"Personas" to Support Development of Cyberinfrastructure for Scientific Data Sharing

Kevin Crowston
Syracuse University
School of Information Studies
348 Hinds Hall
Syracuse, NY 13244–4100 USA

Telephone: +1 (315) 443–1676
Fax: +1 (866) 265–7407
Email: crowston@syr.edu

Draft of 3 August, 2015

"Personas" to Support Development of Cyberinfrastructure for Scientific Data Sharing

Abstract

Objective: To ensure that cyberinfrastructure for sharing scientific data is useful, system developers need to understand what scientists and other intended users do with data as well as the attitudes and beliefs that shape that use. This paper introduces personas—detailed descriptions of an "archetypical user of a system"—as an approach for capturing and sharing knowledge about potential system users.

Setting: Personas were developed to support development of the 'DataONE' (Data Observation Network for Earth) project, which has developed and deployed a sustainable long-term data preservation and access network to ensure the preservation and access to multi-scale, multi-discipline, and multi-national environmental and biological science data (https://www.dataone.org/what-dataone) (Michener et al. 2012).

Methods: Personas for DataONE were developed based on data from surveys and interviews done by members of DataONE working groups along with sources such as usage scenarios for DataONE and the Data Conservancy project and the Purdue Data Curation Profiles (Witt et al. 2009).

Results: A total of 11 personas were developed: five for various kinds of research scientists (e.g., at different career stages and using different types of data); a science data librarian; and five for secondary roles.

Conclusion: Personas were found to be useful for helping developers and other project members to understand users and their needs. The developed DataONE personas may be useful for others trying to develop systems or programs for scientists involved in data sharing.

Type: EScience in action

Keywords: Cyberinfrastructure development, user requirements, personas

Word count: 231 words (abstract); 1832 words (paper)

"Personas" to Support Development of Cyberinfrastructure for Scientific Data Sharing

Introduction

Research in the sciences, social sciences and humanities is increasingly data-intensive, collaborative and computational. Supporting data-intensive multidisciplinary collaborative research requires "new methods for gathering and representing data, for improved computational support and for growth of the online community" (Murray-Rust 2008). As a result, research data management (RDM) is now a critical need, with action needed across the data lifecycle: from data capture, analysis and visualization (Gray 2007), through curation, sharing and preservation, to support further discovery and reuse.

In data-driven research, researchers often interact with information and data through via networked computational tools. The computational tools developed to support research are often referred to collectively as cyberinfrastructure (Atkins et al. 2003) and the practices around them as eScience. Development of cyberinfrastructure for eScience offers much promise for data-intensive research. However, systems developers face a problem that has long troubled software development, namely ensuring that they understand the needs of users properly in order to build usable systems. As Brooks put it more than 30 years ago:

> The hardest single part of building a software system is deciding precisely what to build. No other part of the conceptual work is as difficult as establishing the detailed technical requirements, including all the interfaces to people, to machines, and to other software systems. No part of the work so cripples the resulting system if done wrong. No other part is more difficult to rectify later. (Brooks 1987)

The topic of developing system requirements is of concern to eScience librarians because they are often at the front line of data management and fill an important role bridging between technology and scientists (Crowston et al. 2015). Being able to convey to developers what they know about users is therefore an increasingly important skill for eScience librarians. To do so successfully, it is critical is to be able to capture and describe scientists' needs in a systematic way rather through the traditional face-to-face, anecdotal style of learning about users. In this paper, we report on the use of a technique called personas to communicate user needs for eScience, in this case of the needs of scientists and others involved research data management to be supported by a novel cyberinfrastructure.

Setting

The requirements work reported in this paper was done for the 'DataONE' (Data Observation Network for Earth) project (Allard 2012, Michener et al. 2012), which has developed and deployed a sustainable long-term data preservation and access network to ensure preservation of and access to multi-scale, multi-discipline, and multi-national environmental and biological science data (https://www.dataone.org/what-dataone). It was established in 2009 with funding from the United States National Science Foundation (NSF) and from mid-2014 commenced its second phase of development.

The DataONE project has several unique features: (i) it was designed to expand on existing infrastructure, (ii) it had a global mandate to offer tools and solutions that would promote science and knowledge-creation, and (iii) it needed to facilitate evolving communities of practice based around the ever-developing cyberinfrastructure and the use of it (Michener et al. 2012). In addition to developing cyberinfrastructure, the project also created tools for the research

community, such as training materials, a database of researcher tools and a catalog of best practices. The DataONE mandate was daunting: the environmental and biological science community is notoriously diverse with great variation in scales, discipline paradigms and data types, alongside substantial organizational and geographical diversity. To achieve its goal required innovative solutions to deliver a product that was usable and inter-operable across a wide range of disciplines including environmental, computer and human sciences.

