources from Oregon State University, and a BS in biology from Michigan State University.

Life or career goals, fears, hopes, and attitudes

Tina has been working for MNC for 12 years. She started as a field technician doing plant inventory, but when a position was created for science education through program expansion from a family foundation grant, she was successful at demonstrating how her field skills and science communication background combined for effective outreach to multiple age groups. Eight years ago, MNC decided to start a citizen science project named "Multnomah Weed Watchers" focused on invasive species monitoring at the nature center and in adjacent public lands and Tina was tasked with organizing this project, in addition to providing support for visitor programs, including school groups. She is enthusiastic about the dual benefit for both MNC's conservation work and involving MNC and community members in scientific research.

---

[7] https://www.flickr.com/photos/neonzu1/5610014543/ (CC BY-NC-ND 2.0).

A day in her life

Tina's work involves developing training materials and a website for data entry and reporting for both researchers at MNC and project participants. On a daily basis, she reviews recent observation reports, manages the project email and phone communications, and develops recruitment materials. She arranges quarterly training sessions for new volunteers as well as a volunteer mentor pairing program that matches experienced volunteers with new recruits to help ensure that initial training is supported and skill development can continue, and to provide social support for ongoing participation.

Tina knows that there are other invasive species monitoring project in the state and in the Pacific Northwest that may make good partners for extending the reach of the program and increasing the value of the data being collected. In the past, she has had summer interns helping with database development, marketing materials, and development of educational resources, both for ongoing project participants and for school groups that participate on a one-time basis. However, she does not have regular support beyond the volunteer resources that she can coordinate herself, and she is always strapped for time to fully implement project plans.

The project's target monitoring species include Tansy ragwort (*Senecio jacobaea*), Kudzu (*Pueraria lobata*), Giant hogweed (*Heracleum mantegazzianum*), Spotted knapweed (*Centaurea maculosa*), and Dalmatian toadflax (*Linaria dalmatica*). All of these species are classified as "T" weeds, meaning that they are considered an economic threat to the state of Oregon. This makes the project compelling to diverse groups within the region who are interested in preserving farmlands and public natural spaces, and individuals who have strong environmental conservation interests. Project participants include senior citizens who choose their own monitoring sites, school groups who participate at the nature center and their schools, a singles organizations whose members volunteer for monthly surveys at several local parks, and an assistant professor at Oregon State University along with his graduate students.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

Tina has just heard about DataONE from her Oregon State colleagues. She thinks this might be a resource that would help her find out how to start improving project data management. She is excited about the potential to share her data and make it more usable for research. Because she knows very little about data management, she finds most of the materials bewildering, and therefore has a hard time understanding how DataONE can help her.

Intellectual and physical skills that can be applied

MNC is very supportive of the project and showcases it regularly in annual reports and media releases. However, there is no dedicated funding in the annual budget to support the project, beyond covering Tina's salary. Tina therefore spends a disproportionate amount of her time seeking sources of funding to cover basic operational costs for the project. Although MNC has a Development Officer, that person's focus has been on raising operational funds for whole centre,

so she can only provide advice and pointers to resources that Tina then has to pursue.

Technical support available

Tina has taught herself basic HTML and SQL database management over the years so that she can support the project's online data entry functionality, which was initially set up by a short-term web development volunteer. She is an "accidental techie" and struggles to keep up with the ever-growing task list as expectations for technology sophistication continue to increase. The MNC's Marketing Director manages the organization's website, but she has no time or resources to offer Tina to help support the project.

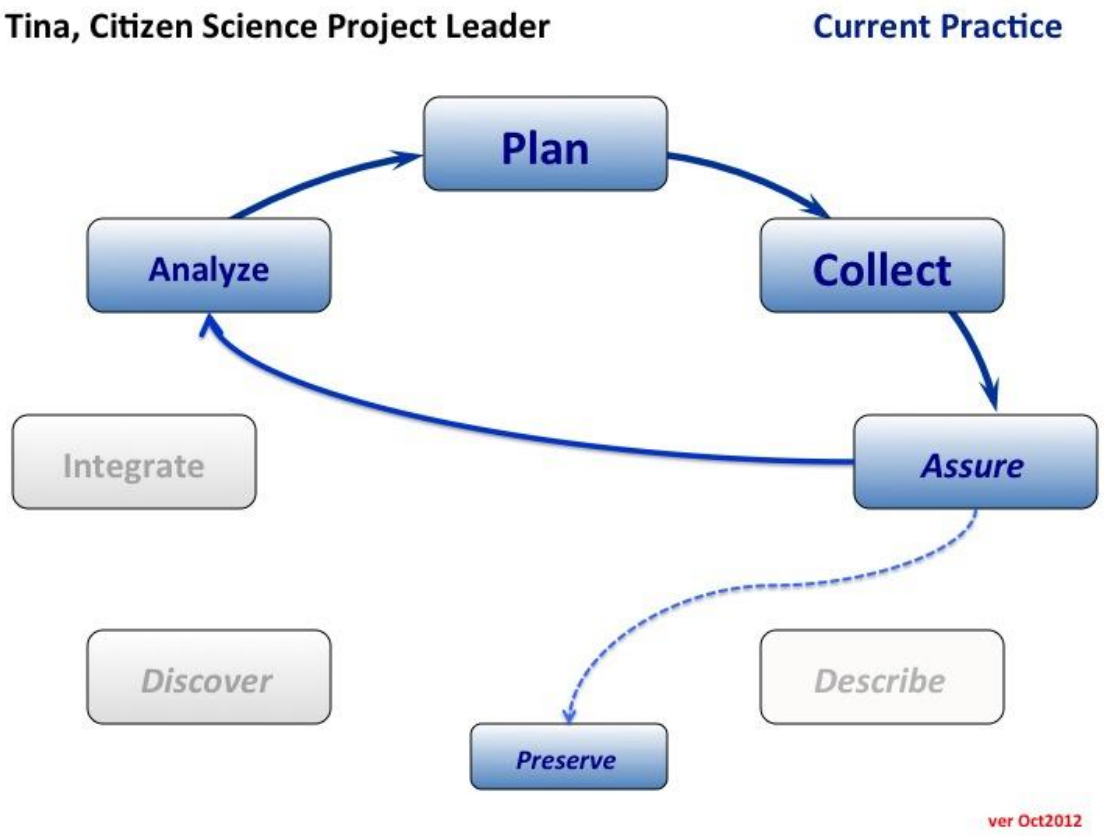
Because location of the invasive weeds is important, she has worked with a work-study student to enable Google Maps location resolution for online data entry, as well as GPS coordinates. With a small grant to support public communication of science, she also hired a developer to create a simple mashup map of locations for each species sighting.

