Environmental Data
& Governance Initiative

envirodatagov.org ■ EnviroDGI@protonmail.com

May 6, 2017

EDGI Introduction and Accomplishments

The Environmental Data and Governance Initiative (EDGI) is an international network of academics and environmental professionals that advocates for evidence-based environmental policy and robust, democratic scientific data governance. EDGI builds web-based tools, facilitates public events, and coordinates research networks to respond quickly to current political attempts to undermine the important work of federal environmental and energy agencies and policy.

EDGI began as a small group of concerned researchers in the weeks after the U.S. elections in November 2016. It has since grown into a network of 116 people from 30 institutions. EDGI members include social scientists with expertise in environmental science, history and policy; physical and life scientists with expertise in climate and environmental health; lawyers with expertise in environmental regulation; librarians and archivists with expertise in digital preservation; and open-access technical communities dedicated to public access to scientific data and analysis. EDGI collaborates with a growing list of organizations including Public Lab, 314Action, DataRefuge, the Internet Archive, and Climate Central.

EDGI projects include monitoring and analyzing changes to federal environmental agency websites, and researching the effects of environmental deregulation and changes at federal agencies, such as EPA, DOE, NASA, NOAA, and OSHA. EDGI also co-coordinates "DataRescue" events to proactively archive public environmental data and ensure its continued public availability. More information about our core working groups and emerging research projects can be found in the pages below.

EDGI is generously supported by the Doris Duke Charitable Foundation, The David and Lucile Packard Foundation, and by individual donations.

Highlights of our work:

- **Dramatically improved the Internet Archive's "End-of-Term" project of archiving of federal environmental websites.** Thanks to our effort, webpage nominations increased by 10-fold from those in 2012, and the EPA is now the Internet Archive's most

comprehensively archived federal agency. As of May 1, 2017, DataRescue events using EDGI's Chrome Nomination Tool nominated 63,076 web pages to the Internet Archive.

- **Established a project of monitoring federal environmental websites on a daily basis.** We started with approximately 25,000 web pages at agencies including EPA, DOE, DOI, NASA, NOAA, and OSHA, as well as whitehouse.gov and data.gov. This project will soon expand to monitor tens of millions of pages. Our team of 10 analysts has produced 15 reports that have informed over 20 new articles at outlets including *The Washington Post, ProPublica*, *Scientific American*, *The New York Times,* and *Business Insider*.

- **Launched a project of confidentially interviewing long-time employees at EPA and OSHA** (primarily retired employees), to illuminate the human side and historical context of the current political transition and its effects on federal agencies. Trained interview teams in Washington DC, Boston, the San Francisco Bay Area and elsewhere are currently conducting confidential interviews.

- **Provided rapid academic analysis on proposed regulatory changes** such as the HONEST Act and events such as Administrator Scott Pruitt's first address to the EPA. We are developing an overarching report on changes to environmental data and governance made in the first 100 days of the Trump administration.

EDGI aims to serve the environmental community and its allies, and to enable them to hold the new administration accountable by preserving and improving public access to at-risk government environmental data, documents, and digital interfaces, and by monitoring, documenting, and analyzing change to federal environmental agencies. We also aim to create an open, collaborative network of individuals, non-profits, universities and companies who believe that science and evidence-based governance are vital for human and environmental well-being.

For more information, including ways to get involved, please see the pages below, visit our website https://envirodatagov.org, or email us at EnviroDGI@protonmail.com.

Many Thanks,

The Environmental Data and Governance Initiative

Table of Contents

1. Website and Dataset Archiving

With the help of the Internet Archive and DataRefuge, we seek to preserve publicly accessible scientific data and archive webpage from EPA, DOE, NOAA, OSHA, NASA, USDA, DOI, and USGS. Between December 2016 and the end of April 2017, EDGI coordinated over 30 DataRescue events at cities across the U.S. and in Canada, with its partner, DataRefuge. Building on this incredible outpouring of support, EDGI convened a well-attended online townhall meeting on April 1st to discuss lessons learned from previous events and plan for the future. EDGI also hosts weekly "community calls" with its network of archivers.

