Terabytes of Government Data Copied

The latest stats on volunteers' efforts in recent months to preserve environmental information generated by federal science agencies

By Kerry Grens | March 8, 2017

Agulhas Return Current (ARC) Ocean Climate Station mooring
NOAA

In the past several months, a movement has sprung up among librarians, environmental and computer scientists, and supporters of access to public data to create archives of environmental information. While the volume of material is daunting (we're talking hundreds of millions of webpages), volunteers have made considerable headway, collecting terabytes (TB) of data to date, with more being collected all the time.

One team, called the Azimuth Climate Data Backup Project, had by February 11 backed up 19 TB of data from NASA, the National Oceanic and Atmospheric Administration (NOAA), and other federal agencies that collect climate-related data. "I know there is a gap, because there are several large datasets which are still being downloaded, mostly by me," Jan Galkowski, a statistician who contributes to the group, told *The Scientist* in an email. "We have the capability and will probably fill 40 TB of storage with the data we have replicated. This will take some time to move to its eventual homes, simply because network transfer speeds are not that high."

The election of President Donald Trump, and expectations that climate-related projects in particular would face cuts during his administration, has stoked fears that access to government environmental data may be in peril. "We're not going to see any disappearing data overnight," said Michelle Murphy, director of the Technoscience Research Unit at the University of Toronto. "One way to lose data is to close a program. . . . . [Its dataset] doesn't have to be deleted, it just becomes uncared for, goes offline, goes into a drawer."

In response to those concerns, Murphy helped start the Environmental Data and Governance Initiative (EDGI) several months ago. The volunteer-run project coordinates the collection and archiving of US government data through so-called data rescue events, in which people meet up to allocate expertise and computing power to saving particular datasets or websites. Dozens of these events have been held around the U.S. and Canada so far, with more scheduled.

EDGI partners with the DataRefuge network—an initiative launched by the Penn Program in the Environmental Humanities at the University of Pennsylvania—and the Internet Archive to store the webpages and data that participants collect. According to Dawn Walker, who works with EDGI, 1.7 Tebibytes (TiB) have been downloaded for inclusion in DataRefuge, including 158 datasets that are now available for download. Data rescue volunteers have also nominated 73,500 URLs from government

websites to be crawled by the Internet Archive. (Crawling is the process of downloading a website, then fanning out to each link from that site and downloading those, and so on.)

Jefferson Bailey, the director of web archiving at the Internet Archive, said some of these would likely have been included in the organization's already-scheduled End of Term crawl, which collects URLs from .gov and .mil at the turnover of each presidential term. But the efforts of EDGI are complementary. Four years ago, Bailey said, people nominated only around 1,400 URLs to crawl. "We can crawl, scale, and store a lot," he told *The Scientist*. "But we don't the have time and staff to host events." Working with EDGI, he said, has "been a good pairing."

In addition to the organized efforts of EDGI and Azimuth, concerned citizens have made their own individual contributions. Bryce Lynch, a Bay Area security specialist who previously worked at NASA, has been downloading sensor-buoy data from NOAA continuously for the past two months. He also participated in a local data rescue event.

"I've got maybe 29-30 terabytes here on my rack. . . . I'm devoting half of my bandwidth to downloading this stuff," he said. "I've been going at it for two months now. I'm not stopping anytime soon."

Nor is the guerilla archiving momentum. At Rice University in Houston last Saturday, around 75 volunteers spent the day searching for websites to be archived, or writing code to harvest data. Kathy Hart Weimer, head of the Kelley Center for Government Information, Data and Geospatial Services at Rice's Fondren Library, said she was inspired to organize the event knowing what had happened in previous administrations when the Environmental Protection Agency's budget was cut. "That caused some libraries to close," she said. "Librarians who remember that are attuned to federal budgets. We want to make sure information is maintained so not only scientists can access the data, but the public as well."

As people were packing up to leave the data rescue event, said Weimer, people were asking: "Are we going to do it again?" Maybe they will, she told *The Scientist*. "We'll see what happens next."

See "Science Policy in 2017"

Tags

trump, noaa, NASA, medical archives, EPA, environment, database, data storage, data sharing, climate change and climate

Add a Comment

You

Comments

Dr Edo
Posts: 30

March 9, 2017

The language and system into which the information will be stored must have some available Rosetta stone otherwise in a short time it becomes garbled. When the USGS was forced to consolidate its regional offices back into Washington (circa 1980s), and I was a USFS geologist (Los Padres) doing a groundwater recon of the forest, I happened to be at

down at the USGS Regional Office at Laguna Niguel. I noted racks of punch cards on the loading dock. Asking what this was, I was told it was all the driller logs for all the water wells in the State of California. What was going to happen to these cards, I asked? Answer----no provision to move them, thus into the adjacent dumpster. I requested that they first be read off onto a 16 mm magnetic tape and sent to the Supervisor's Office for Los Padres and this was done. But by the time we got the tapes, the language and format into which they were transcribed became obscured. I still have access to the tapes but no one knows the underlying basis of language or format that was used. Highly useful data, given the State's current emphasis on groundwater issues, but the data, although being in hand, are hidden.

Dr Edo McGowan

Sign in to Report

True Scientist
Posts: 59

March 9, 2017

Concerned environmental scientists collected TBs of information noise and want to keep it forever. I would like to see statistics about public access of these data bases. When number reaches 10 (if ever) I expect to see an article in this publication.

Sign in to Report

True Scientist
Posts: 59

Replied to a comment from Dr Edo made on March 9, 2017

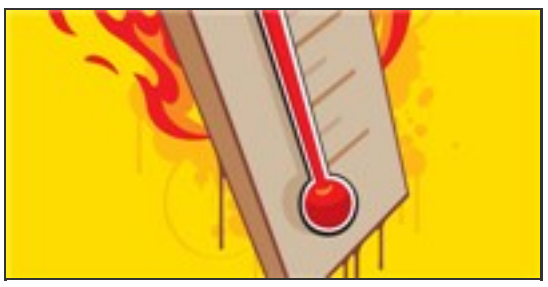
March 9, 2017

Nobody is talking about really important data in this case. It is all TBs of climate change garbish..

Sign in to Report

Related Articles

Congress Agrees to Give NIH $2 Billion



Notable Science Quotes

By *The Scientist* Staff

EPA Scrubs Climate Change Page from

Cur

View

Sub

Extra

By Kerry Grens

The proposed spending plan for 2017 includes money for Alzheimer's and cancer research.

Climate change, research funding, race, and much more.

Website

By Kerry Grens

The US Environmental Protection Agency removed information about global warming and greenhouse gas emissions that doesn't jibe with the Trump administration's views.

Al
Th
Ne
Ca