Approach: Personas

To communicate user needs to project developers and other personnel, researchers involved with DataONE developed a set of persona documents (Cooper et al. 2014). In essence, a persona is a written description of a potential system user. The idea is that software will be more successful if it is designed with a specific user's needs in mind. Some software development methodologies go so far as to suggest that a user representative be always available to answer questions (e.g., the product owner in scrum development (Schwaber and Sutherland 2013)). However, this approach is not always practical. A persona document acts instead as a kind of user stand in, helping developers to understand the users even in their absence. Furthermore, a single person may not fully represent the range of users or may impose his or her own idiosyncrasies. In contrast, a persona does not describe a particular user or an average but rather describes an archetypical user of a system (Cooper et al. 2014, 62).

Personas have some features in common with other requirements documents that are in common use, such as use cases and scenarios. However, personas have some advantages over these approaches. Use cases treat all interactions as equally important, while personas provide information to understand user priorities. Scenarios focus on tasks, rather than users (Madsen

and Nielsen 2009, 59). Personas add detail about interests, emotions, settings and needs, including the goals of the people in using the software.

There are several kinds of personas that are relevant to system development: primary personas (the main user or users of the system); secondary (those who will be served as long as doing so does not 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) (Rind 2007).

Personas have been used by other projects for cyberinfrastructure development. Specifically, the Data Conservancy project developed a set of personas (Davis et al. 2010). More broadly, the Cornell library developed a set of persona for library users (Cornell University Library Web Vision Team and TKG Consulting LLC 2007).

Method: Developing a persona

Personas are built based on detailed data collected about users addressing activities, attitudes, aptitudes, motivations and skills (Cooper et al. 2014, 83). To develop the DataONE personas, we drew on data from the researcher surveys carried out by DataONE researchers (e.g, Branch et al. 2010) and additional interviews conducted by the persona developers. We also drew on the Data Conservancy personas (Davis et al. 2010), DataONE usage Scenarios developed by the DataONE Sustainability and Governance Working Group, and the Data Curation profiles from Illinois and Purdue (http://datacurationprofiles.org).

The description of a persona for DataONE includes (Rind 2007):

- 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

Some of the details (e.g., where the person described works or went to school) are essentially fictional, but they are carefully chosen to be representative of a typical user and help increase the verisimilitude of the persona description. Similarly, personas are given a name for ease of reference and a photograph to make them more real to the developers who use them.

To address data management more specifically, for each primary persona, we described which of the stages in the DataONE data lifecycle (shown in Figure 1) the researcher performs currently (in blue) and which might be performed using tools provided by DataONE (in red). Solid lines represent workflows performed by the persona; curved 3D lines represent flows of data from one researcher to another (as shown in Figure 2). 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 the stages of preserve and discover, as the persona preserves data for others to (potentially) use and discovers other peoples' data that they have preserved. Processes shown shaded out in the persona descriptions 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).

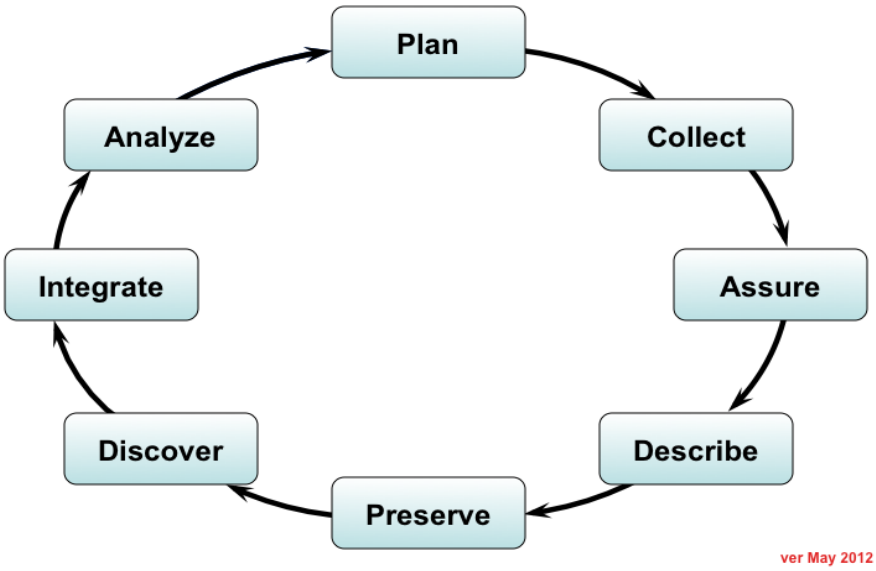


Figure 1. The DataONE data lifecycle, from Figure 7 Michener et al. (2012).

