Coders at #DataRescueDavis archive federal, at-risk databases

EMMA SADLOWSKI — SCIENCE@THEAGGIE.ORG FEBRUARY 9, 2017

CHARLES MIIN / AGGIE

UC Davis volunteers compile archive of scientific data at hackathon

UC Davis community members gathered at the Peter J. Shields Library on Feb. 2 for #DataRescueDavis, a hackathon intended for archiving reliable scientific data. Volunteers spent the day on their laptops backing up data specifically related to climate change and the environment.

The hackathon was sponsored by DataRefuge, a public and collaborative project by the University of Pennsylvania that aims to safeguard scientific, federal data and ensure that data remain accessible to researchers around the world. The project has inspired communities across the country to host DataRescue events and to contribute to preserving public databases.

“These events are meant to pull down federally funded and publicly available data on federal websites and back them up in a location that isn’t under control by the government,” said Kevin Miller, the university archivist at UC Davis. “We want to ensure that any digital form or research dataset survives and is accessible over the long term.”

#DataRescueDavis began at 10 a.m. with an introduction of the project’s purpose and goals followed by a coding workshop. The volunteers were then split up into groups and worked on different coding tasks for the remainder of the day under the guidance of Shields Library staff members.

The DataRefuge Project kickstarted in December 2016 in anticipation of the U.S. presidential transition. Federally available data, such as the information found on the White House and U.S. Environmental Protection Agency (EPA) websites, are often updated whenever a president begins a new term, which can result in lost data. Shields Library staff members were inspired by the project's objective of archiving these data and began organizing an event for the UC Davis community shortly thereafter.

"Libraries are committed to ensuring the preservation of and public access to knowledge," said MacKenzie Smith, a UC Davis university librarian in an email interview. "It is critical that valuable scientific data remain available to researchers, so the UC Davis Library is offering its space and expertise in managing data to facilitate this important effort."

Volunteers at #DataRescueDavis were divided into two main groups. The "nominator" group identified federal websites that had information about environment, climate, datasets or public research. These volunteers picked out websites that were "web crawlable," websites that can properly index downloaded pages. When a web page was loaded with large data or media (an uncrawlable website), it was flagged and sent to the "scraper" group, who reviewed the websites and dug deeper to "scrape" out and manually download the data. The web crawlable data was backed up into the Internet Archive, a nonprofit digital library based in San Francisco, while the scraped data was backed up into an archive provided by DataRefuge.

Fernando Espinosa, a third-year neurobiology, physiology and behavior major, volunteered as a scraper at the event. He saw the project as a great opportunity for him to apply his knowledge of data science and coding toward a good cause.

"We know that we can prevent important scientific data from disappearing — data that can help us predict future climate change and other issues," Espinosa said. "We want to create a large enough database that can be accessible to not only Davis residents, but to data scientists around the world."

The project is currently collaborating with the Environmental Data & Governance Initiative and the Internet Archive's End of Term Project to archive data involving climate change and the environment. According to Miller, preserving this information is crucial to research institutions like UC Davis.

"A lot of strength and focus here at UC Davis is on the environmental sciences, and that's a part of the reason as to why we want to focus on those types of websites," Miller said.

Some volunteers at the event expressed concern about the possible censorship of these public databases. Rachel Baarda, a physics graduate student and another scraper at the event, emphasized the importance of protecting reliable data so that future generations may use them with assurance.

“I remember it was only last quarter when I was trying to do research on my own about climate change, and I was using the EPA’s website to look at their climate change links,” Baarda said. “At that point, I started thinking, ‘In a year, these data might not be here if they’re not backed up.’ If you Google it, you should be able to go to these government websites with reliable data because there’s so much misinformation out there. We want to back up what we know is accurate and reliable data so that they can be accessible in the future.”

#DataRescueDavis’s dedicated volunteers made the event a true success; even at full capacity, volunteers continued to code and web crawl while sitting in groups on the floor and up against walls.

“I think what’s driving these numbers is that people are sensing that they have a skill that they want to use for good,” Miller said.

Upcoming DataRescue events are scheduled to take place at UC Berkeley, Georgetown University, Haverford College and MIT. No coding experience is needed in order to volunteer with the project.

Written by Emma Sadlowski — science@theaggie.org