Personal biases about data sharing and reuse (and data management more generally)

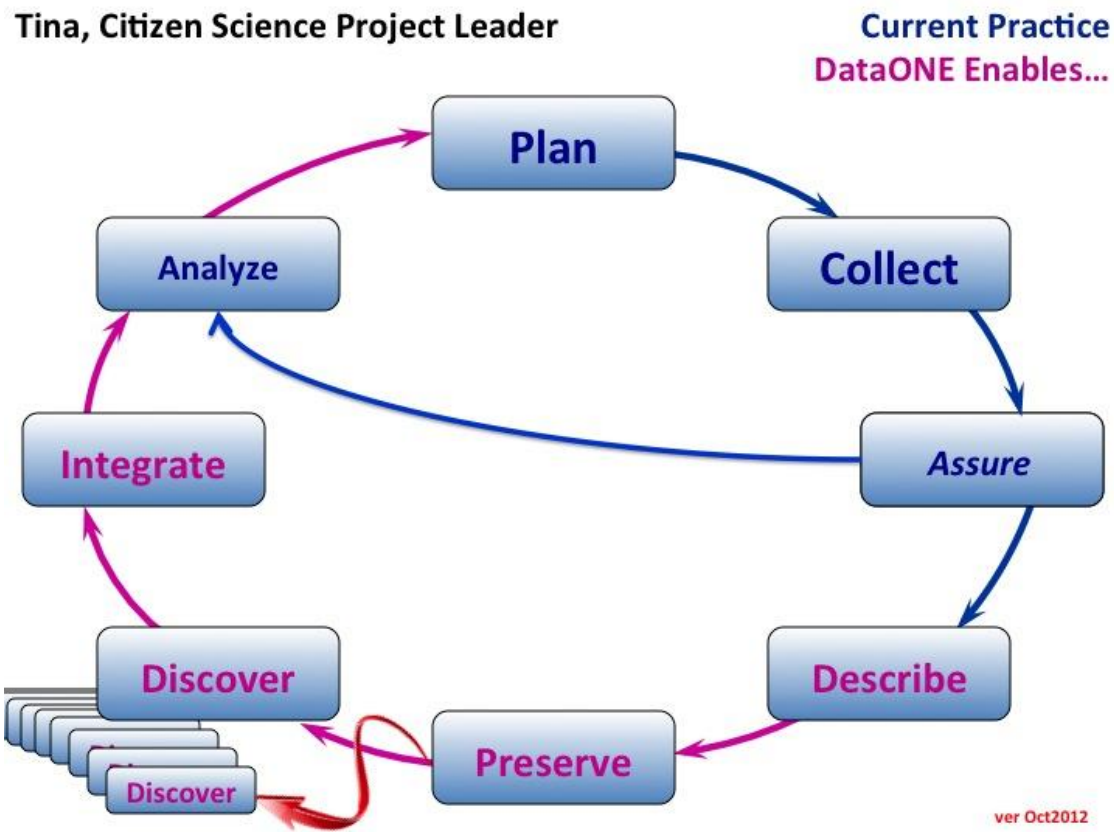
Tina knows she doesn't know much about data management but always has more pressing priorities to handle and so hasn't been able to make much time to learn. However, she is becoming worried that the project data may be lost if her *ad hoc* funding sources are not renewed or are compromised. She is also interested in data sharing with other organizations and researchers who can use the Weed Watchers data for larger studies, though she does not have contacts beyond Oregon State to use the data. She feels it's very important to make sure the data are put to scientific use, as this is part of her commitment to the project participants, and also a way to fulfill MNC's mission.

Tina is only vaguely aware of data repositories, and because academic publication is not a priority, there is little internal pressure to ensure data management standards are met. However, in meetings with her partners from Oregon State, she has realized that this is a point of concern for both project credibility and the potential of the data to be used for conservation research, part of MNC's mission. She is not sure how to ensure that the data that are collected by her volunteers are interoperable with other data sources, whether her protocols align with those used by other groups, and never has time to seek out relationships for developing ongoing partnerships, although she knows this is an important aspect for long-term project sustainability.

Tina's volunteers collect data which do get used by Oregon State researchers and local land managers. She does all of the data review herself using Excel spreadsheets. She has never tried to document the database with metadata, and in fact does not know what metadata means. While she is aware that there are data repositories for this type of data, she has no idea how to figure out where she could have her data archived and shared. Currently, the data is backed up weekly to an organizational backup drive, and she also stores it in Dropbox. She does not currently work on data discovery, integration, or analysis, but hopes to support others who can do so by providing quality data.



DataONE could provide Tina with a repository decision tree tool to figure out where she can deposit her data. She needs tutorials on the basic expectations for data documentation so she can prepare a description of the dataset, and guidance on how to maximize data interoperability. Access to tools that provide feedback during the data upload and description process would be especially helpful. Ideally, DataONE would also offer a tool that would provide a quick and easy set of basic data visualizations that would update whenever she uploads fresh data, and which she can embed in her project web pages. Support for data deposit and documentation would also make the data discoverable.



Comparison of current and DataONE-enabled practices:

Project Planning

- *Management Planning:* Develops a project Data Management Plan following examples provided on the DataONE portal.

Current data collection:
Collects field data.
DataONE enabled data collection:
No change.

Current data assurance:
Validates data using own standards.
DataONE enabled assurance:
Could apply more broadly-used data quality standards and assurance tools.

Current data description:
Describes data for her own purposes, using her own data description techniques.
DataONE enabled description:

- *Training:* Learns how to use *Morpho* (a metadata management editor) based on instructional materials available in the DataONE Best Practices Database and associated

downloadable video instructions.
- Creates metadata for datasets following best practices.

Current data preservation:
Tina publishes summary and analysis results but does not deposit data or have arrangements for long-term preservation.
DataONE enabled preservation:
Tina might deposit data with a DataONE member node for long-term preservation.
- *Data Preservation:* With colleagues, submits a research paper to an ecological journal associated with Dryad—a DataONE Member Node. Upon acceptance, she submits the publication-relevant data, metadata, and model to Dryad where they are given a DOI (digital object identifier) and preserved in the Dryad repository.
- *Citation:* Upon publication, she adds the publication reference and the data citation (including DOIs for both; provided by Dryad and the journal) to her CV.

Current data discovery:
Does not use other researchers' data.
DataONE enabled discovery:
Could use DataONE tools to discover relevant data from other researchers.

Current data integration:
Does not use other researchers' data.
DataONE enabled integration:
Could use DataONE tools to integrate her data with data discovered from other researchers.

Current data analyses:
Uses standard desktop data analysis tools.
DataONE enabled analysis:
- *Data Visualization:* Uses data analysis and visualization tools identified through DataONE Tools Database or available as part of the Investigator Toolkit to analyze existing data and develop initial model parameters that she will use in her own research.

Data Visualization: Creates graphics using tools identified via DataONE.