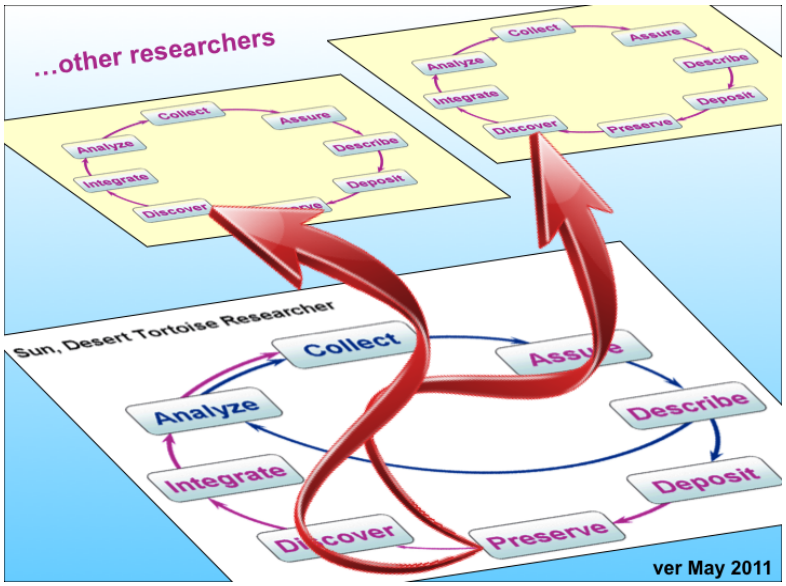


Figure 2. Red wavy lines represent data flows from the focus persona to others.

Results: Personas for DataONE

To date, we have developed 11 personas for DataONE (see Table 1). Six of these are primary personas, describing the main intended users of the DataONE cyberinfrastructure and project

tools. Five of the six are research scientists. These five scientist personas were developed so as to cover differences among scientists along multiple dimensions that were considered to affect how scientists would share or reuse data and so their likely use of DataONE. Dimensions covered are:

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

Table 1. DataONE personas

- Primary personas
  - Research scientists
    - Sun: Early-career herpetologist
    - Jean: Agricultural scientist at a field station
    - Laura: Mid-career oceanographer
    - Andreas: Biochemical modeller
    - William: Late-career plant taxonomist
  - Abby: Science data librarian
- Secondary personas
  - Tina: Citizen science project manager
  - Rick: Citizen scientist
  - Elizabeth: University administrator
  - Mr. McMillin: K-12 educator
  - Gretta: College educator

The sixth primary persona developed was for a science data librarian. The final five personas are secondary personas, describing other kinds of people who might be clients for the DataONE cyberinfrastructure and tools, but whose needs would only be served if doing so did not get in the

way of serving the primary personas. These personas included citizen scientists, administrators and educators. To date we have not described any negative or buyer personas. The full set of personas can be found at http://bit.ly/D1Personas. One example—Sun, an early-career government herpetologist—is given in an appendix to this paper.

Conclusion

Personas provide a tool to help developers and others involved in a system development project develop a shared understanding of users to guide development. Sharing a set of personas helps developers maintain a common vision of that user and promotes agreement between different stakeholders. The personas developed from DataONE proved to be helpful in communicating the research done with users and were well received by project members.

While these personas were developed specifically for DataONE, the descriptions are of researchers' (and others') work and lives more generally. As such, they may be useful for others developing systems or programs for those involved in research data management, either as they are, or as a starting point for further development. EScience librarians in particular may find the personas to be useful in planning products and services for researchers. By better understanding the wants and needs of users through tools like personas, developers can create cyberinfrastructure that is more responsive to their needs, thus improving the impact of these systems and of eScience more generally.

References

Allard, Suzie. 2012. "DataONE: Facilitating eScience through collaboration." *Journal of eScience Librarianship* 1 (1):e1004. doi: 10.7191/jeslib.2012.1004.