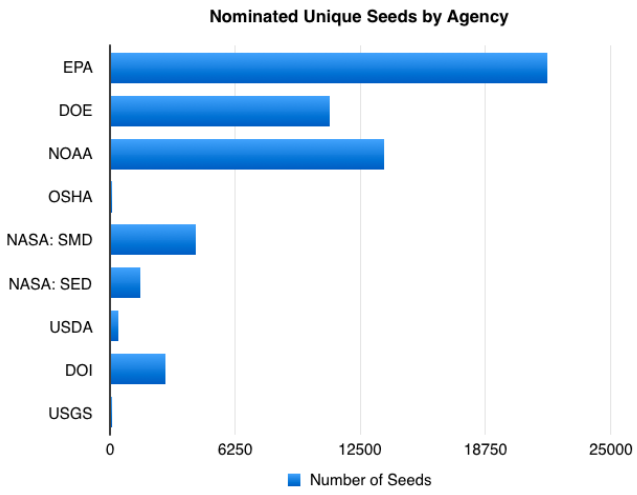
As of May 1, 2017:

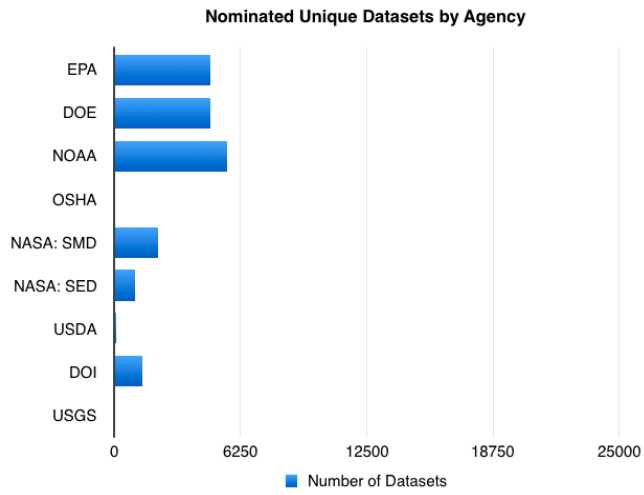
Number of Unique Webpages Seeded (i.e. added to the Internet Archive using the EDGI Chrome Extension): 63,076

Number of Pages with Datasets to be Archived*: 21,798 (34.6% of Unique Webpages Seeded)

*This number includes web pages that were previously referred to as "Uncrawlable", but also includes other types of datasets or repositories of information to be archived. Here is the breakdown:

Number of Pages with Many Files: 6,792 (31.2%)
Number of Pages with Visualization/Interactive Features: 7,129 (32.7%)
Number of Pages with a Database: 3,576 (16.4%)
Number of FTP pages: 693 (3.2%)





Nominated Unique Seeds and Datasets by Agency

Agency	Number of Unique Seeds	Nominated Unique Datasets	Percent of Datasets within Seeds
EPA	21835	4785	21.9
DOE	10942	4781	43.7
NOAA	13671	5580	40.8
OSHA	113	45	39.8
NASA: SMD	4282	2152	50.3
NASA: SED	1489	1042	70.0
USDA	398	108	27.1
DOI	2747	1418	51.6
USGS	113	19	16.8

The EDGI volunteer tech team of over 30 contributors has built the following tools and projects:

- A Github organization to develop, maintain, and share tools for archiving government data. Github is an open source software development and version management system. All of the software EDGI develops is open source and freely available to the public on Github.
- The Nomination Tool Chrome Extension, used by archiving volunteers to nominate nearly web pages to the Internet Archive's End of Term 2016 project, as "seeds" for further web crawls.
- A technical toolkit and documentation for volunteers planning or attending archiving events.
- Three iterations of scripts used by our Website Monitoring team to improve the process of tracking changes to government websites. The current tool enables rapid comparison of versions of webpages to identify significant changes.
- Other tools and scripts for scraping/downloading information from specific agency data sources. A web application to enable rapid transfer of large files using Amazon Web Services to assist our archiving workflow.

Our archiving work has primarily occurred at public "DataRescue" events, which have also been called "archive-a-thons." We avoid using the term "hacking" to describe EDGI work because all of the websites and databases we download are publically available. This project has received a good deal of media attention. It can be difficult to understand, and so we have provided an "Archiving Glossary" below, and tried to describe the process as

best we can. If you have further questions please reach out to the Archiving of Data contacts listed on the EDGI website.

