ISSN 2161-3974

JeSLIB 2015; 4(2): e1082
doi:10.7191/jeslib.2015.1082

Journal of eScience Librarianship
putting the pieces together: theory and practice

EScience in Action

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

Kevin Crowston

Syracuse University, Syracuse, NY, USA

Abstract

Objective: To ensure that cyberinfrastructure for sharing scientific data is useful, system developers need to understand what scientists and other intended users do as well as the attitudes and beliefs that shape their behaviours. 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.

Correspondence: Kevin Crowston: crowston@syr.edu
Keywords: cyberinfrastructure development, user requirements, personas

JeSLIB 2015; 4(2): e1082
doi:10.7191/jeslib.2015.1082

Introduction

Research is increasingly data-intensive, collaborative, and computational. As a result, research data management 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 discovery, and reuse. In data-driven research, researchers often interact with data via computational tools, referred to collectively as cyberinfrastructure (Atkins et al. 2003). Cyberinfrastructure developers face a problem that has long troubled software developers, namely ensuring that they understand the needs of users. As Brooks put it:

> *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*. (Brooks 1987)

Developing system requirements is of concern to eScience librarians because they are often at the front line of data management. In some cases, libraries develop cyberinfrastructure themselves: for example, Lage, Losoff, and Maness (2011) report on the development of a library-led institutional repository for research data. In other cases, librarians link between scientists and developers of systems (Crowston et al. 2015). In either case, being able to convey to developers what they know about users is an increasingly important skill for eScience librarians. It is critical is to be able to describe users' needs systematically rather than through anecdotes. Without such knowledge, developed systems may fail to meet user needs (e.g., Dombrowski 2014). In this paper, we describe the application of a technique called "personas" (Cooper 1999) to communicate user needs.

Setting

The 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). DataONE 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 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 cyberinfrastructure (Michener et al. 2012), specifically, a search engine for data stored in diverse repositories. The project also created tools for the research community, such as training materials, a database of researcher tools, and a catalogue of best practices. The DataONE mandate was daunting: The environmental and biological science community is notoriously diverse with great variation in scales, disciplinary paradigms, and data types, alongside substantial organizational and geographical diversity. To achieve its goal, DataONE required innovative solutions that were usable and inter-operable across a wide range of disciplines, which required understanding user needs and sharing that understanding across the developers working on the project.

JeSLIB 2015; 4(2): e1082
doi:10.7191/jeslib.2015.1082

Approach: Personas

To communicate user needs to project developers and other personnel, researchers involved with DataONE developed a set of personas (Cooper 1999, Ch. 9). 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 always be 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 as a kind of user stand in, helping developers to understand 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 describes an archetypical user of a system (Cooper 1999, Ch. 9), and there can be multiple personas with different needs.

Personas have some features in common with other commonly used requirements documents, such as use cases and scenarios (both of which were used for DataONE). However, personas also have some advantages. 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). As well, personas add details about interests, emotions, settings, and needs, including the goals of the people in using the software, thus providing additional insight into user needs.

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, but who do not use the system themselves and so do not drive the interface) (Cooper 1999, Ch. 9).

Personas have been used by cyberinfrastructure development projects, including the Data Conservancy project (Davis et al. 2010). The personas for DataONE and the Data Conservancy have some similarity given the similar goals and target users of the two projects, though the DataONE personas include details specific to the use of the DataONE system. Lage, Losoff, and Maness (2011) developed personas for researchers who might be clients for a proposed library role in data curation. They note the value of personas for representing a range of users and the "disciplinary, institutional, and perhaps even departmental cultures in which [they] work," (p. 933).

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). For DataONE, we drew on data from the researcher surveys carried out by DataONE researchers (e.g, Branch et al. 2010) and additional interviews we conducted. 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).

One example persona — Sun, an early-career government herpetologist — is given in an appendix to this paper. As can be seen in the example, the description of a persona for DataONE includes:

- 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 works or went to school) are fictional, but they have been carefully chosen to be representative of a typical user and to 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.

To address data management more specifically, for each 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). Processes shown shaded out are not performed by the persona; those shown in smaller or italicized font are performed but at less than best practice. Solid lines represent workflows performed by the persona. Curved 3D lines represent flows of data from one researcher to another. Note that the data 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 data that others have preserved (shown by the red curved arrows in Figure 2).

The figures in the example persona show that Sun — at present — analyzes data, plans for data collection, collects data, does data assurance and description to a lesser degree, and only a limited amount of data preservation. With DataONE tools, Sun could do a better job of data assurance and description and preserve her data so that other researchers can discover and integrate them with their own data for their own analyses; and conversely, she could discover and integrate data from others.