Atkins, D.E., K.K. Droegemeier, S.I. Feldman, H. Garcia-Molina, M.L. Klein, D.G. Messerschmitt, P. Messina, J.P. Ostriker, and M.H. Wright. 2003. Revolutionizing science and engineering through cyberinfrastructure: Report of the Blue-Ribbon Advisory Panel on Cyberinfrastructure. Washington, DC: National Science Foundation.

Branch, BD, C Tenopir, S Allard, K Douglas, L Wu, and M Frame. 2010. "DataONE: Survey of Earth Scientists, To Share or Not to Share Data." AGU Fall Meeting Abstracts.

Brooks, Frederick P., Jr. 1987. "No Silver Bullet: Essence and Accidents of Software Engineering." *IEEE Computer* 20 (4):10–19.

Cooper, Alan, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About Face: The Essentials of Interaction Design*. Indianapolis, IN: John Wiley & Sons.

Cornell University Library Web Vision Team, and TKG Consulting LLC. 2007. "Cornell University Library Personas." http://hdl.handle.net/1813/8302.

Crowston, Kevin, Alison Specht, Carol Hoover, Katherine M Chudoba, and Mary Beth Watson-Manheim. 2015. "Perceived discontinuities and continuities in transdisciplinary scientific working groups." *Science of The Total Environment*. doi: 10.1016/j.scitotenv.2015.04.121.

Davis, Lynne, Tim DiLauro, Mark Evans, Siri Jodha Singh Khalsa, Ruth Duerr, and Anne Thessen. 2010. "Moving From Users, Through Use Cases To Requirements." http://dlsciences.org/research/DataConservancy/DC+Requirements+White+Paper.pdf.

Gray, Jim. 2007. "Jim Gray on eScience: A transformed scientific method." In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, edited by Tony Hey, Stewart Tansley and Kristin Tolle, xvii–xxxi. Redmond, WA: Microsoft.

Madsen, Sabine, and Lene Nielsen. 2009. "Exploring persona-scenarios: Using storytelling to create design ideas." the Second IFIP WG 13.6 Conference, Human Work Interaction Design: Usability in Social, Cultural and Organizational Contexts, Pune, India, 7–8 October.

Michener, William K., Suzie Allard, Amber Budden, Robert B. Cook, Kimberly Douglass, Mike Frame, Steve Kelling, Rebecca Koskela, Carol Tenopir, and David A. Vieglais. 2012. "Participatory design of DataONE: Enabling cyberinfrastructure for the biological and environmental sciences." *Ecological Informatics* 11:5–15. doi: 10.1016/j.ecoinf.2011.08.007.

Murray-Rust, Peter. 2008. "Chemistry for everyone." *Nature* 451:648-651. doi: 10.1038/451648a.

Rind, Bonnie. 2007. "The power of the persona." *The Pragmatic Marketer Magazine* 5 (4):18–22.

Schwaber, Ken, and Jeff Sutherland. 2013. "The Scrum Guide™: The Definitive Guide to Scrum: The Rules of the Game." http://www.scrumguides.org/.

Witt, Michael, Jacob Carlson, D. Scott Brandt, and Melissa H. Cragin. 2009. “Constructing data curation profiles.” *The International Journal of Digital Curation* 4 (3):93–103. doi: 10.2218/ijdc.v4i3.117.

Acknowledgements

The DataONE personas were developed by a team of members from the DataONE Sociocultural Issues Working Group with support from other members of the DataONE team. DataONE is supported by US National Science Foundation Awards 08–30944 and 14–30508.

Appendix: Example primary persona

Sun

(Primary persona)

Source: Data Conservancy Sun persona: 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

Photo credit: U.S. Army Environmental Command https://www.flickr.com/photos /armyenvironmental/2650014187

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 number of her research subjects are radio tagged, giving her a latitude/longitude 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.

