Beyond Data Management: Designing User-Driven Data Services at UCSF Library

Ariel Deardorff
University of California San Francisco, USA

Abstract

As the biomedical sciences grow more data intensive, scientists and researchers are increasingly being expected to work with larger, more complicated datasets. UCSF Library, the only library on the health sciences campus, wanted to expand its data services to ensure that the university's students, staff, and faculty were prepared to work with their research data throughout the data lifecycle. Because this was a relatively new area for the library, the data services team decided to assess the data needs of the university in order to determine which programs they should offer. The data needs assessment relied on a mixed methods approach combining informal community feedback with focus groups. As the goal was to let the community guide the creation of a new service, the data services team used open-ended questions to reveal service gaps and data challenges, as well as resources and tools that the UCSF community desired. Findings indicate that the UCSF community is very interested in workshops and classes on programming with R and Python, as well as data organization and assistance finding open datasets. The findings of this needs assessment will help the UCSF Library's data services team design and prioritize new programs.

Data and the Biomedical Sciences

It is no secret that biomedical and health science research is more data-intensive than ever before. On the basic science side, researchers now have the ability to analyze large genomic datasets to reveal the cause of diseases at the molecular level. On the clinical side, researchers are turning to electronic health records as a source of patient data that can be mined for insight into how diseases spread and are cured. In order to be proficient in these new areas, researchers are increasingly required to have programming or other technical skills in order to run large-scale analyses, query large datasets, or mine patient data.

As biomedical research data becomes increasingly complex, funders like the National Institutes of Health are putting pressure on researchers to properly manage and share the data they collect, recommending data management plans and requiring certain kinds of research data (including human genomic data) be deposited into publically accessible data repositories.[1] On the publishing side, journals like *PLOS*, *Science*, and *Nature*, have created data sharing policies[2] that require researchers to make the data underlying their publications openly available. These data policies and requirements put new pressures on researchers to properly document, track, organize, and store their research data.

UCSF Library

The University of California San Francisco is a graduate only, health sciences university that includes four professional schools (dentistry, medicine, nursing, and pharmacy) and 17 biomedical graduate programs. In addition to the 3,100 enrolled graduate students, UCSF serves more than 1,500 clinical residents, and 1,600 postdocs.[3] As a research-intensive university, UCSF is particularly affected by the growing data-intensive nature of the sciences. While current graduate students may have topics like programming and database design built into their coursework, many of the university's postdocs, faculty, and staff were never taught these essential skills, and are constantly playing catch up in order to be effective researchers.

The UCSF Library, as the sole library on the UCSF campus, saw this lack of data-related skills as an opportunity for the library to offer essential services not provided elsewhere on campus. To ensure that new programs or services truly fit the needs of the UCSF community, library management decided to enlist the help of the assessment librarian to perform a data needs assessment.

Methods

The data needs assessment was originally designed to have a mixed methods approach consisting of

three stages. First, an informal idea-generating stage consisting of pop-up whiteboards around campus, then a formal survey to solicit campus-wide feedback, and finally a series of focus groups in order to gather feedback on potential service models. Once the project was launched it soon became clear that the chosen assessment methods were not well suited to the kinds of information and feedback that was desired. The initial idea-generating stage—which consisted of white boards and easels with questions like "What is your biggest data challenge?" and "What tools would you like access to at UCSF?"—generated only a few superficial answers (i.e., "too much stuff" and "not enough money"). This was surprising as the whiteboards were prominently located at places where people often congregate and were likely seen by many people. After trying different versions of the questions and various locations around campus, the assessment team decided that these kind of questions required more reflection than could be demanded of someone waiting in line for their coffee. Unlike answering a simple yes or no question, the whiteboard questions required respondents to think deeply about their workflows and research processes. The lack of response might also have been due to the competitive nature of UCSF, where people might feel uncomfortable describing their research challenges in a public forum. In order to get the truly rich information that they were looking for, the assessment team decided to proceed to the focus group stage of the project.

Because the assessment team thought that people would be more willing to participate in informal gatherings, they designed the focus group as informal "data discussions," where the goal would be to meet with various groups on campus (over lunch) to learn more about their research data needs and challenges. During the focus groups two members of the assessment team met with groups of between one and three people and asked them to describe their research. Participants were asked to talk about the structure of a typical day, who they collaborated with, the kinds of data that they worked with, and what kinds of tools and services they used. One of the team members took notes and the other listened carefully to the speakers and prompted them to elaborate on any frustrations or challenges they described. The format of the informal focus group worked particularly well in this situation as the relaxed small-group setting made participants feel comfortable sharing their struggles and allowed them to build off of each other's remarks. In all, the team met with three faculty members, five research staff, one graduate student, two postdocs, and two clinical research fellows.

