Ecological Informatics 11 (2012) 5–15

Contents lists available at SciVerse ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Participatory design of DataONE—Enabling cyberinfrastructure for the biological and environmental sciences

William K. Michener [a,*], Suzie Allard [b], Amber Budden [c], Robert B. Cook [d], Kimberly Douglass [b], Mike Frame [e], Steve Kelling [f], Rebecca Koskela [c], Carol Tenopir [b], David A. Vieglais [g]

[a] University Libraries, MSC04 2815, University of New Mexico, Albuquerque, NM 87131-0001, United States
[b] College of Communication and Information, University of Tennessee, 1340 Circle Park Drive, 423 Communications Bldg., Knoxville, TN 37996-0341, United States
[c] DataONE, 1312 Basehart Drive SE, Albuquerque, NM 87106, United States
[d] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37830, United States
[e] U.S Geological Survey, 230 Warehouse Road, Oak Ridge, TN 37831, United States
[f] Cornell Lab of Ornithology, 159 Sapsucker Woods Rd, Ithaca, NY 14850, United States
[g] Biodiversity Research Center, University of Kansas, Lawrence, KS 66045-7593, United States

ARTICLE INFO

Article history:
Received 15 June 2011
Received in revised form 20 August 2011
Accepted 24 August 2011
Available online 3 September 2011

Keywords:
Data centers
User-centered design
Stakeholder assessment
Usability testing
Data integration
Data life cycle

ABSTRACT

The scope and nature of biological and environmental research are evolving in response to environmental challenges such as global climate change, invasive species and emergent diseases. In particular, scientific studies are increasingly focusing on long-term, broad-scale, and complex questions that require massive amounts of diverse data collected by remote sensing platforms and embedded environmental sensor networks; collaborative, interdisciplinary science teams; and new approaches for managing, preserving, analyzing, and sharing data. Here, we describe the design of DataONE (Data Observation Network for Earth)—a cyberinfrastructure platform developed to support rapid data discovery and access across diverse data centers distributed worldwide and designed to provide scientists with an integrated set of familiar tools that support all elements of the data life cycle (e.g., from planning and acquisition through data integration, analysis and visualization). Ongoing evolution of the DataONE architecture is based on participatory, user-centered design processes including: (1) identification and prioritization of stakeholder communities; (2) developing an understanding of their perceptions, attitudes and user requirements; (3) usability analysis and assessment; and (4) engaging science teams in grand challenge exemplars such as understanding the broad-scale dynamics of bird migration. In combination, the four approaches engage the broad community in providing guidance on infrastructure design and implementation.

1. Introduction

Numerous grand environmental challenges face humankind in the next decades including climate change, decreased water availability, and loss of biodiversity and ecosystem services. The science required to understand and mitigate these challenges will require unprecedented access to data that cross science domains, distance (meters to the biosphere), time (seconds to millenia), and scales of biological and physical organization. Principal data challenges lie in discovering relevant data, dealing with extreme data heterogeneity, and converting data to information and knowledge. Addressing these challenges requires new approaches for managing, preserving, analyzing, and sharing data.

In this paper, we describe the DataONE cyberinfrastructure platform and how it is designed to serve as the basis for integrative biological and environmental research. Next, we describe how stakeholder communities were identified and further discuss four participatory, user-centered processes influencing the design of DataONE: (1) understanding community perceptions, attitudes, and user requirements via stakeholder baseline and repeat assessments; (2) creating personas and user scenarios that focus on how stakeholders (e.g., scientists) spend their time and the ways in which they may benefit from DataONE infrastructure and resources; (3) employing usability studies to better understand user requirements and to improve user interfaces and system interoperability; and (4) engaging scientists in grand challenge research exemplars to better understand the challenges they experience in discovering, acquiring, managing, and analyzing data. Finally, we conclude with a vision for how stakeholders can be involved throughout the lifespan of community cyberinfrastructure development efforts, highlight some lessons learned, and provide recommendations for future efforts so that stakeholders may directly influence system architecture design and implementation approaches.

* Corresponding author. Tel.: +1 505 814 7601; fax: +1 505 277 2541.
E-mail addresses: wmichene@unm.edu (W.K. Michener), sallard@utk.edu (S. Allard), aebudden@dataone.unm.edu (A. Budden), cookrb@ornl.gov (R.B. Cook), kdougla2@utk.edu (K. Douglass), mike_frame@usgs.gov (M. Frame), stk2@cornell.edu (S. Kelling), rkoskela@unm.edu (R. Koskela), ctenopir@utk.edu (C. Tenopir), vieglais@ku.edu (D.A. Vieglais).

1574-9541/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2011.08.007

2. DataONE as an integrative platform for biological and environmental research

DataONE is designed to provide an underlying infrastructure that facilitates data preservation and re-use for research with an initial focus on biological and environmental sciences. DataONE is unique in that it: (1) builds on existing data centers, leveraging the global investment in scientific data preservation; (2) creates a global, federated data network by focusing on interoperability solutions, providing tools and services to enable new science and knowledge creation; and (3) facilitates evolving communities of practice enabled by the DataONE cyberinfrastructure (CI) and informed by best practices, exemplary data management plans, and tools that support all aspects of the data life cycle.

