EDGI

Environmental Data and Governance Initiative

Introducing the Environmental Data and Governance Initiative

The *Environmental Data and Governance Initiative* (EDGI) is an international network of academics and non-profits that has arisen to address potential threats to federal environmental and energy policy, and to the scientific research infrastructure built to investigate, inform, and enforce it. Dismantling this infrastructure -- which ranges from databases to satellites to models for climate, air, and water -- could imperil the public's right to know, the US' standing as a scientific leader, corporate accountability, and environmental protection. In response, EDGI is building online tools, research networks, and hosting public events to proactively archive public environmental data and ensure its continued public availability. We also are monitoring changes to federal regulation, enforcement, research, funding, websites and general agency management at agencies including EPA, DoE, NASA, NOAA, and OSHA.

EDGI began on Nov. 11th, 2016 as an email sent to 14 researchers, and has grown rapidly to a network of some 65 people from 21 Institutions. Our 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 environmental data and analysis. EDGI also collaborates with a growing list of organizations including: Public Lab, 314Action, DataRefuge, The Internet Archive, and Climate Central.

Highlights of our work over the last month and a half:

- dramatically improved the saving of federal environmental websites, placing them in the Internet Archive's "End-of-Term" project. Thanks to our efforts, nominators of links needing to be saved at the end of the last Obama administration rose to ten times those in 2012, at least 60% of them working in EDGI-sponsored public events and efforts. Nearly seven times as many important links were designated for automatic saving; and 1591 datasets for non-automated downloading.

- set up real-time monitoring of government environmental websites. We started with approximately 25,000 pages at agencies including EPA, DoE, DoI, NASA, NOAA, and OSHA, as well as whitehouse.gov and data.gov. We are expanding to many more for a total of tens of millions of pages, and have already produced four reports that have informed news articles.

- launched a project of confidentially interviewing those recent agency retirees, to illuminate the human sides of this versus earlier transitions. Trained interviewing teams in Washington, Boston, the San Francisco area, and elsewhere, are currently conducting secure interviews.

We aim 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 data are vital for environmental governance. For more information, please see the pages below.

We invite your assistance in all of our efforts, especially:

- your suggestions about data and websites that need saving
- your time and skills:
  - helping out with data rescue events
  - planning your own data rescue event
  - joining EDGI, to lend a hand with real-time agency monitoring and analysis
- your memories, if you are a recently retired federal environmental agency employee
- your financial support in our crowdfunding effort at bit.ly/EDGI314

To offer suggestions or volunteer in any of these ways (including for an interview), please fill out this Google Form. Those wishing to share their experiences at environmental agencies can also contact us via phone, at (917) 887-4244. If you have a fully encrypted email account (meaning that only the people communicating can read the messages; see *note below for how to set up), you can copy the full text of the survey, fill out, and send in the body of an email to: EnviroDGI@protonmail.com

For more information on us, please visit our website: https://envirodatagov.org/

Contact us at a secure email: EnviroDGI@protonmail.com or on Twitter @EnviroDGI

*Note: end-to-end encryption is only available *between * encrypted accounts. If you do not have such an account register for a ProtonMail account (it is free) and also note that some metadata cannot be encrypted, so be please be careful with subjects and attachment names.

Many thanks!

Steering Committee, Environmental Data and Governance Initiative

Table of Contents of EDGI's Work and Plans:

1. **Archiving:** With the help of the Internet Archive and Data Refuge, we have built a campaign to preserve publicly accessible government data on the environment that may in the future be deleted or rendered inaccessible. To date, we have completed the rescue of thousands of websites, webpages and datasets, across seven environmental agencies, aided by archiving tools crafted by EDGI's own developers. To carry this endeavor forward, plans are consolidating for many future archiving events and related work.

2. **Website Monitoring:** We are expecting many changes to the websites of federal environmental agencies as the new administration moves to take charge of them. An EDGI team is now in place to follow these, using special software as well as a careful vetting process to confirm what changes we find. Our website monitoring team will issue regular reports on the significant alterations they discover.

3. **Interviewing Project:** Another EDGI team has launched a project to confidentially interview long-term employees at the EPA, OSHA, and other environmental agencies, especially those that have recently retired. In the first round of interviews, we have been interested in the ways memories of earlier presidential transitions may shape expectations and worries about the present one. As our interviews with departing employees continue through the first months of the Trump administration, we also hope to learn more about how this transition is looking and proceeding from the inside.

4. **Analysis and Outreach:** As the transition proceeds, EDGI plans to follow impacts on agency operations, capacities, and policies, and in combination with findings from website monitoring and interviews, disseminate searching and synthetic analyses of significant transformations. Among the public outreach activities in the works are roundtables, biweekly reports, and an assessment of the new administration's first hundred days.