Rick

 [8]

(Secondary persona)

Source: Written by Ahrash Bissell

Tags: citizen scientist, non-professional, community

See also: Rebecca Leaking D1 Scenario

Background

Name, age, and education

Rick is a middle-aged male with an advanced degree in civil engineering. And in fact, Rick is a civil engineer by trade, but he has long been an enthusiastic outdoorsman, nature photographer, and gardener, and is keenly interested in local and native flora. He loves photographing what he sees on his nature walks, which mostly take place in local reserves but occasionally on family vacations. He has supported the local chapter of the Nature Conservancy for years both financially and as a participant in guided walks and as a volunteer for management activities at their local preserves. In the last few years, he has participated in an effort coordinated by Nature Conservancy staff to regularly survey the presence, abundance, and life-history characteristics of several different plant species that are resident in local reserves. He is not a computer or data expert but is perfectly comfortable using computers and following standardized rubrics and protocols.

Life or career goals, fears, hopes, and attitudes

Rick sees his citizen science activities as one of his hobbies. However, that doesn't mean that he doesn't take this work seriously. He believes that the data he is helping to collect are important,

[8] https://www.flickr.com/photos/mmoorr/233040583 (CC BY-NC 2.0)

both for local resource management needs but also as part of the larger body of evidence regarding the role of nature preserves in protecting endangered and native species. His sense of the exact role his contributions play is vague though, and he often wishes that he had a better idea of how the information is being used, and also how important his specific contributions are to the enterprise. Regardless, because he participates in data collection for fun, he expects to keep participating as long as there are opportunities to do so.

A day in his life

On any given weekend, Rick is likely to be found taking a stroll through one of the local preserves, taking photographs along the way. He is now proficient with the data collection protocols, having gone through an initial brief training program with reserve staff and also met up with other data-collection volunteers to hear updates on the project and share notes. The protocol is fairly simple, consisting of noting the presence/absence of any target species he finds, noting the location (if possible, usually using his smartphone's GPS or by noting the location on a map with a grid of coordinates), noting the life-stage of the plant (e.g., germinating or leafing out, flowering), and adding any other notes that seem relevant. Rick records all of this data in waterproof pocket notebooks and turns those notebooks into the reserve staff as he fills them up. He will often, but not always, take a picture of the specimen at the time of the observation, noting the filename of the picture in his notes. He copies all of the pictures onto a disk and turns those in to reserve staff as well. The reserve staff take responsibility for digitizing the data and linking the pictures to each observation record. When the staff finds errors or cannot read something, they have Rick's contact info and follow up with him for clarification. Occasionally they have to throw out a record because they cannot resolve its accuracy to their satisfaction, but this process tends to be *ad hoc* and subjective.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

Rick first becomes aware of DataONE at a meeting with other volunteers when the reserve staff show how the repository they have been using has become a DataONE node, which allows them the use of the visualization and other analysis tools. The reserve staff also showed how the data from their local project is being integrated with comparable datasets on those same species in other places, providing a more comprehensive picture of the natural histories of the target species and their current conservation status. Rick was inspired by this presentation, which confirmed for him that his volunteer efforts were contributing meaningfully to the science. He also became more interested in exploring related data himself, as well as in improving his data collection methodology to reduce the numbers of errors and the effort required by the reserve staff to digitize the data. To the extent that DataONE could provide templates, best practices, or other data acquisition tools, he might find find that useful. And to the extent that the visualization and data discovery tools are designed for a layperson, he might find that an interesting way to spend some of his time. And perhaps these tools and processes will help him become a leader of sorts for the local effort, serving to train future volunteers and assist reserve staff.

Intellectual and physical skills that can be applied

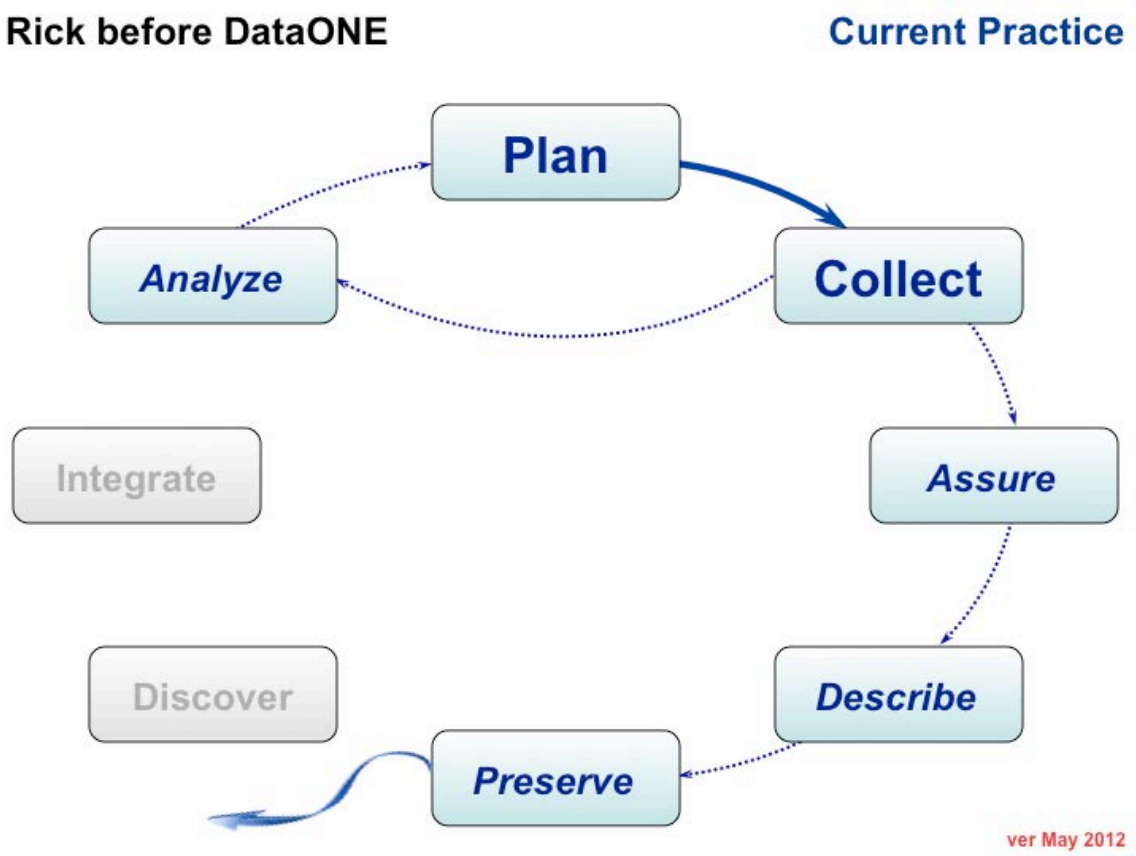
At this point, Rick is quite expert at spotting and cataloging the target plants as an enhancement of his regular walks. Having seen the presentation using DataONE, he is willing to make some extra effort searching for and cataloging the plants in a more systematic manner, though probably only some of the time. He is also willing to learn how to use new tools, such as hand-held GPS units (perhaps available for loan from the reserve staff). And he is willing to assist with the process of entering his data (and those of others) into databases himself, especially if it means that less data will be lost.

Technical support available

To date, he has not had to deal with the technology directly, for the most part. Even with DataONE, he may get some limited technical support from reserve staff, but for the most part he will be on his own, particularly when it comes to any exploratory analyses he might perform using DataONE tools.