Findings

The needs participants shared in the data discussions touched on all of the different aspects of the research lifecycle and can be summarized as difficulties with collecting data, processing/analyzing data, storing data, documenting data, and sharing data.

Collecting Data

Most of the participants that worked with clinical data specifically mentioned the difficulty of extracting data from APeX, UCSF's electronic health record system. A research staff member shared that they never knew what was in the system and what could be extracted, while another staff member told the group about the difficulty of extracting the same data each time the system was queried. Those who were not using UCSF data were not any better off; one faculty member told us it had taken months to receive data they had requested from the California Department of Health, a delay that severely impacted the timeline of their research project.

Processing/Analyzing Data

The comments related to processing and analyzing data often spoke to a lack of expertise in statistical tools and programming languages that made it hard for researchers to clean and analyze data on their own. A faculty member shared that their lab runs all experiments in R (programming language) and it can be a high bar for new grad students who often come in with little to no exposure to the language. Another theme that emerged was the frustration with MyResearch, UCSF's virtual research environment. At least four participants specifically told us how much they dislike using MyResearch and one clinical fellow even said it would be easier to drive across town and hand deliver a dataset rather than trying to upload and share it via the tool!

Storing Data

Participants who worked with clinical data were especially frustrated with the tools available to them for storing their data in HIPAA-compliant environments. Research staff and faculty members shared that restrictions on cloud-hosted software have led them to FedEx external hard drives and store confidential information in their (secure) e-mail inbox. What is more, data storing restrictions

make it harder to manage datasets within labs and with collaborators at other universities and in industry.

Documenting Data
Documenting data was a challenge that almost every participant discussed. The assessment team heard about labs where everyone organizes and describes their datasets differently, and where standards and protocols are passed down almost as an oral tradition. Postdocs talked about attempting to build on the data of a former lab member and not having any metadata or documentation to tell them how the experiments were run or what analysis was already performed. A grad student who had been in several labs reported that every lab was different and that it was necessary to rely on other grad students and postdocs to figure out the various system and protocols.

Sharing Data
Data sharing requirements are still relatively new and therefore were not on the radar of many of the participants. Those who had been required to share their data complained about how much time it took to reformat their datasets to meet the file formats and standards of their intended repositories. Other faculty knew they were supposed to share but reported that no one really did because there were not yet any real penalties for not sharing.

New Library Services

The data discussions gave the assessment team insight into several areas of need related to research data. While the library could not address all the issues raised in the meetings (MyResearch and APeX are not run by the library, for example), it could expand its educational offerings in areas like programming, data organization and storage, and data sharing. Since the needs assessment was conducted, the data services team has recruited instructors from inside and outside the library to offer quarterly Software Carpentry R/Python programming workshops along with monthly R/Python work sessions to provide opportunities for people interested in improving their programming skills. These workshops have been incredibly popular; registration for the first four sessions filled up immediately and there are often more than 40 people on the waitlist. On the data storage side, the library is currently planning an SQL workshop that will teach participants how to work with databases. In order to highlight tools for data organization and documentation, the library recently held an electronic lab notebook fair that was attended by over 45 researchers from across the university. While the library has yet to address the unique data challenges of clinical researchers, there might be an opportunity to partner with MyResearch and APeX developers to share feedback or even just offer training and orientations on those tools.

Although the data discussions did reveal real areas of need on campus, there were only 13 participants, and there are likely several other issues that were not discussed. The data needs assessment must therefore be a continual process to ensure that the library's offerings are consistent with campus needs. Ongoing assessment strategies include measuring attendance at workshops and classes, monitoring requests for new classes, and continuing to engage with campus discussions around research data.

Endnotes

1. "NIH Data Sharing Policies," *U.S. National Library of Medicine*, last modified October 21, 2014, https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_policies.html.

2. "Data Availability," *Public Library of Science*, accessed December 1, 2016, http://journals.plos.org/plosone/s/data-availability; "Science: Editorial Policies," *Science*, accessed December 1, 2016, http://www.sciencemag.org/authors/science-editorial-policies#dataavail; "Availability of Data, Materials, and Methods," *Nature*, accessed December 1, 2016, http://www.nature.com/authors/policies/availability.html.

3. "UCSF Overview," *University of California, San Francisco*, accessed December 1, 2016, https://www.ucsf.edu/about/ucsf-overview.