1. Archiving

Archiving Completed

Our campaign preserves government environmental data now publicly available on the web, but that may in the future be deleted or rendered inaccessible. Our archiving work primarily happens at public events called archive-a-thons, which have received a good deal of media attention. Where possible we "seed" (digitally designate) websites for copying by the Internet Archive, a vast, freely accessible digital repository. However, much data online, such as pdf documents and databases, has to be identified and harvested by people using several more steps: they are then either put in the repository kept by a partner group, Data Refuge, and/or are converted into a format (called WARC, or WebARChive format) that can be added to the Internet Archive. To keep track of what's being rescued, we use tools we built that allow large groups of people to collect seeds for the Internet Archive, while also creating a systematic record of what was archived, and a spreadsheet of uncrawlable material and datasets that needs to be acquired with other tools. Between archiving events, EDGI members seed additional data we determine to be

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 many organizations including the Library of Congress as well as the Internet Archive, this project captures and saves federal government websites at risk of changing during Presidential transitions; has done so since the transition of 2008.

Internet Archive: a San Francisco-based non-profit digital library with 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 pages so that a webcrawler will be sure to copy material on that url (http://webaddress) itself or its subdomains ("below" it; i.e., http://webaddress/subaddress).

Uncrawlable url: a website/page containing data or other information that cannot be downloaded or stored automatically by a webcrawler; it must be saved 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; also known as a "spider" or "bot."

vulnerable through on-going research into transition plans and work being done at the agencies themselves.

Seeding Progress (as of 5 p.m., January 25, 2017)

Seeds are websites/pages sent to the webcrawler to copy material into the Internet Archive; pages and information then saved by the webcrawler amount to far more than just these urls, since those urls "below" any seeded are also crawled and saved.

Seeds Submitted for Crawling at Events *				**# Seeded****	**Uncrawlable Urls Noted**
Toronto Event				**3169**	
Subsequent Events (Philadelphia, Indianapolis, Los Angeles, Ann Arbor, others)		**Agency-Specific Breakdowns Below**		**25575**	**Total for these events: 5247**
		NOAA		**4262**	**1758**
		DOE		**646**	**159**
		EPA		**466**	**62**
		NASA and NOAA		**19000**	
		Remainder: NASA, DOI, USDA, OSHA		**unknown**	
Seeds Submitted for Crawling Independent of Events					
EDGI End-of-Term Submission 11/19				**4122**	
EDGI End-of-Term Submission 1/18/17				**24238**	

		Total Seeded	**54478**	

*** These seeds are produced through crowd-sourcing methods.**

****These numbers are not filtered for duplicates.**

Seeding Stats from the Internet Archive reflecting our effort to support the End-of-Term Project

As of January 25, 2017, here is a comparison to previous End of Term crawls by numbers of seeds added by nominators.

- End-of-Term 2008: 457 seeds from 26 nominators
- End-of-Term 2012: 1476 seeds from 31 nominators
- End-of-Term 2016: over 10,200 seeds from over 300 nominators (so far, at least 60% of these nominators were from EDGI-sponsored events and efforts)

The Internet Archives' counts and visualization of End-of-Term progress can be viewed online.

As of January 25th, the End of Term crawl was estimated at 110 million pages, a 10X expansion of the size.

Uncrawlable Datasets Progress (as of 5 p.m., January 25, 2017)

These are datasets that cannot be copied into the internet Archive and need to be harvested by tools used by our coding community. Harvested datasets are then sent to the DataRefuge repository, which is cloud computing storage organized at the University of Pennsylvania.

Uncrawlable Datasets Identified = **1,591** (From the Master Uncrawlables list, January 25, 2017)

Uncrawlable Datasets Harvested; An Overview (January 25, 2017)**:**

- **By Size:**
  - Toronto: 84 GB
  - Michigan 1TB
  - DataRefuge Repository: 21GB
- **By Number**
  - Network Overall: 96
  - In the DataRefuge Repository 18

Events Sponsored for Archiving (Data Rescue) and Coding

- **Past Archiving (Data Rescue) Events**
    - December 17, Toronto: 150 people
    - January 13-14th, Philadelphia: 81 people
    - Jan. 19th, Indianapolis
    - Jan. 20th, Los Angeles 60 people
    - Jan. 26-27, Ann Arbor, Michigan, 200 people
- **Past Coding Events**
    - 7 coding nights in Toronto with affiliate Civic Tech TO
    - 2 coding nights in Chicago with the Code For America affiliate ChiHack

