DATA SHARING AS A "BROADER IMPACT": RESULTS FROM THE SITE-BASED DATA CURATION PROJECT AT YELLOWSTONE NATIONAL PARK

CONTACT DETAILS:
Project Name: Site-Based Data Curation at Yellowstone National Park
NAME: Carole L. Palmer1, G. Sayeed Choudhury2, Andrea K. Thomer1
ORGANISATION: [1] Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign; [2] Digital Research and Curation Center at the Sheridan Libraries, Johns Hopkins University
EMAIL: clpalmer@illinois.edu; sayeed@jhu.edu; thomer2@illinois.edu

Objectives:

To date, research on scientific data curation has tended to focus on the curation requirements of researchers as data producers and the interests of institutions, disciplines, and data archives. Less attention has been paid to the curation needs of scientifically significant sites, such as national parks and other federally managed lands, where many scientists conduct research and collectively generate large amounts of data every year. Data curation best practices are needed that accommodate the special scientific and organizational considerations of these sites and improve access and reuse of the highly valuable data produced over time. Sites of data production, data producers, and data centers need to work together to develop effective divisions of labor and workflows that support both reuse for scientific research and coordination for site resource management.

The Site-Based Data Curation (SBDC) project is developing a framework of guidelines and processes for the curation of research data generated at scientifically significant sites, concentrating first on geobiology data produced at Yellowstone National Park (YNP). We aim to improve the preparation and description of data by researchers, reduce curation burdens for data centers and archives, and advance the preservation, access, and utility of data resources for scientific and site management purposes.

On-going activities:

The project is a collaboration among information scientists and geobiologists at the University of Illinois, data archiving experts at Johns Hopkins University, and resource managers at Yellowstone National Park (YNP). Project work to date has focused on stakeholder analysis, development of a data practices questionnaire, and analysis of an extensive and diverse set of YNP data produced by one prominent geobiologist over a ten-year period. Currently the team is developing model curation processes and workflows for the transfer of data from scientists to data centers for preservation and access. Three core research questions are under investigation:
What curation units and data parameters are needed for scientists vs. resource managers?
How should continuing series be curated and managed?
How can site and repository policies and processes be aligned?

Stakeholder analysis is informing SBDC requirements through engagement with geologists, geochemists, and microbiologists conducting research at YNP and park personnel, including managers of research permitting and reporting, and information professionals from the YNP research library and archive. A workshop held in April 2013 brought together nine researchers and seven YNP representatives where the SBDC team conducted focus groups with researchers and resource managers, an exercise on integrative science and data sharing, and roundtable discussions.

Results:

Stakeholder analysis shows agreement on the high scientific value of aggregating data produced at YNP for: study of dynamic systems, variability in geological features, and site evolution, and for assessing anomalies. Researchers emphasized data sharing as an important "broader impact" of their research, a way to make progress on "big picture" research questions, and improve collaboration among the many independent investigators working within

YNP. Park resource managers expect benefits from improved coordination and transparency of data collection activities and easier identification of trends and connections across projects.

Focus groups and roundtable discussions determined that data reuse will depend on detailed and consistent records of sampling processes and that SBDC guidelines must not inhibit individualized approaches to data collection, laboratory experimentation, or analytical and computational methods. The team identified minimum and optimal parameters for data description, to be vetted by the community, with core metadata consisting of GPS coordinates, sampling date, geological feature name, precise sampling position in relation to the geological feature, and temperature and Ph data where appropriate. In addition, a strategy has been developed for organizing data around sampling events, with photographs as a key component for reliable capture of location context and conditions. Next steps will include promotional materials for YNP researchers on data curation and broader impacts of data sharing.

URL http://cirssweb.lis.illinois.edu/Project/project-details.php?id=41