Personal biases about data sharing and reuse (and data management more generally)

Rick has not thought much about the challenges of sharing data. He has some increased appreciation for the importance of data quality, but he doesn't otherwise have any reason to withhold data or actively seek other data. He has been operating on the presumption that the data he contributes to the project are useful and are contributing to the larger scientific enterprise. If his data are not being shared (e.g., to protect the locations of specimens of sensitive species), he would probably appreciate knowing about that and may even feel privileged to be part of a trusted circle of participants who have special knowledge about sensitive species.



D1 usage scenarios

- **Data collection:** Rick might be willing to adopt slightly more robust or automated data collection methods based on DataONE templates once he has seen how his data integrate with other data available through DataONE to enable scientific inquiry.
- **Data assurance:** Rick does not know much about data assurance but has developed a renewed appreciation for accuracy in data collection and transcription after seeing how the data get integrated into the DataONE member node. Were the project to implement some kind of expert data filtering similar to eBird, Rick might be willing to volunteer as a first-level editor.
- **Data discovery:** Rick is only likely to engage in data discovery activities as a casual exercise, but the fact that such activities are possible is a major motivational aspect of his continued interest in being a volunteer for the project.
- **Data analysis:** Rick is only likely to engage in data visualization activities (which comprise one component part of analysis) as a casual exercise, but the fact that such activities are possible is a major motivational aspect of his continued interest in being a volunteer for the project.

Elizabeth: University administrator (department chair)

 [9]

(Secondary persona)
Source: Written by Ahrash Bissell

Background

Name, age, and education

Elizabeth is a mid-to-late-career woman who is a professor of biology and the department chair for a regional comprehensive university in Washington State. This university historically focused on teaching. While faculty consider research to be an important part of their work, their research activities are generally restricted to smaller, minimally-funded projects, and they usually have time for research activities only during the summer. Elizabeth was recruited to this university from a research-intensive state university in order to expand the research activities in the biology department, both by changing existing reward structures and by recruiting new, research-focused faculty. However, her efforts are hampered by limited research support and facilities. She is unlikely to be able to recruit faculty who excel according to traditional metrics for the field, such as numbers of publications and impact factors, since those faculty will likely be able to find positions at other institutions which will provide greater support for their research activities. Instead, Elizabeth would like to recruit faculty who excel at new metrics that have demonstrable impact on the field of biology. She consulted with the dean of the college, her current faculty and colleagues elsewhere, and there was broad agreement that the department can and should evaluate the scholarly impact of raw datasets which are contributed to the field. She would also like to perform such evaluations on possible new hires.

Life or career goals, fears, hopes, and attitudes

While Elizabeth still maintains a limited research program, she has mostly transitioned to being an administrator. She believes all faculty in higher education should be doing serious, impactful research, which was a primary motivation for accepting the position at this regional comprehensive university. At the same time, she believes that the standards by which we measure "impact" are both flawed and out of date, and they are particularly ill-suited to the circumstances of faculty at a university like hers. She is passionate about finding new ways to support such

[9] https://www.flickr.com/photos/cobalt/2672600427/ (CC BY-NC-SA 2.0).

faculty and to elevate the profile of her new institution, and she believes that the expanding role of cyberinfrastructure in research activities offers some possible solutions, if she can just identify them. She fears that the faculty will resist all of these changes and will resent the push towards greater research productivity, towards new institutional and field-wide metrics, and towards the changing make up of the department. She also fears that the funding and other support for new cyberinfrastructure may not last, especially if few scientists gravitate to these new tools and capacities. These fears mean that she is taking a risk by orienting the department to these new tools and metrics, but she feels strongly that these are risks that must be taken, especially for a department like hers.

A day in her life

The biology department has been advertising for new faculty positions, and all job announcements highlight research expectations. The hiring committee narrows the pool of candidates down to a few applicants, but this process is typically based on traditional profiles and metrics. Elizabeth is then tasked with determining the extent to which each of these candidates utilizes key cyberinfrastructure research tools, collaborates with peers located at different institutions, and contributes impactful data to the field. Some candidates have addressed these issues in their applications, so she has some information at hand regarding the locations of their deposited data and any impact metrics they employed, but she has to perform independent analyses as well. She will need to be able to produce some summary statistics of these data-deposition and impact metrics for the hiring committee, so she needs a tool which enables such queries with a minimum of hassle and subjectivity.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

In the absence of DataONE, Elizabeth would have to navigate to many different repositories, sign in or create accounts on each one, learn how to query each system appropriately, and then manually compile the disparate data into a coherent document. For Elizabeth, DataONE offers many possible advantages. First, presuming that all of the relevant repositories are member nodes for DataONE, she can access all of them via a single sign-on. She only needs to learn how to work with one set of tools, and she can query multiple repositories and datasets at once. Second, it may be possible to establish some standardized methods of extracting the information she seeks which will expedite such analyses in the future, even as the diversity of member nodes and the data within continues to grow. Third, DataONE may offer the ability to produce summary tables and other outputs based on her analyses, saving considerable time and effort. Fourth, DataONE might make it possible to extract consistent information about data reuse (e.g., citations, derivative datasets, etc) which would otherwise be very difficult to obtain and compare. Elizabeth believes that the DataONE project, and similar ventures, are key to her current needs and to the future of the biological sciences. This also means that Elizabeth has high expectations for DataONE and is likely to react to shortcomings harshly.

Intellectual and physical skills that can be applied

Elizabeth is proficient in working with data and databases and has pushed herself to stay on top of new developments in the field. However, she has very limited time and must develop methods to streamline both the analyses and the reports. Because the analytical tasks are not standardized, she feels she must perform the work herself, though she is optimistic that DataONE will make it possible to create rubrics or scripts which will allow her to automate or delegate much of this work in the future. To the extent that Elizabeth is able to extend the functionality of DataONE and create potentially valuable templates for data-related impact metrics, she is interested in sharing those insights and products back to the field.

Technical support available

Elizabeth has access to some department-level technical support, as well as support for constructing queries which will extract the information she needs. However, her expectations that she will not really need such support, presuming DataONE has been built to meet the needs of someone like her.

Personal biases about data sharing and reuse (and data management more generally)

Elizabeth believes that data sharing is crucial, and that researchers should be encouraged to generate and share data and be rewarded accordingly. She also recognizes that such beliefs are not universal, and that changing the culture of science when it comes to tenure and advancement is hard. She is hoping that other peers and institutions will also move in this direction, though she is willing, if not exactly eager, to be a leader in the field. It is her belief that scientists generally lack good data management skills, to the detriment of the profession, and that elevating the professional impact of data sharing, courtesy of projects like DataONE, is the best opportunity to change attitudes and practices.