Agencies/Offices Whose Websites and Data We've Been Seeding and Harvesting

1. Through Archiving (Data Rescue) Events So Far

- **DOE: Department of Energy**
    - Office of Energy Efficiency & Renewable Energy
    - DOE - National Labs:
        - Argonne National Laboratory
        - Ames Laboratory
        - Brookhaven National Laboratory (BNL)
        - Fermi National Accelerator Laboratory (FNAL)
        - National Renewable Energy Laboratory (NREL)
    - Office of Science
    - Energy Information Administration
- **NOAA: National Oceanic and Atmospheric Administration**
    - National Environmental Satellite, Data, and Information Service
    - National Marine Fisheries Service
    - Office of Oceanic & Atmospheric Research
    - National Center for Environmental Information
    - National Weather Service
- **NASA: National Aeronautics and Space Administration**
    - Science Mission Directorate
    - Sciences and Exploration Directorate
- **EPA: Environmental Protection Agency**
    - Note of accomplishment: According to End-of-Term project statistics, prior to the January 20, 2017 Inauguration, the EPA had become the most thoroughly archived agency at the Internet Archive thanks in part to our efforts (based on searches by agency name using IA's search engine).

2. Through non-event EDGI Seeding

- **EPA**, **DOE**, **NOAA**, **NASA, OSHA, USDA, DOI** (so far)

What Else We've Done to Support Archiving

Agency Primers

To support events, our Capacity, Finance and Budget team has created a set of what we call "Agency Primers" which are available on our website. These primers outline the structure of agencies, their programs, websites and data, and serve as guides for our archiving and monitoring activities. The scope of many agencies is incredibly vast, and certain offices or programs may not be as at-risk as others. For example, the Department of Energy has over 30 offices, over 20 labs, and over 20 other associated centers and departments, covering topics including nuclear energy, renewable power, and regulatory enforcement. The primers play an important role in systematizing archiving efforts to both guide seeding efforts and ensure coverage of at-risk programs.

Coding and Tool Building Work

The EDGI volunteer tech team of over 25 coders has built or contributed to 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.
- **EDGI Github Stats**
  28 people, 13 projects (called repositories)

  103 commits to Workflows Repository

  25 commits to Overview

Tools and Scripts Built

- The Nomination Tool Chrome Extension has been used by over 380 volunteers to nominate nearly 29,000 web pages as seeds for further crawls in the Internet Archive's End of Term 2016 project. The extension has greatly enhanced the ability of volunteers to contribute to the project and continues to be used in our data harvesting efforts.
- Various incarnations of a technical toolkit and documentation for volunteers planning or attending archiving events. These tools are used to both capture the seeds to go to the Internet Archive, and to download datasets.

- A JavaScript tool used by our monitoring team to dramatically speed the process of tracking changes to government websites. This tool enables rapid comparison of versions of webpages to identify significant changes.
- Seven tools for scraping/downloading information from specific EPA data sources, one of which, the EIS WARC Archiver, has evolved into a general-purpose model for downloading and packaging difficult-to-access web resources.
- A web application to enable rapid transfer of large files using Amazon Web Services to assist our archiving workflow.

Copying Data.gov

Finally, we are creating a mirror copy of the central federal repository of open-access data, data.gov, (~30Tb). This mirror is also the biggest test yet of IPFS, the next-generation distributed-web hypermedia protocol. It is an important experiment towards developing a future internet protocol to replace http that would better keep datasets and information open access over the long term. This project is a collaboration between EDGI, the Technoscience Research Unit at the University of Toronto, Stanford University, and ipfs.io.

Archiving Work Remaining:

Planning through Primers

This list of primers reflects EDGI's current determination of the 33 highest priority at-risk offices and labs within 7 government agencies. It also maps our priorities for next steps in archiving work.

Primers and seeding completed and planned:

- DOE: 8 Primers (2 fully seeded; 3 partially seeded; 3 not seeded)
- NOAA: 5 Primers and 2 Draft Primers (5 fully seeded; 2 not seeded)
- NASA: 2 Primers (2 fully seeded)
- EPA: 3 Primers and 8 Draft Primers (3 fully; 8 not seeded)
- DOI: 2 Draft Primers (2 not seeded)
- USDA: 2 Draft Primers (2 not seeded)
- OSHA: 1 Draft Primer (1 not seeded)

Total: 18 Final Primers with 15 Draft Primers (12 fully seeded; 3 partially seeded; 18 not seeded)

While agency offices differ drastically in size and scope, making it very difficult to compare across them, we have completed an initial review of the scope of 18 of these 33 high priority offices, represented by completion of those primers, which leaves 15 remaining offices. In addition, we have finished a thorough

examination of the websites of 12 of those 18 offices, archiving individual web pages and identifying important data and document repositories within those sites. While there are, of course, many other important agency offices to examine and archive, this list gives a good sense of the high priority work that remains.