The cyberinfrastructure implemented by DataONE is composed of three principal components (Fig. 1): *Member Nodes* which are existing or new data repositories that install the DataONE Member Node application programming interfaces (APIs); *Coordinating Nodes* that are responsible for cataloging content, managing replication of content, and providing search and discovery mechanisms; and an *Investigator Toolkit* which is a modular set of software and plug-ins that enables interaction with DataONE infrastructure through commonly used analysis and data management tools. Multiple instances of these three components operate together to provide a reliable fabric from which data may be retrieved by persistent identifiers, and contributions are guaranteed to be available indefinitely.

Data are principally acquired and maintained by Member Nodes that are located throughout the world. Member Nodes are envisioned to include a wide variety of institutions and organizations including natural history collections, Earth observing institutions, research projects and networks, libraries, universities, and governmental and non-governmental organizations. Each Member Node supports a specific constituency through its own implementation and often provides value-added support services (e.g., user help desk, visualization services). It is, therefore, expected that DataONE will accommodate highly geographically distributed and diverse Member Node implementations. Member Nodes extend the functionality of existing repositories by adding a standard set of application programming interfaces (APIs) that enable consistent user and Coordinating Node interactions. Adding functionality to any service requires resources for implementation and ongoing maintenance, and so the APIs are kept as minimal as possible. Additional feedback from current and potential Member Node operators indicated a preference for a staged approach for implementing the DataONE services, and as a result the APIs are now grouped into four tiers with increasing services. Tier 1 APIs provide the minimum necessary functionality for a Member Node to participate in the DataONE network as a read-only Member Node with little authentication or access control for users. Member Nodes conforming to the Tier 4 APIs offer full support for DataONE services including replication and access control.

Exemplars of Member Nodes offering a wide range of data and existing functionality are summarized in Table 1. Member Nodes are selected based on evaluation criteria that include factors such as diversity of data holdings, readiness to participate, community leadership, and resource availability. These early participants in the DataONE federation cover different information domains of relevance to integrative biological and environmental research. Participation in DataONE simplifies user access to these valuable resources through a common set of service interfaces and helps to ensure long-term access to the information through the use of persistent unique identifiers for all data and metadata.

Coordinating Nodes are designed to be tightly coordinated, stable platforms providing network-wide services to Member Nodes. These services include network-wide indexing of digital objects, data replication services across Member Nodes, mirrored content of science metadata present at Member Nodes, management of access control rules, and mapping of identities among different identity providers. The three initial Coordinating Nodes are located at Oak Ridge Campus

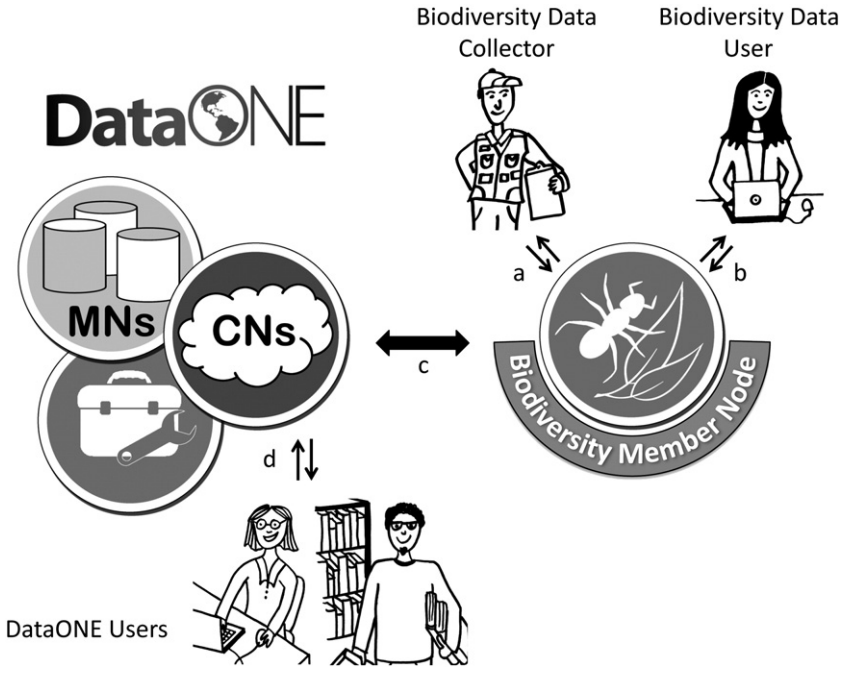

Fig. 1. Overview of DataONE cyberinfrastructure elements illustrating value derived to users associated with a Biodiversity Member Node. Biodiversity data collectors (a) deposit data into a biodiversity repository, which is a Member Node of DataONE. Biodiversity data users (b) can access these data directly from the biodiversity repository. The crescent around the repository represents the software stack that enables the Member Node functionality for the repository. This software stack is developed and installed by DataONE staff, making use of the characteristics of the biodiversity repository system and metadata. Biodiversity users can continue to access data as they did before (b). However, by linking up to the DataONE cyberinfrastructure (c), these data are accessible to a broader community of DataONE users (d) through DataONE enabled search and retrieval. The cyberinfrastructure components include the Coordinating Nodes (that maintain a metadata catalog of data held in the biodiversity repository), additional Members Nodes (for data replication and preservation and for integrated data search and retrieval) and the Investigator Toolkit (which provides increased functionality through integrated software elements such as Mendeley, Zotero, VisTrails, Kepler, Excel, R and others).