CSWS *Research Matters*

Diving into Big Data

Humanists bridge the qualitative vs. quantitative divide with help from Knight Library Data Services

BY JENÉE WILDE
CSWS Research Dissemination Specialist
Career Instructor, Department of English
University of Oregon

"I'm a very qualitative person, which is why it gets weird with all the quantitative stuff," said CSWS affiliate Courtney Cox, assistant professor of indigenous, race, and ethnic studies.

An ethnographic researcher of identity, technology, and globalization through sport, Cox has recently ventured into a new world of advanced analytics and algorithms in sports biodata, collected through wearable technology.

"I started thinking about high-level athletes and how that data could be used for them in terms of injury prevention and against them in terms of free agency," Cox said about her current project. "On the collegiate level where a lot of this new technology gets tried out, athletes don't have the choice to opt out of having this data tracked in a particular way. I haven't seen a whole lot of people doing that sort of work, so I wasn't even sure what I needed."

For help thinking through these issues, Cox contacted the Knight Library Data Services department for an initial tour of their data management offerings and to see how her work might fit into that space. She wanted to learn coding terminology that would enable her to interview biotechnology developers, as well as how to better manage transcripts, screenshots from websites, and "all the other stuff that sort of lives on my desktop," she said. "I'm a mess when it comes to that, so I decided when I came to Oregon I was going to get all that sorted."

Library staff in Data Services helped Cox through consultation sessions on best practices for data management and introductory classes in statistics and RStudio, a free software environment for statistical computing and graphics that has become increasingly popular among researchers in many academic disciplines as a tool for reproducible research.

"It was intimidating," Cox said of her first class, "because it brought me back to that statistics class in college that you kind of hate. I had one quantitative class that was required, and I just kind of cried all the way through it silently at home."

But Data Services' classes were an entirely different experience, Cox said. In addition to getting step-by-step coding help, learning work-arounds in open source software, and discovering add-ons to apps that could help her do her job more efficiently, she felt comfortable being in a room with people having a broad range of experience, all working together to answer questions. "It was a judgment-free zone, which was the opposite of my college stats experience," she said.

Courtney Cox, Assistant Professor, Indigenous, Race, and Ethnic Studies.

Cox came away from her classes with a needed confidence boost. "As someone who is not a numbers person," she said, "I feel like I'm the testimony of 'If I can do it, everyone can do it' because of my inherent mental block toward statistics and numbers."

When qualitative researchers see words like data services, data science, or statistical programming, said Jonathan Cain, interim director of digital strategies in the Knight Library, they may feel intimidated or discouraged about pursuing research

UNIVERSITY OF
OREGON

outside of their comfort zones.

"Just because it's data services doesn't mean that it's only statistics, that it's only quantitative," Cain said. "We want to help people figure out how to handle their research product for qualitative inquiry as well. There's no minimal requirement of knowledge with these tools."

The library's Data Services department formed about three years ago with a pilot project to help researchers submit their data to an open repository for evaluation before publication.

"We didn't really have an infrastructure for that at the time," Cain said. "Now we are developing a whole suite of services that can help people wherever they are, across the board."

Jonathan Cain, Interim Director of Digital Strategies, Knight Library.

With support from the library and deans in the College of Arts and Sciences, the initial pilot program grew into a cohesive set of services for research data management and reproducibility, as well as training and support for the range of ways that data is being mobilized by researchers, teachers, and students across campus.

Today, Data Services offers free consultation help with statistical software, GIS/ArcGIS, aerial photography, R/RStudio, survey software, Git, Docker, Excel and more to faculty, staff, graduate students, and undergraduates. In addition, they offer free training in a variety of tools for processing, cleaning, and visualizing data, version control with Git and GitHub, basic statistics, and using Excel and SPSS for statistical analysis, as well as best practices for data management across platforms.

Moreover, Data Services offers personalized help installing popular open-source apps such as R, Python, and MySQL and software available through the UO Software Center. They also provide help with citation tools such as Zotero and EndNote, as well as how to manage qualitative data with tools like Taggette, an open source software for tagging ethnographic interview transcriptions for themes of interest to the researcher.

"We're always trying to ask who is invited and who is not in data spaces," said Gabriele Hayden, research data management and reproducibility librarian in Data Services. With a background in English literature and Romance languages, as well as open source software and enterprise database reporting in the non-profit and private sectors, Hayden understands the importance of bridging disciplinary divides in today's digital landscapes.

"In a way," Hayden said, "they hired me and not a physicist because they wanted to be able to pull the humanities into this."