Nominating web pages to the Internet Archive is also referred to as "seeding." The Internet Archive is a vast, freely accessible digital repository. Through its Wayback Machine and its "End-of-Term" project, it maintains a web crawler that periodically captures government websites, allowing users to view web pages that have disappeared (which can be a normal part of political transitions) or compare changes across different dates. However, the Internet Archive web crawler cannot capture an entire agency website— some web pages are "uncrawlable" (see Glossary below), and must be nominated (or "seeded") by hand. Uncrawlable web pages includes important datasets, such as those containing pdf documents and or which are dynamic (rather than static) databases. These must be identified and harvested by skilled volunteers through a multi-step process. EDGI offers webinars and online resources to train people hosting DataRescue events in this process.

Archiving Glossary

Github: an open-source software development and version management system used by our archivers, where many tools they've built are stored.

Harvesting: the process by which datasets and files such as pdfs are downloaded and saved.

End-of-Term Web Archive: sponsored by the Library of Congress, the Internet Archive, and other organizations, this project capture and saves federal government websites at risk of changing during Presidential transitions. It began in 2008.

Internet Archive: a San Francisco-based non-profit digital library with the goal of "offering permanent access for researchers, historians, scholars, people with disabilities, and the general public to historical collections that exist in digital format."

Repository: digital storage for large files and datasets; usually cloud-based.

Seeding: marking websites or webpages so that a webcrawler will copy the material on that url (http://webaddress) itself or its subdomains ("below" it; i.e. http://webaddress/subaddress).

Uncrawlable url: a website or webpage containing data or other information that cannot be downloaded or stored automatically by a webcrawler; it must be archived by other means.

Webcrawler: a program that visits websites and reads their programs and information to provide indexes of them, in this case for inclusion in the Internet Archive. Webcrawlers are also known as "spiders" or "bots."

This data collected through harvesting is either put in the repository kept by a partner project, DataRefuge, and/or are converted into a format (called WARC, or WebARChive format) that can be added to the Internet Archive. To facilitate and keep track of has been rescued, EDGI built online tools, such as our Chrome Nomination Tool. These online tools also create a systematic record of archived web pages (from events across the country and in Canada), and a list of "uncrawlable" web pages and datasets that need to be downloaded and archived in other ways.

Because data archiving is a dispersed effort, taking place in cities across the country, we needed a way of systematically working through vulnerable web pages and datasets. To direct this broad effort, we initially developed reports on at risk federal agencies, which we called Primers (these are are available on our website). These Primers outlined the structure of federal environmental agencies, their programs, including vulnerable websites and data, and they served to guided our archiving and initial monitoring activities.

2. Website Monitoring

Using version-tracking software, **EDGI monitors changes to federal environmental agency websites on a daily basis**. This work involves documenting and analyzing data that disappears from public view as well shifts in rhetoric and presentation, which reflect the new administration's priorities.

For example, on January 28, 2017, we found that the Bureau of Land Management had deleted a section of its web site that outlined the Methane and Waste Prevention rule, an Obama-era policy to reduce greenhouse gas emissions. The rule had been opposed by the oil and gas industry. Our documentation of these changes, along with the removal of a section of Department of Interior's website on hydraulic fracking, was reported on in *The Intercept*. As another example, we found that on February 2nd, the U.S. State rewrote its website on climate change, removing text such as, "The United States is taking a leading role by advancing an ever-expanding suite of measures at home and abroad," along with references to climate change mitigation. Our subsequent report informed an article in Climate Central.

As of May 1, 2017, our website monitoring team includes 10 analysts that **actively track over 25,000 web pages**.

We have compiled 15 reports that document and analysis socially significant changes to federal websites, including EPA, DOE, DOI, OSHA, NIH, NASA, DOT, GAO, CDC, the State Department, and the White House. We publish these reports on our website.

Our monitoring team partners with journalist to communicate and publish our findings. Our reports (including reposts) have appeared in *The Washington Post*, *Propublica*, *The Atlantic*, *Scientific American*, *The New York Times*, *The New Republic*, *Quartz*, *The Intercept*, *The Independent*, Climate Central, *E&E News*, *Business Insider*, Mashable, *The Christian Science*

Monitor, EcoWatch, and Michigan Radio.

Please visit the Website Monitoring on the EDGI website for more information. Our in-development open-source website monitoring platform is made possible by a collaboration with the company, PageFreezer.