JeSLIB 2015; 4(2): e1082
doi:10.7191/jeslib.2015.1082

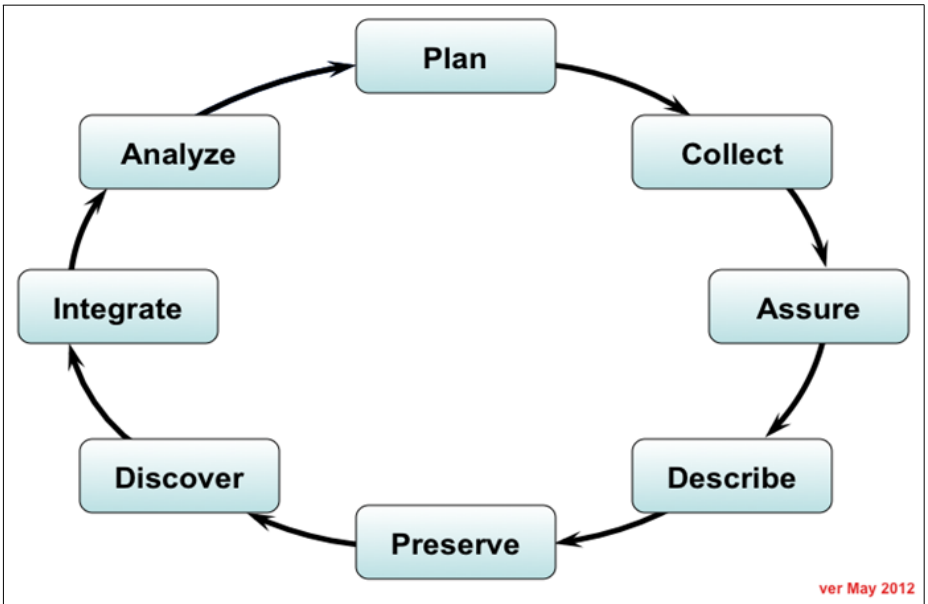

Figure 1: The DataONE data lifecycle, from Figure 7 Michener et al. (2012).

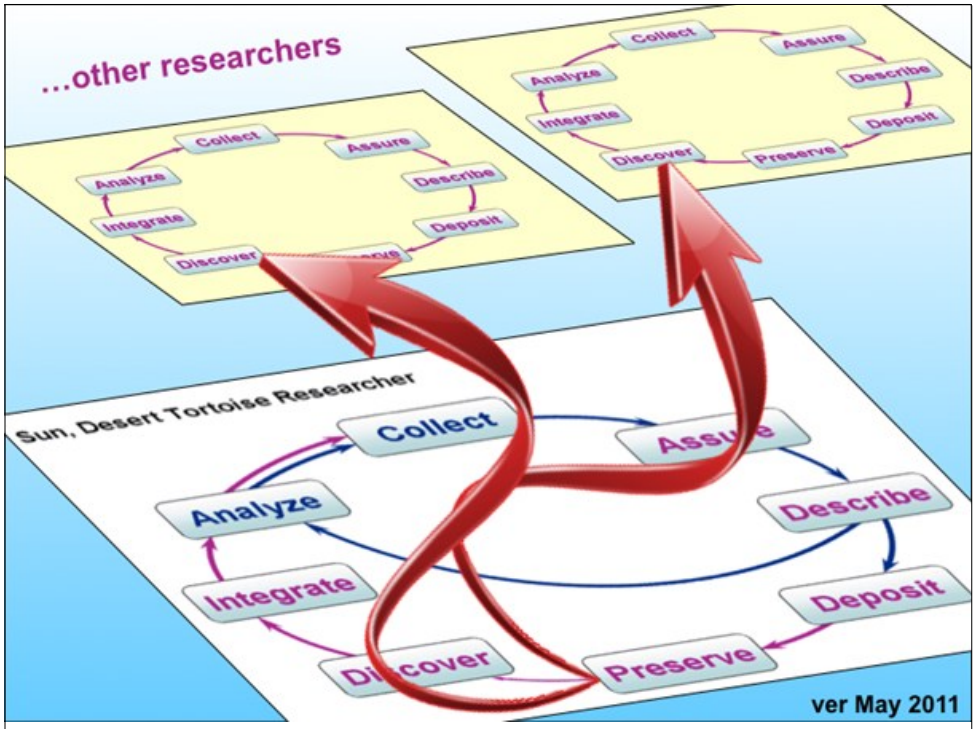

Figure 2: Red wavy lines represent data flows from the focus persona to others.

JeSLIB 2015; 4(2): e1082
doi:10.7191/jeslib.2015.1082

Results: Personas for DataONE

We developed 11 personas for DataONE (see Table 1). The full set of personas can be found at https://www.dataone.org/user-personas. Six are primary personas, describing the main intended users of the DataONE cyberinfrastructure and project tools. Five of the six are research scientists, developed to cover multiple dimensions that affect how scientists would share or reuse data and 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, modeler
- Data: Human vs. machine-collected
- Data management skills: novice to expert

The sixth primary persona is for a science data librarian.