D1 usage scenarios

Researcher Support

- Data discovery: Elizabeth needs more efficient and robust data discovery tools, with special emphasis on metadata and paradata about the datasets of interest.
- Data integration: Again, Elizabeth's interest is in metrics and metadata about the contributed datasets, not the actual raw data contributed by the applicants, so she doesn't really need to integrate data in the same manner as researchers might want to. However, to the extent that DataONE enables comparisons in these variables for different researchers and their data, this would be enormously helpful.
- Data analysis: For Elizabeth, being able to analyze the metadata across datasets and researchers is key. If the metadata can be integrated for ease of analysis, that would be great, but otherwise she is likely to just export the relevant data and analyze them using standard statistical packages.

Mr. McMillin: K-12 educator

[10]

(Secondary persona)

Source: Written by Bruce Grant, Kevin Crowston and Miriam Davis with input from Kely Lotts

Tags:

See also:

Background

Name, age, and education

A native of Knoxville, Mr. McMillin graduated with a dual degree in Chemistry and Education from Berea College, with coursework supporting teaching certification in Kentucky. Family ties led him to return to the middle class urban/suburban neighborhood in the East Knoxville School District near where he grew up. Upon his return to Knoxville, Mr. McMillin sought employment as a High School Chemistry teacher, but the district only had an opening in Middle School Earth Science. Somewhat begrudgingly he took the job, but quickly grew to like it, and has been teaching at his school now for over a decade. Mr. McMillin now chairs the Science Faculty, and new teachers, as well as his older colleagues, at least some of them, have come to depend upon him for new ideas, programs and innovations.

Life or career goals, fears, hopes, and attitudes

Mr. McMillin feels successful in his career but his school district has seen hard times over the past few years. Pressures for compliance and accountability under "No Child Left Behind" (NCLB), exacerbated by the economic downturn and reduction of the East Knoxville school district's limited property tax base, have resulted in a 22% staff reduction and the evisceration of their in-service (i.e., professional development) programs for science faculty and curriculum development.

Mr. McMillin struggles to gain his students' attention in the face of many distractions. He seeks to engage them, to show them that science is pertinent to their lives, not just a subject in school. To this end, he values project- and inquiry-based learning and wants his students to do meaningful projects. But he needs help to continue to interest and engage students in the face of increased work and reduced resources. Regardless of the morale of the neighborhood and the

[10] https://www.flickr.com/photos/jblmpao/6215109547 (CC BY-NC-SA 2.0). Picture is of Dr. Michael Page, Stevenson High School, Stevenson, WA. Image was cropped to fit.

school, it is essential that Mr. McMillin give his students hope that they can contribute meaningfully, and that their learning matters.

A day in his life

Mr. McMillin teaches five sections of Earth Science daily. With the addition of his administrative responsibilities as chair of the faculty and advisor to the Science Club, Mr. McMillin is at school from 7am–5pm with only 55 minutes of class prep time each day. At night he grades papers, prepares lesson plans, and generates lab materials. He no longer has the time or money to take his students on field trips for observation or data collection; there is a small wetlands on the edge of the campus that is a partial substitute. He has basic lab equipment, though lab supplies are short.

Instead of field trips, he attempts to grab students' interest with the technology, e.g., computerized analysis and visualizations. He seeks to connect them to global science communities, showing how science can use the same kinds of social media tools that engage his students' attention. However, he is required to teach to the state science standards, meaning every item in his curricula must directly match a state mandate. His students' success on standardized tests directly impacts the rating of his school, his and the school's success, budget and student-to-teacher ratio.

Needs and expectations of DataONE tools

Mr. McMillin needs access to data his students can use to augment the limited data they can collect on their own in labs or outdoors. He has also found analysis of existing datasets to be a great way to involve special needs students with physical disabilities that keep them from participating fully in fieldwork. However, to be useable, he needs an educational module with teaching materials, mapped to educational standards, that identifies the appropriate data for the students to access and the tools for them to use. Ideally, these exercises will provide examples of how science and math matter in people's lives. He has been using the National Science Digital Library (NSDL), so he wants similar kinds of support for finding material. He gets a lot of ideas from friends, so it could be useful for him to connect with other teachers using similar materials to trade tips and ideas.

Finally, Mr. McMillin seeks to demonstrate how all people can contribute meaningfully to the scientific process. Given his students' interest in online interaction, he feels it might be possible to get them interested in participating in an on-line research project. Given the right materials, he could have students collect data and contribute it to a citizen science project, then follow that project to see how the data are used. He's been wanting to do something like this for years, but hasn't yet found a project that he could fit logistically and topically into his classes. On the other hand, he is certainly not going to create and deposit his own datasets, much less the metadata to describe them.

Intellectual and physical skills that can be applied

Mr. McMillin is familiar with scientific data collection and analysis from his undergraduate education. He has had some additional training since finishing his degree (though not in data science), and does his best to stay current with scientific developments, mostly from the popular

press. He is proficient with computer applications and comfortable navigating the Web and getting data into Excel.

Technical support available

The school and the district has a small IT staff, but they have their hands full keeping the school's computers and networks running smoothly and so are unable to provide much personal support for faculty. However, the IT staff is willing to install new software for Mr. McMillin, as long as it's open source software, as the district has no budget for buying new software.

Personal biases about data sharing and reuse (and data management more generally)

As an educator and mentor, Mr. McMillin believes in open access to science information and processes and would like his students to see how even their seemingly small efforts can integrate with those of the larger scientific community through technology.

D1 usage scenarios

Researcher Support

Project Planning

- *Professional development and training:* Download curriculum units and explore online tutorials in data management practices.
- *Professional development and training:* Access online tutorial datasets and visualizations for lesson planning.
- *Data evaluation, analysis and visualization:* Create a simple example workflow in preparation for student application.

Project Activity

- *Data discovery, access, use and dissemination:* Engage students in sample data exploration and visualization.
- *Data interoperability, standards and integration:* Work though data integration practices with students utilizing dataset from multiple member nodes (per tutorial guidelines)
- *Data evaluation, analysis and visualization:* Create species distribution graphics and simulations of species movements through time.

