Campus Support Systems for Technical Researchers
Navigating Big Data Ethics

Complex data sets raise challenging ethical questions about risk to individuals who are not sufficiently covered by computer science training, ethics codes, or Institutional Review Boards (IRBs). The use of publicly available, corporate, and government data sets may reveal human practices, behaviors, and interactions in unintended ways, creating the need for new kinds of ethical support. Secondary data use invokes privacy and consent concerns. A team at Data & Society recently conducted interviews and campus visits with computer science researchers and librarians at eight U.S. universities to examine the role of research librarians in assisting technical researchers as they navigate emerging issues of privacy, ethics, and equitable access to data at different phases of the research process.

New Ethical Dilemmas
As noted, computer science researchers face new ethical dilemmas when they conduct big data research, especially research that uses social media data or scrapes “public” information off the web. The traditional model of seeking informed consent at the beginning of a research study is often insufficient when it comes to big data research. In addition, secondary use of human subjects data collected by a third party falls into a gray area: it is considered “exempt” and not reviewed by IRBs because the data was already collected. However, some researchers consider that a loophole and advocate for greater oversight of this frequent practice due to the threat of reidentification or privacy violations that become possible through the continued analysis or aggregation of the data.

The acquisition of online public data carries terms of service (TOS) requirements that raise logistical and ethical challenges such as replication, identification, and consent. Growing interest in web scraping of online data raises questions about the use of online information, rules of mass downloading of data, copyright, and legal access to data.

When making decisions about data storage, researchers must take into account current security issues as well as unknown future possibilities for data breaches and reidentification. Finding the right repository involves many factors. One university we visited offers a resource that matches project characteristics with the appropriate storage. Regardless of the storage location chosen, there is widespread and prevalent concern over whether data is truly secure. One researcher we interviewed bought his own servers to store data rather than using university servers, which can be accessed by IT staff. Once the data is anonymized and aggregated, he stores it on the university supercomputers. Beyond the most sensitive and well-protected data, ambiguity surrounds what instructions or criteria a researcher should follow in deciding when to take more protective measures. IT and other departments offer guidance, but inconsistencies and confusion remain, since advice may not always be sought, followed, or clearly conveyed.

There is a growing set of requirements for sharing raw data with journals for replicability and for sharing and disseminating federally funded research with the public for potential reuse. Fulfilling data-sharing mandates is complicated, ambiguous, and potentially risky. Sharing requirements cause concern about potential privacy issues such as reidentification. Some researchers fear that sharing will lead others to misinterpret or draw different conclusions from their data. Regarding these issues, Christine L. Borgman writes: “They [scholars] need tools, services, and assistance in archiving their own data in ways they can reuse them, which increases the likelihood that their data will be useful to others later.”

Formal Research Support and Mandates
The IRB is often seen as the campus legal and ethics oversight mechanism for protection of human subjects. While researchers may learn ethical principles through the restraints of the IRB and value its legal and procedural oversight, many researchers say the IRB is not the best mechanism for considering potential ramifications of big data ethics overall, since human subjects protections are just one component of ethics. IRBs struggle with questions such as whether deidentified data is human subjects data, how to assess whether data can be reidentified, and how to deidentify data while still retaining its research value. Secondary data use is generally considered exempt by IRBs and not part of traditional review, but changes to research methods resulting from big data have drawn this exemption into question as the distinction between primary research and secondary research has become increasingly blurry. In our interviews, IRBs were often criticized as lenient, bureaucratic, and slow, all of which can tempt researchers to cut corners.
Funder requirements for Data Management Plans (DMPs) were meant to encourage researchers to think through their work with data, but many see this as a hurdle. Assistance is available—such as the DMPTool (https://dmptool.org/)—and at one university we visited, representatives plan to have the library review all DMPs before proposals go to the funding agency. Many researchers we interviewed said they informally swap and copy language and see this as another item to check off a list.

Informal or inconsistent policies of publications and conferences leave researchers unsure, which affects their ability to publish or present their work. There are mixed opinions on what the role of journal review boards and conference program committees should be in determining whether submitted work is ethical and on what should be required to make their review process fair and consistent. Professional associations often lack review policies, leaving the protocol up to individual reviewers. Program committee reviewers have inconsistent approaches and often simply trust the researcher.

Informal Networks
Computer science researchers receive little to no formal or systematic ethics training during their education, compared with researchers in medicine or psychology. The former often use informal networks or conversations to make ethical decisions in their work, or they learn from their advisors in an apprenticeship relationship as they encounter issues for the first time. Requirements such as IRB training or Responsible Conduct of Research (RCR) provide some basics. However, researchers generally learn ethics on the job, through good and bad experiences, and from ad hoc conversations with other graduate students or peers.

Various formal and informal structures and services help to fill this gap on campus. Yet knowledge of these mechanisms is often shared simply through word of mouth; they are not always universally used and sometimes are made visible only following a violation.

Libraries’ Unique Position
Many research libraries have increased their Research Data Management (RDM) services in recent years. From what we saw in our project, libraries have several straightforward ways to increase their support for researchers. The legal use of information is sometimes complex to navigate, but libraries have been providing copyright, IP, and Creative Commons resources on campuses for a while. There may be a role for libraries to help researchers navigate murkier areas such as data ownership, TOS violations/advice, and web scraping concerns.

Libraries have increased their support of larger and more diverse files in their repositories over time. As needs for safe, secure, and long-lasting research repositories increase, more libraries will host robust data repositories or will partner on campus or with a consortium of organizations to create data repositories, especially for potentially sensitive data. As catalogers of knowledge, libraries need to be creating and thinking through metadata to safeguard the security and privacy of sensitive data sets. This metadata can help ensure that any sensitive data is wrapped with the proper descriptive information for future sharing.

When libraries advocate for open access, open science, and open data, they must take the next step and help support the means for making data open and sharable—they must have the difficult conversations about ensuring privacy and confidentiality and protecting against potential unintended future uses of data. As a profession concerned about privacy, intellectual freedom, and the public good, librarians have a unique role to play as we all figure out how we should handle data being collected about us, how we think about future uses of it, and where we go from here.

Training and Partnering
Across our interviews, we heard concern about these emerging ethical issues. Some institutions have started lecture series or are including a segment on data ethics in their classes. The Council for Big Data, Ethics and Society, a Data & Society initiative funded by the National Science Foundation, recommends embedded training within computer science classes as early as possible, integrating ethics training with the course materials and course projects rather than as a separate module, training, or course.4 At one campus we visited, a data clinic grew out of the statistics consulting clinic. Could a data-focused drop-in location be a checkpoint for helping researchers with their questions about the legality, privacy, or reproducibility of their work?

We also see a need for a centralized organization or initiative that can support researchers’ needs throughout the research lifecycle. This may be an opportunity for the research library, as a central hub. Librarians have a key set of values and skills. From offering training in data science to helping clarify gray areas, research librarians can benefit and support technical researchers as they navigate the emerging issues of big data ethics.

Notes
1. The Alfred P. Sloan Foundation funded this pilot project.

Bonnie Tijerina (bonnie@datasociety.net) is researcher and former fellow at Data & Society (http://www.datasociety.net) in New York City.

© 2016 Bonnie Tijerina. The text of this article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.