Archiving Events Ahead

Confirmed DataRescue Events Upcoming:

- Boston: Feb 1st
- UC Davis: Feb 3rd
- NYC: Feb 4th
- SF Bay: Feb 11th

There are 31 proposed upcoming DataRescue events.

2. Website Monitoring

As the Trump administration transforms agencies to align with its campaign promises and goals, we expect many changes to agency websites. We have devoted considerable effort over the last few weeks to setting up the means for monitoring these changes. The EDGI website tracking team selected a substantial baseline of vulnerable webpages in the last weeks of the Obama administration, and is now following changes in these pages as Trump "beachhead teams" arrive and transform the digital face of agencies. To publicize these changes, we are working with journalists and are planning biweekly bulletins that will summarize what we've found.

At present, we are monitoring approximately 25,000 pages at agencies including EPA, DoE, DoI, NASA, NOAA, and OSHA, as well as cross-agency domains such as whitehouse.gov and data.gov. Thanks to donations of server space, crawling time, and analytical software use from PageFreezer.com, we have also begun tracking the entirety of more than 150 .gov sub-domains, for a total of tens of millions of pages. We are working to build an open source portal to enable public access to this data, which we hope to make available in late 2017 or early 2018.

3. Interviewing

Another wing of EDGI is confidentially interviewing retired longtime employees, staff, officials and political appointees at EPA, OSHA, and other environmental agencies. These interviews will help us gain a

human perspective into the impacts of the current transition on federal environmental work. They also will enable us to explore earlier transitions between presidential administrations, and ways these may be compared to the current one. Prior to the inauguration, we conducted fifteen interviews with former EPA and OSHA employees that will serve as a baseline for many more to come, especially if shrinkage of agency staff proceeds as this administration has apparently planned. Interviewers probe what respondents remember about how this and earlier transitions affected and perhaps transformed the agency's sense of mission, rationales, capabilities, public interface, and politics of evidence—the status and uses made of different kinds of science and scientists.

We are now planning on conducting interviews on a continuing basis throughout the months and perhaps years of transition into the Trump administration. With sufficient support we will gain perspectives not just on the many parts of EPA and OSHA but the DOE, DOI, NASA and other agencies with environment missions, perhaps on a scale of a hundred interviews or more. Already, we have trained interviewers based in multiple sites around the country, from Boston to Washington, D.C., to Colorado to the San Francisco area; we also conduct many interviews by phone. We have developed strict procedures for keeping the audio and transcripts of the interviews, as well as names of interviewees, safe and secure, including several different tiers of confidentiality that interviewees may choose.

In the coming weeks, we will begin synthesizing findings from the interviews to be combined in public reports, in ways that also respect interviewee choices about confidentiality. We also offer interviewees the option to make their own public statements, and are currently exploring ways of publicizing these. We are also in negotiation with well-known oral history collections at Columbia University and University of Pennsylvania Libraries over an ultimate home for the transcripts, which we anticipate will become an important historical archive.

If you have spent substantial time at an environmental agency, we invite you to nominate yourself for an interview, via this Google Form. You can also contact us via phone at (917) 887-4244. If you have a fully encrypted email account (meaning that only the people communicating can read the messages; see *note below for how to set up), you can copy the full text of the survey and fill it out in the body of an email, sending it to: EnviroDGI@protonmail.com

4. Analysis and Outreach

A working group in EDGI has also begun monitoring changes to federal environmental programs, funding, policy and regulation, with a view to reporting on what we find. We began by developing a baseline account of what the scientific and regulatory missions and capacities of these agencies looked like in the Obama administration. Against that baseline we are now monitoring the remaking of agencies and their mandates through budget, staffing and policy changes. We plan to combine this information with findings from website monitoring and interviews and make our findings available through digests and reports, including one on the Administration's first 100 days. EDGI's unique combination of archiving, networking and monitoring puts us in an exciting position to develop concrete positive visions for alternative approaches to environmental governance that combine data with human experience. Our collection and

analysis of data will be conducted and disseminated with a view to exploring and recommending possibilities for improving environmental governance.

Notes

Participating Institutions: Harvard University, Indiana University, Johns Hopkins University, New York University, Northeastern University, Rice University, Stanford, SUNY Stony Brook, UC Davis, UC Santa Cruz, University of Pennsylvania, University of Toronto, and Yale University. https://envirodatagov.org/about/

News articles informed by our website monitoring in the past week include articles in Business Insider and The Intercept.

Environmental Data and Governance Initiative, February 1, 2017.