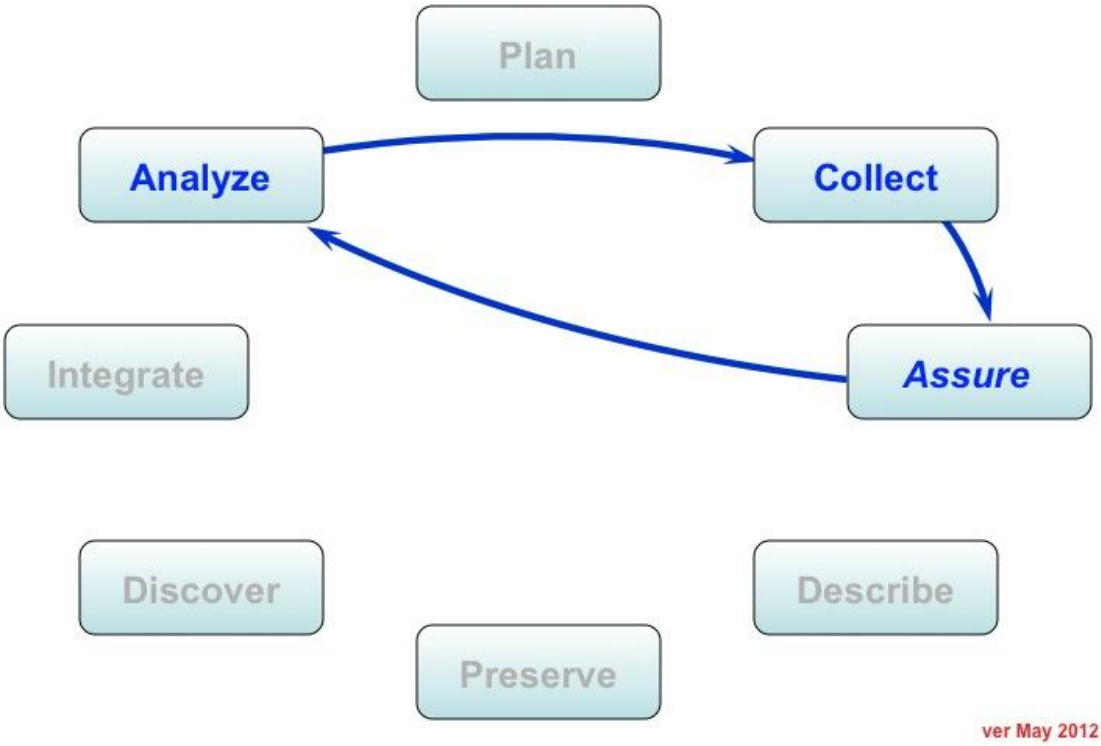
Publication and Data Preservation

- *Data deposition/acquisition/ingest*: Upload class data from feeder observations to the Avian Knowledge Network.

Additional Activities

- *Data discovery, access, use and dissemination:* Use the class data and integrated datasets as the basis for county science fair projects.

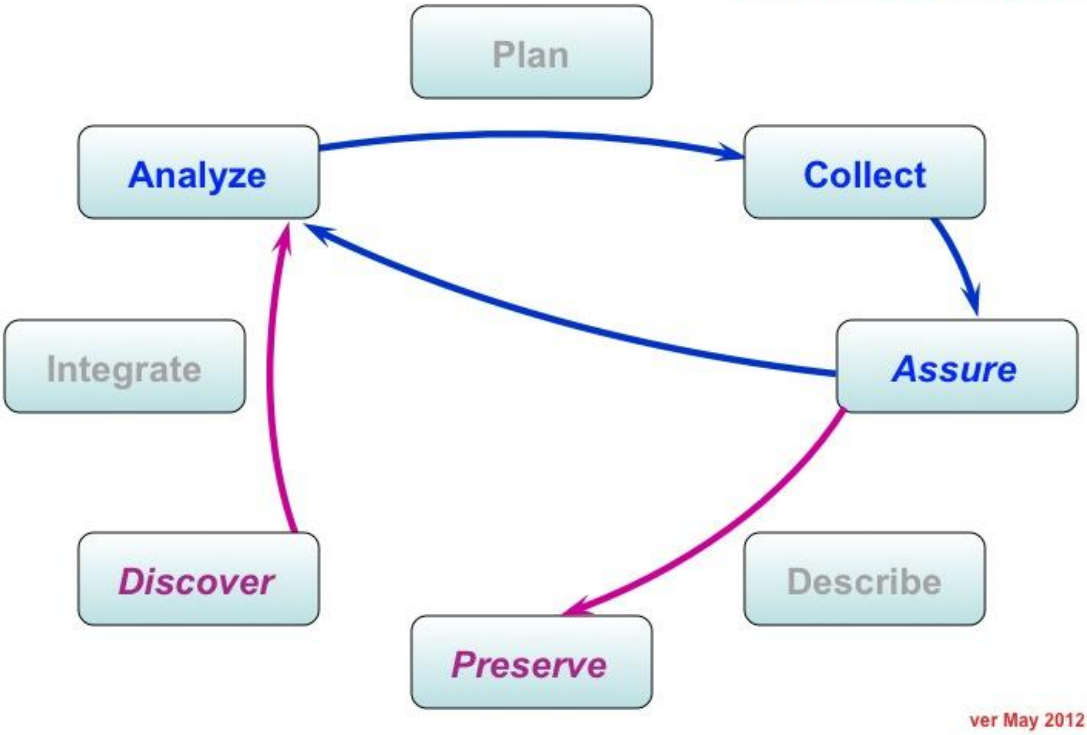
Mr. McMillin, K-12 educator **Current Practice**



ver May 2012

Mr. McMillin, K-12 educator

Current Practice
DataONE Enables…



ver May 2012

Gretta: College educator

 [11]

(Secondary persona)

Source: Kevin Crowston and Miriam Davis, with input from Bruce Grant and Gretchen LeBuhn

Tags: college educator

See also:

Background

Name, age, and education

Gretta received her PhD from University of Wisconsin Madison in Zoology six years ago at the age of 29. She did her dissertation fieldwork on the Mosquito Coast of Honduras studying physiological and population ecology of bats. Her research linked individual bioenergetics to landscape level meta-population genetics.

At Wisconsin, Gretta was a Teaching Assistant for general biology lectures and labs, and taught several different sections of a variety of elective courses but she did not specifically study or receive more than minimal training in teaching or education. However, during a postdoc at the La Selva Biological Station in Costa Rica she interacted with a group of undergraduates doing a field study. This positive experience led her to consider a career with a greater focus on teaching than she had previously been considering.

As a result, Gretta accepted a position at Kendall College in the Cuyahoga Valley of Ohio, and has been teaching in the Biology department for four years. She also has an appointment in a new interdisciplinary Environmental Studies program. Kendall College is a primarily undergraduate institution located in a distant suburb of Cleveland. Named after its first benefactor, Kendall College occupies approximately ¼ of a large forested preserve, with a pond, stream, unsurveyed caves and karst features. It was well out in the country when it was founded, but in the intervening decades, the suburbs have grown to the edges of campus.

---

[11] https://www.flickr.com/photos/usfwshq/5687793836/ (CC BY 2.0).

Kendall College enrolls approximately 2000 undergraduates plus about 800 students in MS programs in teaching, nursing and allied health fields (but there is no master's degree in biology). Approximately 50% of students come from within 50 miles of the school, so support from the community is important. The biology department has about 15 faculty; it is large compared to other departments because of the number of service courses offered and the large number of pre-med biology majors.

Life or career goals, fears, hopes, and attitudes

Gretta's main concern at the moment is to get tenure. The reward system at Kendall College is heavily based on teaching, with research as necessary but secondary. Gretta's first year of teaching was disastrous—she didn't fully appreciate how different the students and their needs were from those she had encountered when she was a TA. But through hard work, she has overcome this rocky start and now receives excellent teaching evaluations. She has also stayed active in her research, so tenure should be no problem. Nevertheless, she is not confident that she fully understands the politics of her department and the college. She feels a need to be careful about her relationship with her colleagues, particularly with senior faculty who do cellular and molecular biology and who do not always appreciate her ecological interests.