Cain agrees. "The humanities, the social sciences, the hard sciences— they're all taking a look at these materials as new ways of exploring interesting questions, but not everyone has necessarily had that opportunity within their formal or informal training, and not every department has access to the same resources to do that work. One of the roles of the library is to be this central node that can provide that aspect of support and skill building irrespective of where you're coming from as part of our community."

Cain added that really interesting questions are being explored now by humanists who may not have done quantitative or spatial work before, "but they are making use of our services, which is one of the things I'm extremely happy about," he said. "When we designed the department, it was

Allison Kunerth, Data Manager, SNAP Lab, Center for Translational NeuroScience.

very much with the mindset that these are skills that are extremely important for everyone to have a handle on."

Data Services is also committed to increasing access for women and minority communities who are underrepresented in the data science profession or who may have experienced barriers to getting those skills, Cain said. In addition to hiring diverse staff and GEs, they offer classes in languages other than English and work with McNair

Gabriele Hayden, Research Data Management and Reproducibility Librarian / photo by Randy Sullivan.

Scholars and the Undergraduate Research Symposium as ways of offering access to people who may not have a place to learn these skills or showcase their work in the field.

"There's lower representation of alternative voices and perspectives in developing algorithms and developing the parameters that are being used to make financial and administrative decisions," Cain said. For example, it's not surprising, he said, that self-driving cars have a hard time recognizing Black people because the images used to train algorithms are normalized with sample sets that contain mostly white faces.

"We're making more people conversant with the tools so they can have a stake in those sorts of discussions," Cain said.

On the quantitative side of research, Data Services provides support and instructional training to data science experts who are managing data in various ways for science research labs and departments on campus.

"I attended most of

Understanding Data Management
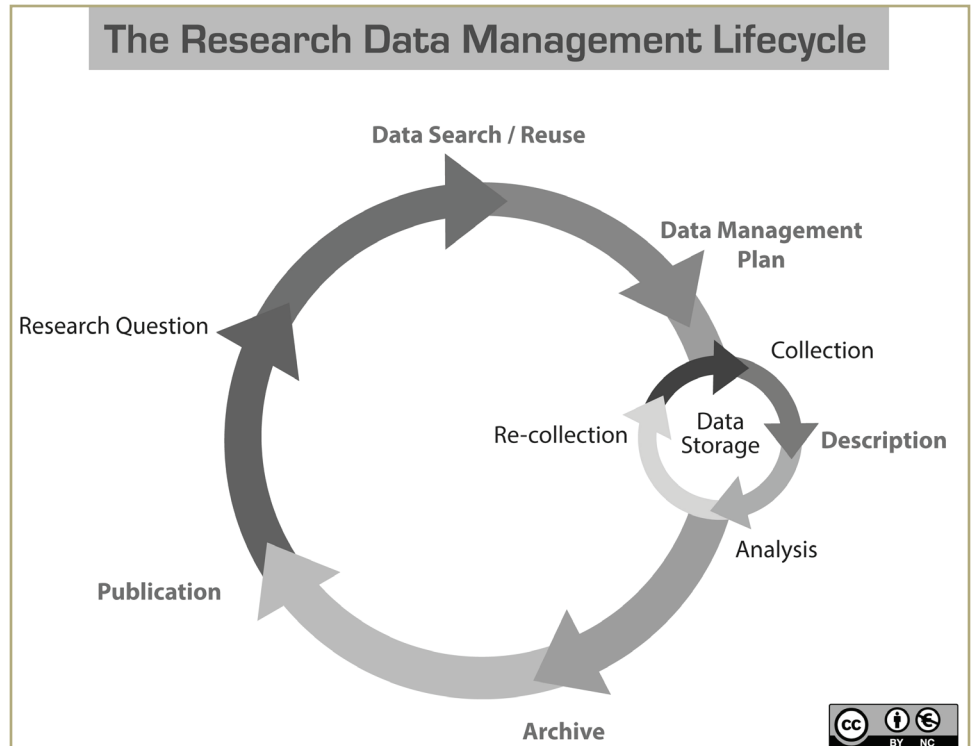## Researchers get help planning for the data lifecycle

Biocultural medical anthropologist Lesley Jo Weaver, assistant professor of international studies, recently sought help from the Knight Library Data Services with writing a data management plan for a National Science Foundation grant.

"Starting about five years ago, NSF started requiring that every proposal include a data management plan," Weaver said.

