Surf's Up: Riding the Big Data Wave

Ashley Sands[1], Christine L. Borgman[1], Laura A. Wynholds[1], Sharon Traweek[2]
University of California, Los Angeles, 1. Department of Information Studies, 2. Gender Studies and History

From afar, the surf appears calm. By watching steadily, over long periods of time, it becomes apparent that the waves of big data are creating dangerous conditions for researchers, users, funding agencies, and public policy.

We use ethnographic methods to follow the builders and users of the Sloan Digital Sky Survey (SDSS), known as "the human genome project of astronomy." SDSS was designed as an open data project and it is among the most successful big data projects in science. We are spotting the indicators of how well astronomers are riding the big data wave - and when they are wiping out - with consequences for other creators, users, analysts, managers, and funders of big data systems.

Shooting the Curl
Ethnographic Research Methods

Get Inside the Pipeline Follow the data and the people from the design of instruments through processing, analysis, publication, and curation.

Unload the Trunk The forgotten odds and ends at the bottom may be the treasures.

Be There Making and using big data are social practices with shared tacit knowledge.

Beginners Data practices are learned and circulated among practitioners.

Rankings Leadership is earned on the waves.

Board and Body Surfing Older technologies and techniques are embedded in the newer forms.

Storm Surf and Dirty Water Strategies in changing ecologies are negotiated among those on the waves.

Stories Matter Knowledge is transmitted in narratives; learn who tells what kinds of stories, who listens, when and where.

Shark Attack
Preliminary Findings

Despite the view from afar, even big data professionals bite it every now and then:

Wipeout Just "backing up" big data is insufficient for sustainability. Important datasets are being lost due to lack of curation.

Shark Attack While astronomy has few privacy and confidentiality concerns, other fields of big data are susceptible. Data integrity is a concern everywhere.

Rip Tide Managing big data leads to unforeseen traps—big data is often comprised of many small and complex datasets, each with their own problems of interpretation.

Hang 10 Data Managers ready for the long haul keep scalability in mind—long boards are the only reasonable investment at this stage in the game.

Generations The demographics of big data elites are changing globally.

Taking Turns Defining and maintaining standards is an ongoing social practice among a distributed global community.

Surfing the Wedge
Policy Implications

No Lifeguard on Duty **Who is responsible for caring for data? When do these responsibilities begin?**

Don't be a Beached Whale **Researchers want to secure their data in repositories, but they don't know where to begin.**

Keep the Beach Clean **Data must be processed before they can be usable by others. SDSS has a pipeline for cleaning data, but individual researchers do not.**

The Tide is Always Changing **Research data are rarely "finished products" and it is unclear how to manage the ebb and flow.**

Public Beaches **Open Data get used. Astronomy researchers now share data and other resources such as code and repositories. SDSS has spawned big data innovations such as Galaxy Zoo for citizen science and its architecture has become the basis for other big data repositories.**

Charging for Parking **The scope of "open" remains negotiable.**

The Tide has Turned **Closed data is a bogus localism.**

The Big Kahuna **Long-term, in-depth research on big data practices can identify policy payoffs and policy wipeouts.**

Krill: This research is funded by the U.S. National Science Foundation ("Data Conservancy" OCI0830976, S. Choudhury, PI, Johns Hopkins University, and "Knowledge & Data Transfer: the Formation of a New Workforce" # 1145888. C.L. Borgman, PI; S. Traweek, Co-PI) and the Alfred P. Sloan Foundation ("The Transformation of Knowledge, Culture, and Practice in Data-Driven Science: A Knowledge Infrastructures Perspective" # 20113194. C.L. Borgman, PI; S. Traweek, Co-PI).

Knowledge Infrastructures http://knowledgeinfrastructures.gseis.ucla.edu/

Photo courtesy of puuikibeach http://www.flickr.com/photos/puuikibeach/3957417065/