

UC Office of the President

UCOP Previously Published Works

Title

Data Mirror: Complementing Data Producers

Permalink

<https://escholarship.org/uc/item/2n1715ff>

Journal

Against the Grain, 29(6)

Author

Chodacki, John

Publication Date

2017-12-01

Peer reviewed

Data Mirror: Complementing Data Producers

by **John Chodacki** (Director, University of California Curation Center) <john.chodacki@ucop.edu>

Data Mirror is a collaborative project between the **University of California Curation Center (UC3)** and **Code for Science & Society (CSS)**, a non-profit organization committed to improving access to data for the public good. We are interested in preserving federal data because we know that the research produced, collected, or funded by the federal government are an integral part of the rich tapestry of the nation's cultural and scholarly record, and are critical resources for advancing scholarship, public policy, and governmental transparency and accountability. However, we in the library and preservation community often forget that the data producers within the federal government have comprehensive preservation strategies and workflows of their own. Although we are focused on helping solve problems, many times we unnecessarily create duplicative or parallel solutions that cut the federal research groups out of the conversation and can cause



additional issues down the road. The Data Mirror project (datamirror.org) is working to exemplify a different possible path forward.

Data Mirror is a complete, and routinely updated, copy of the main federal government research data portal, *data.gov*. Hosted by the UC3 at the **California Digital Library (CDL)**, Data Mirror points back to the “datasets of record” on federal agency websites for routine access. Why? Because those are the copies that are cared for and handled by the data producers themselves, and therefore, those copies should be referenced and used by researchers. However, should these access paths become interrupted or inaccessible, Data Mirror also includes pointers to

CDL-managed copies, as well as additional registered replicas hosted by other institutions. In this model, *data.gov* and the mandates that it works under remain the center of the workflow. Basically, Data Mirror works as a back-up of the

existing systems and offers redundancy to the *data.gov* metadata catalog and preservation services to its underlying datasets. Providing alternative search and retrieval opportunities helps to ensure that these important data remain available for study and use in perpetuity while keeping existing Federal workflows intact. Without building entirely new systems or processes, government research groups can continue to rely upon their existing workflows.

We have worked directly with the team at *data.gov* to ensure we are respecting their existing workflows. With the support of the wider library and preservation community, we would like to enhance the Data Mirror portal to include the ability for our communities to propose enriched metadata or the addition of new datasets through the portal, which would be communicated back to the agencies and *data.gov*. It is that round-tripping of federal data preservation (through existing channels!) that would truly build long-term collaboration between those producing government data and those focusing on the preservation of government data. 🐾