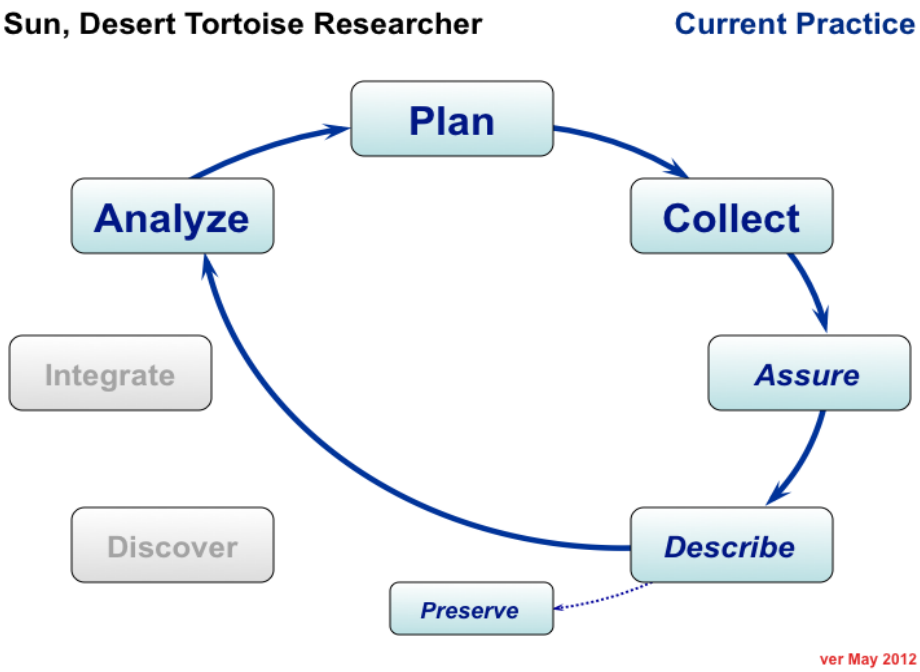
Technical support available

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.

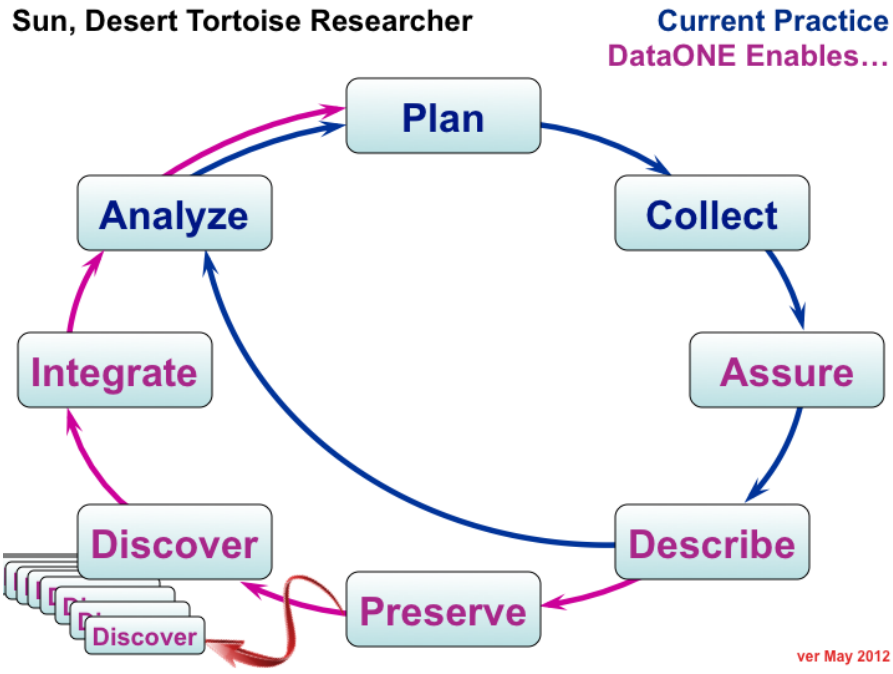
Personal biases about data sharing and reuse (and data management more generally)

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 geo-located 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:

Current project planning	*DataONE-enabled project planning*
No explicit attention to data issues in project planning.	*Management Planning:* Develops a project Data Management Plan following examples provided on the DataONE portal.
Current data collection:	***DataONE-enabled data collection:***
Collects tortoise field data.	No change.
Current data assurance:	***DataONE-enabled data assurance:***
Validates data using own standards.	Could apply more broadly-used data-quality standards and assurance tools.
Current data description:	***DataONE-enabled data description:***
Describes data for her own purposes, using her own data description techniques.	• *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:	***DataONE-enabled data preservation:***
Sun publishes summary and analysis results but does not deposit data. Data preservation is done only within her lab.	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 data 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 data integration:*** Use DataONE tools to integrate her data with data discovered from other researchers.
Current data analysis: 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.