Gretta also wants to make a difference in her students' lives, which is why she came to Kendall instead of going to a more research-focused university. She particularly wants to help students learn to think more systematically about the ecosystem and the ways it may be changing. She thinks that fieldwork is very helpful way to engage students and to get them to think more broadly, but her time for field research pursuits is limited and many of her students, even biology majors, are focused on a health career rather than the environment.

A day in her life

Gretta teaches 3 courses a semester, a mix of service courses for professional health and science education students as well as courses for majors. She teaches the ecology and evolution sections of the introduction to biology course, where she faces the problem of keeping the interest of pre-med students who do not see the relevance of ecology to their careers. For biology majors, she teaches mammalogy, where she faces the same problem. Ironically, she gets the most satisfaction teaching non-major electives, such as animal diversity and issues in conservation. A small but growing number of students are interested in conservation biology and the new environmental science program. The growth of this program is welcomed by mostly newer younger faculty but seen as a potential threat to the department by older more-established faculty. Teaching occupies about sixty percent of her time with a significant part of the rest of her time being committed to administrative matters, leaving little time for research.

Nevertheless, for her research, Gretta remains concerned about the physiological and population ecology of bats, especially with regards to invasive species, white nose bat syndrome, and the impact of sprawl from Cleveland. As she has only a token amount of funding from the college to support her research, she takes advantage of the campus preserve and nearby areas for research sites. In the summer, she actively involves undergraduates in collecting data about bat habitat and occurrence. They also collect some specimens (e.g., scat, prey insects, adult and larvae, diseased

bat corpses). She can collect enough data during the summer to keep busy with analysis and writing during the rest of the year, with a few periodic visits to local sites for on-going monitoring. Her work provides many opportunities for student involvement in both data collection and analysis, though this occasionally puts her in competition with other faculty for good students.

Gretta would like to get students involved in surveying the campus preserve, e.g., doing something like the All Taxa Biodiversity Inventory (ATBI), and comparing these results to other areas along the corridor to Cleveland. As a step in this direction, Gretta partnered with an Education faculty member on an experiential science education summer course series that brings college students, inner-city high school students and college faculty together to examine bat habitat use on the preserve and along the corridor. This course also provides her with much-needed summer salary and some infrastructure for managing students. Officially such collaborative activities are encouraged, but she is aware that some of her senior colleagues do not think the work is real science. She is also beginning to realize that a greater volume of data collected will require more work to manage and is not sure she and her students are adequately prepared for that.

Because of her interest in invasive insects, Gretta was recently named to the county's Climate Change Adaptation Planning Committee as an expert in the impacts of climate change on the area's ecology. The college greatly values this kind of community engagement and she was initially flattered to have been invited ahead of her more senior colleagues and hopes that the data she has collected will be useful. However, she is concerned about the time required and the potential for political missteps that might affect the college and her career. For example, the local community is concerned about the potential impacts of the West Nile virus and has been spraying insecticide to control mosquitoes but Gretta is worried about the impact of spraying on the wildlife in the college's preserve and in the habitat corridor between Cleveland and the college.

Reasons for using DataONE to share and to reuse data

Needs and expectations of DataONE tools

Gretta wants data to inform her own research, e.g., on similar populations and regions or complementary data on the area she studies. She wants to stay involved with zoology community by contributing her findings. She hopes to find relevant data to inform discussions of the County Planning Committee, but feels that digested data would be more useful for them.

After initially learning about DataONE as a place to look for data, she found the educational materials on the DataONE website and is interested in using DataONE data for class exercises. For example, the ecology course could be enhanced by having students look at data on habitats and species distributions. It would certainly help to have well-designed class modules: she has developed a lot of her own materials, but it's hard to make time for such innovations given the other demands of her busy day.

Gretta is currently thinking of redesigning her research to collect data more systematically across a wider area and over time, rather than her current approach of addressing a project at a time. This

approach would require more effort but would produce a longitudinal dataset that she could contribute to a DataONE member node. However, her senior colleagues have made it clear that a dataset does not count as a publication. Before she can devote effort to the redesign and increased data collection, she will need some assurance that she and her students will receive appropriate credit for the work, at least a publication about the data and ideally joint authorship on papers that use her data.

Intellectual and physical skills that can be applied

Gretta's graduate and postdoc work included training in data collection and analysis, but not in data management more generally. She is skilled with general data collection approaches but mostly relies on Excel to store her data. She also relies on students to do much of her data entry and Excel works well for that purpose.

Technical support available

Kendall College has an IT staff that can handle routine issues and Gretta is provided with a PC and routine software. However, she does not have further support with the technology. She has a colleague in the department who helps her with statistical analyses.

Personal biases about data sharing and reuse (and data management more generally)

Gretta observed the value of shared data in her graduate education and postdoc work. She appreciates the ability to access others' data and would like the opportunity to reciprocate, but is not sure that she can justify the time and energy it would take or that it would be rewarded.

D1 usage scenarios

Project Planning

- *Professional development and training:* Access online tutorial datasets and visualizations for lesson planning.
- *Data evaluation, analysis and visualization:* Get help setting up longitudinal study along a habitat continuum. See example workflows. Use workflow tools. Create a data management plan.