The other five personas are secondary personas, describing other kinds of people who might be clients for the DataONE cyberinfrastructure and tools, but whose needs are only served if doing so does not get in the way of serving the primary personas. These personas include citizen scientists, administrators and educators. We have not described any negative or buyer personas (e.g., a grants officer who might fund researchers using DataONE). A number of additional personas were suggested (e.g., a scientist at a non-profit, a policy maker, a member node manager, a graduate student), but these were a lower priority for the development team to understand (indeed, most were secondary personas). Developing the 11 personas took a group of five core researchers about eight days of work, with lesser contributions from a number of others, spread out over several years of the DataONE project, not counting the time taken to do the original user research.

Table 1: DataONE personas

Primary personas	**Secondary personas**

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

Conclusion

Personas provide a tool to create 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. As Allard (2012) stated, the personas "allow developers, LIS professionals, and DataONE management to visualize how users from specific communities may use DataONE." The development team is currently using personas to group together related use cases that should be supported in a particular release and for planning future releases, and to identify which kinds of users should be involved in system testing. The community engagement team has found them useful as a way to engage potential new users by showing that the system was designed for people with their current experiences.

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

Supplemental Content

Appendix
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2015.1082 under "Additional Files".

Acknowledgements

The DataONE personas were developed by Ahrash Bissell, Kevin Crowston, Bruce Grant, Maribeth Manoff, and Rebecca Davis a team of members from the DataONE Sociocultural Issues Working Group, with support input and feedback from other members of the DataONE team. Andrea Wiggins and Sandra Henderson developed the citizen science personas.

Funding Statement

DataONE is supported by US National Science Foundation Awards 08–30944 and 14–30508, William Michener, Principal Investigator; Matthew Jones, Patricia Cruse, David Vieglais, and Suzanne Allard, Co-Principal Investigators.

Disclosure

The authors report no conflict of interest.

References

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

Atkins, Daniel E., Kelvin K. Droegemeier, Stuart I. Feldman, Hector Garcia-Molina, Michael L. Klein, David G. Messerschmitt, Paul Messina, Jeremiah P. Ostriker, and Margaret H. Wright. 2003. "Revolutionizing science and engineering through cyberinfrastructure." *Report of the Blue-Ribbon NSF Advisory Panel on Cyberinfrastructure* February. https://www.nsf.gov/cise/sci/reports/atkins.pdf

Branch, Benjamin D., Carol Tenopir, Suzie Allard, Kimberly Douglass, Lei Wu, and Mike Frame. 2010. "DataONE: Survey of Earth Scientists, To Share or Not to Share Data." Abstract IN11A-1062 presented at Fall Meeting, AGU, San Francisco, California, 13-17 Dec ember. http://abstractsearch.agu.org/meetings/2010/FM/IN11A-1062.html

Brooks, Frederick P., Jr. 1987. "No Silver Bullet: Essence and Accidents of Software Engineering." *IEEE Computer* 20:10-19. http://dx.doi.org/10.1109/MC.1987.1663532

Cooper, Alan. 1999. *The Inmates Are Running the Asylum*. Indianapolis, IN: SAMS.

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

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* 534:159-172. http://dx.doi.org/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." *A Data Conservancy White Paper.* http://dlsciences.org/research/DataConservancy/DC+Requirements+White+Paper.pdf

Dombrowski, Quinn. 2014. "What ever happened to Project Bamboo?" *Literary and Linguistic Computing* 29:326-339. http://dx.doi.org/10.1093/llc/fqu026

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. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf

Lage, Kathryn, Barbara Losoff, and Jack Maness. 2011. "Receptivity to library involvement in scientific data curation: A case study at the University of Colorado Boulder." *portal: Libraries and the Academy* 11:915–937. http://dx.doi.org/10.1353/pla.2011.0049

Madsen, Sabine, and Lene Nielsen. 2009. "Exploring Persona-Scenarios: Using Storytelling to Create Design Ideas." In *Human Work Interaction Design: Usability in Social, Cultural and Organizational Contexts*, edited by Dinesh Katre, Rikke Orngreen, Pradeep Yammiyavar, Torkil Clemmensen, 57-66. Second IFIP WG 13.6 Conference, Pune, India, 7-8 October. http://dx.doi.org/10.1007/978-3-642-11762-6_5

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. http://dx.doi.org/10.1016/j.ecoinf.2011.08.007

Schwaber, Ken, and Jeff Sutherland. 2013. "The Scrum Guide™: The Definitive Guide to Scrum: The Rules of the Game." Last modified July. 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* 4:93-103. http://dx.doi.org/10.2218/ijdc.v4i3.117