3. Interviewing

EDGI began confidentially interviewing long-term EPA and OSHA employees in December 2016, and since then **we have conducted 58 interviews**. Our current interview team includes 8 professors and graduate students based in Washington DC, New Jersey, Boston, Colorado, Vermont, and the San Francisco Bay Area. These interviews provide a human and more nuanced perspective on the impacts of the current administration on environmental agencies. They also provide historical context that enable us to better analyze and evaluate the present moment. The interviewing project is an example of the unique value EDGI provides as an academic research network.

All interviews are confidential, and unless the interviewee specifies otherwise, the transcript is de-identified. Our research team has developed rigorous procedures for securely storing the audio files and interview transcripts. EDGI researchers also share files and communicate with an end-to-end encryption program to maintain confidentiality.

These interviews inform our upcoming "100 Days Report" on the Trump administration effects on environmental agencies and policy. For example many long-term EPA employees drew comparisons with the early Reagan administration, as in the following quote:

> "I was in…a small staff office, answering to the assistant administrator for enforcement. Handled all the training, handled a lot of the penalty kinds of issues, a lot of the broad policy stuff the Enforcement Office did. When Anne came in, Anne Gorsuch, one of the first things she did was to demolish the Enforcement Office…. What they did was they stripped all the attorneys out and moved them to the various program offices, and then basically disaggregated the Office of Enforcement."

The Trump administration also threatens to dismantle parts of the EPA, in ways our interviewees have not experienced, at least for the past 35 years. Yet the Reagan years also provide an instructive case for how EPA employees and others resisted attempts to undermine the agency. For example, former EPA Administrator William Drayton established network of environmental professionals called "Save EPA," which monitored cutbacks to the agency and wrote reports that were sent to Congress. These and other insights are explored further in the "100 Days Report," and will be analyzed and written about for future publications.

The interview project is ongoing. We invite people who have spent substantial time at a

federal environmental agency to contact us for an interview. You can contact us using this this Google Form, by phone at (917) 887-4244, or if you have a fully encrypted email account you can copy the text of the survey and fill it out in the body of an email, sending it to: EnviroDGI@protonmail.com.

4. Capacity and Governance

Our Capacity and Governance working group monitors changes to federal agency budgets, enforcement, and research capacities, and to environmental policy and regulation more broadly. Our team meets weekly to coordinate research projects and public responses.

Our accomplishments include:

- A **white paper**, "Public Protections Under Threat at the EPA: Examining Safeguards and Programs That Would Have Been Blocked by H.R. 1430." Based on our research, we conclude that H.R. 1430 (the "HONEST Act"), which was recently passed in the U.S. House of Representatives, would inhibit the EPA's use of important scientific data and the agency's mission of protecting human and environmental health. We also sent a letter to Congress, which was introduced on the House floor as public record.
- Gave **oral public comments** at an EPA hearing on H. J. Res 59, which seeks to "stay" or prevent the implementation of the Clean Air Act's Risk Management Prevention rule, or RMP. The RMP rule was created by the EPA in response to an Executive Order by Obama to improve chemical and facility safety, following numerous dangerous and fatal toxic releases at industrial facilities. After learning about this H.R. Res 59, the Capacity and Governance team quickly mobilized to research the benefits of the RMP rule and write a response, working collaboratively online from different locations across the country.

EDGI has also developed several **Freedom of Information Act (FOIA) projects** to (1) determine the reasons for observed changes to information access/content changes on public-facing government websites occur, (2) gather information to about changes to environmental agencies and their work, (3) collect information regarding agencies' data access and maintenance practices, and (4) collaborate with national environmental FOIA and government transparency efforts. To these ends, EDGI has:

- Filed 19 FOIA requests since February, 2017 to gather records related to website changes, budget and agency personnel, and various agency datasets.
- Formed a partnership with MuckRock, a major FOIA organization and request platform to create a FOIA project for environmental organizations to collect, file, and collaborate on FOIA efforts.
- Worked with environmental advocacy organizations to create a platform and open discussions about environmental information transparency and FOIA efforts, and to

ensure continued environmental information access.
- EDGI will serve as a hub organization for FOIA requests and responsive records.

Lastly, **EDGI periodically mobilizes to quickly respond to important events**, such as Scott Pruitt's first speech as EPA Administrator. EDGI historians and social scientists provided a "rapid analysis" of Pruitt's speech (in this case, an annotated version of the speech), which was also reported on and reprinted in Newsweek. The analysis of Pruitt's speech, published days after the event, is an example of how EDGI seeks to inform current political debates from a scholarly, research-based perspective.