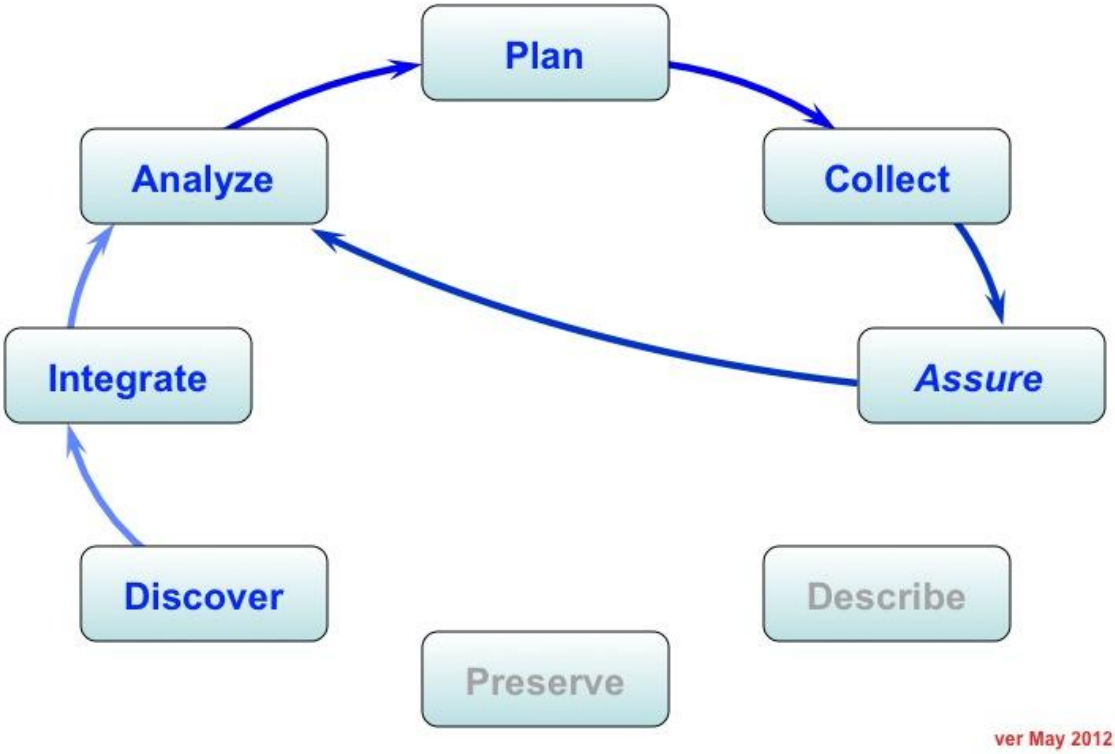
Project Activity

- *Data discovery, access, use and dissemination:* Engage students in data exploration and visualization.Use the class data and integrated datasets as the basis for publications with undergraduates.
- *Data interoperability, standards and integration:* Contribute data from Ohio so that it can be integrated with other studies.
- *Data evaluation, analysis and visualization:* Create species distribution graphics and simulations of species movements through time to help students realize their seemingly small efforts in a summer or semester course do have larger impacts and context.

Publication and Data Preservation

- *Data deposition/acquisition/ingest*: Upload research and class research data for preservation and reciprocation.

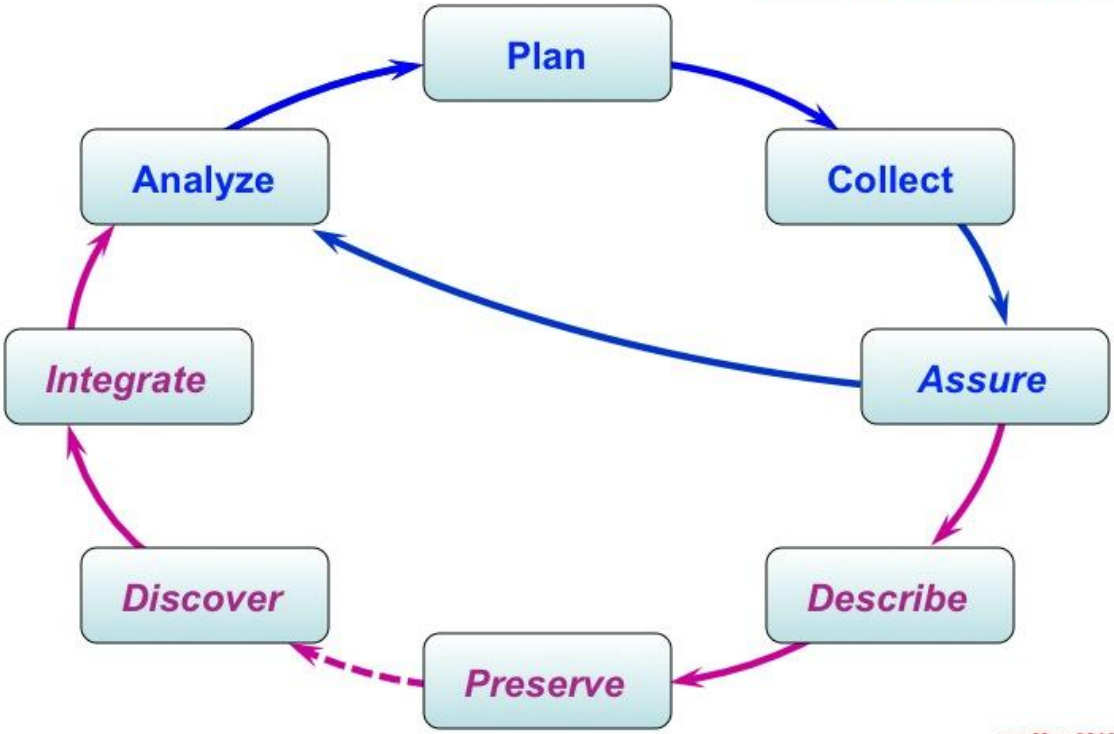
Gretta: College educator **Current Practice**



ver May 2012

Gretta: College educator

Current Practice
DataONE Enables…



ver May 2012

Appendix:
General compilation of compelling reasons and tools for researchers to benefit from DataONE

Project Planning

- Data Discovery, Access, Use and Dissemination
  - Determine 'ideal' and minimum datasets necessary for generating a model with robust outputs.
  - Using DataONE metadata clearing house, discover what data are available that meet the needs of preliminary research.
- Data Integration
  - Standardize / compile / synthesize / integrate datasets (both input data and driver variables).

Project Activity

- Data Visualization
  - Use workflow and visualization tools to make model runs under varying conditions and evaluate with test data.
  - Explore, visualize, and analyze the results.

Publication and Data Preservation

- Data Preservation
  - Upload and publish value-added data products (inputs, driver data, and outputs) to DataONE repository and provide citation.
- Citation
  - Upload and publish model source-code to repository (as record of the model & algorithms used to generate the results) and provide citation.
- Publication
  - Publish and submit for peer-review a co-authored manuscript.

DataONE Value-Added Capabilities

- Supports rapid training in "best practices."
- Ability to search across multiple data sources; DataONE provides a single framework for acquisition of data from disparate sources reducing time engaged in data discovery. [Enables scientists to easily discover and access relevant data that are documented and stored in a wide variety of data repositories.]
- Use standardized metadata to evaluate utility of datasets; Development of standardized semantics for tagging data enhances utility of diverse datasets, reducing time engaged in data management activities. [Provides templates for developing data management plans, as well as tools that support data deposit and preservation.]
- Ability to generate integrated datasets; DataONE output can be yielded as a single dataset with file identification coded within the dataset reducing time engaged in data

management activities. [Supports easier data integration via semantic mediation tools.]

- Use of workflow / visualization / analysis tools; DataONE toolkit facilitates preliminary visualization of data prior to extensive modeling, ability to track data management processes and generation of products for use in presentations. [Promotes Discovery, Access, Use and Dissemination of software tools that facilitate data and metadata management, analysis and modeling, and visualization.]
- Quality attribution associated with DataONE name; Reputation of DataONE infrastructure and data coverage provides end users with quality assurance regarding the data discovery and integration processes.
- Standardized methodology statement / terminology for data extraction protocols; Inclusion of DataONE recommended description of the methods used for extensive data discovery and integration methodologies within the resulting manuscript; provides a citable resource for other researchers.
- Repository for value-added products (secondary datasets) that can be re-used by the community, cited etc; Uploading generated secondary dataset enhances the DataONE resource and provides a data citation within the resulting manuscript. Further facilitating scholarly communication and research development. [Promotes recognition of researchers for their scientific data contributions via internationally endorsed data citation mechanisms.]