In 2013, the White House issued a directive on "Increasing Access to the Results of Federally Funded Scientific Research" in order to enable access to and validation of results by other researchers, to encourage researchers to expand on previous data, and to ensure that grand-funded data is preserved for the future. As a result, federal funders such as the NSF, the National Endowment of Humanities, the National Endowment for the Arts, the Institute of Museum and Library Services, the National Institutes of Health, and the National Oceanographic and Atmospheric Research now require data management plans in grant applications.

"My understanding of that move is that the government pays lots and lots of money for people to collect data that gets lost when their hard drive dies or they move institutions," Weaver said. "The idea is to hold on to the data so they can be reanalyzed later by the researcher, or analyzed by someone else, or archived so that in 30 years when somebody is a famous researcher and someone wants to go back and look at the trajectory of their career and reanalyze what they've done, that can happen."

In general, data management plans

The Research Data Management Lifecycle

Data Search / Reuse · Data Management Plan · Research Question · Collection · Re-collection · Data Storage · Description · Analysis · Publication · Archive

Lesley Jo Weaver, Assistant Professor, International Studies.

include a description of the project and the data that will be produced; how the data will be managed throughout; documentation about the data; plans for short-term data storage, backup, and security; legal and ethical issues; plans for access, sharing, and reuse of data; plans for data retention and disposal; and plans for preservation and archiving.

At Data Services, Weaver consulted with Gabriele Hayden, the department's new research data management and reproducibility librarian.

"One way to think of it," Hayden said, "is when you have data, there's a pipeline that happens with it when first you want to find data or make data, then sort it and keep track of it, and then analyze and preserve it. Then there's the second half of that pipeline where you want to share and communicate it with others. We help with all of those parts."

Hayden provided Weaver with grant application language to describe the UO's data security measures, such as what kind of off-site servers the institution has and what firewalls are set up for privacy, among other technical details.

"I had produced a data management plan for a prior institution," Weaver said, "but I hadn't done one here so I didn't know the kind of hardware and software they had access to for enacting that long-term storage, so part of the consultation was just fact-finding and part of it was identifying what staff people I could work with for long-term storage options."

Hayden consults with students, faculty, and researchers on research data management, sourcing and sharing data, and best practices in

the workshops that Data Services held so I could update some of my skills and learn a few more," said CSWS affiliate Allison Kunerth, data manager for the Stress Neurobiology and Prevention Lab (SNAP) in the Center for Translational NeuroScience. SNAP runs large-scale psychological studies with families in child development programs to determine how they improve over time.

While Kunerth would like to see a more standardized way for data to be managed across the university, "they did help point me in the right direction and help me brainstorm how to develop my own program," she said of Data Services. "I worked with Gabriele to help figure out a better way to develop a database for my research lab."

With some advice from Data Services, Kunerth is building tools for real-time access to the lab's ongoing research studies. She is developing a user dashboard with ShinyApps and Excel Power BI that allows researchers to pull up and track various metrics from their studies as they progress. By having continual access not only to quantitative metrics but also qualitative results, such as participant responses to open-ended survey questions, researchers can decide in the early stages of research if they should make changes in a study's data collection methods or adjust how they are analyzing results.

"It helps them to adjust the questions, dig deeper on something, or say, hey, we need to look at these open-ended questions now and see what's happening instead of waiting until the end when it's too late," Kunerth said. "The biggest take away is if you have your data managed well then you can actually get these researchers more information ahead of time, and it can hopefully inform better their qualitative assessments."

For information on data workshops, go to https://researchguides.uoregon.edu/library_workshops/.

computational reproducibility. She can help with a project at any stage of the research data lifecycle—from looking for ways to locate or reuse data to finding time-saving methods for collection and analysis, as well as thinking through options for securing and sharing raw data and research outputs.

"When I have an intake, part of what I want to find out is, where are you with it?" Hayden said of data management. "What is your level of existing competency and what is your interest in learning new things? Based on that, we can talk about best practices that are purely about how you do your backups, how you name your files, how do you set up your workflow—just project management stuff that will make your life easier in advance."

Hayden also can help researchers consider how they want to share their data and code, such as depositing them in open-access, search engine-optimized repositories like the Inter-university Consortium for Political and Social Research (ICPSR) or Scholar's Bank: Data, a new repository specifically for data and code.

"This sharing of your research is about the development of knowledge, but it also can help you develop your professional profile," Hayden said, adding that some research suggests increased citation advantages to sharing your data and code in open-access repositories beyond journal and book publication. "We as librarians understand a bit about how that works," she said, "so if you're interested in that, we can help in various ways."

For more information about data management planning, contact Hayden at ghayden@uoregon.edu or go to the library's website: https://researchguides.uoregon.edu/data-management/dmp.

—*By Jenée Wilde*