

Developing a Chain of Custody For Government Data

Andrew Battista, Librarian for Geospatial
Information Systems, NYU

Stephen Balogh, Data Services Specialist, NYU

Library of Congress Designing Storage
Architectures Meeting

September 18-19, 2017



Johannes Moreelse, Héraclite (ca 1630)



How can academic institutions play a role in preserving government data?

What does it mean to “claim” a federal agency?

How can one institution develop a “chain of custody” for an agency’s collection of data?



Government data is at risk

- Political
 - Threats to climate change + socioeconomic data (e.g., HR 82)
- Funding
 - Servers going dark
- Obscurity
 - Inability to locate particular datasets
- Format and compatibility issues
 - Obsolete and proprietary formats



Main challenges

- Data is always in flux
- ... so is metadata
- Lack of adequate identifiers for many datasets
- Not knowing what data exists in the first place
- Coordinating preservation responsibilities across institutions / “claiming” a department



“You could not step twice into the same river...”
alt. “All is flux, nothing is stationary.”



Heraclitus (535 - 475 BCE), on the challenges of data modeling
[Plato, *Cratylus*, 402a]

Possible strategies

Mirror **everything**

Mirrors **are** extremely useful, and essential infrastructure



Possible strategies

Focus on catalogs, and mirror **everything**

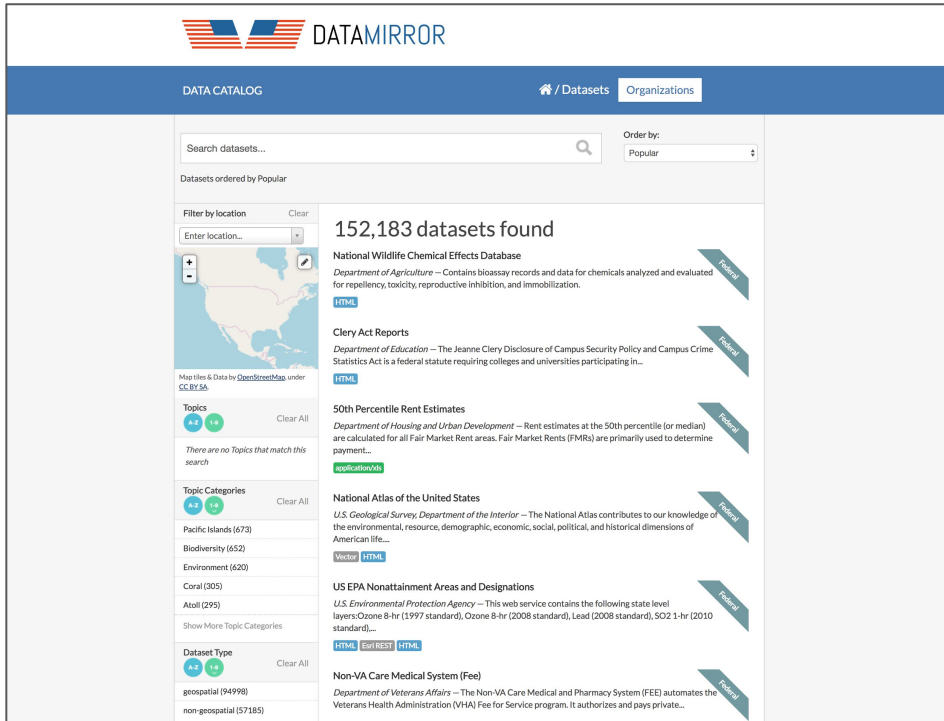
But what is “everything”?

Data.gov + other known catalogs

1. these are incomplete
2. ... they're catalogs, not repositories
3. mismatches between metadata and data
4. records are ephemeral



Excellent mirroring work *is* being done



The screenshot shows the DATAMIRROR website interface. At the top left is the logo with the text "DATAMIRROR". Below it is a navigation bar with "DATA CATALOG" and "Organizations". A search bar is present with the text "Search datasets...". Below the search bar, it says "Datasets ordered by Popular". On the left side, there are filters for "Filter by location" (with a map of the United States) and "Topics" (with a list of topics like Pacific Islands, Biodiversity, etc.). The main content area displays "152,183 datasets found" and lists several datasets with brief descriptions and format icons (HTML, PDF, etc.):

- National Wildlife Chemical Effects Database**
Department of Agriculture — Contains bioassay records and data for chemicals analyzed and evaluated for repellency, toxicity, reproductive inhibition, and immobilization.
- Clery Act Reports**
Department of Education — The Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act is a federal statute requiring colleges and universities participating in...
- 50th Percentile Rent Estimates**
Department of Housing and Urban Development — Rent estimates at the 50th percentile (or median) are calculated for all Fair Market Rent areas. Fair Market Rents (FMRs) are primarily used to determine payment...
- National Atlas of the United States**
U.S. Geological Survey, Department of the Interior — The National Atlas contributes to our knowledge of the environmental, resource, demographic, economic, social, political, and historical dimensions of American life...
- US EPA Nonattainment Areas and Designations**
U.S. Environmental Protection Agency — This web service contains the following state level layers: Ozone 8-hr (1997 standard), Ozone 8-hr (2008 standard), Lead (2008 standard), SO2 1-hr (2010 standard)...
- Non-VA Care Medical System (Fee)**
Department of Veterans Affairs — The Non-VA Care Medical and Pharmacy System (FEE) automates the Veterans Health Administration (VHA) Fee for Service program. It authorizes and pays private...



The screenshot shows the Climate Mirror website interface. At the top left is the logo with the text "Climate Mirror". Below it is a navigation bar with "HOME ABOUT MIRRORS TOOLS JOIN AND DONATE SUBMIT DATA CONTACT PARTNERS". The main content area features a large image of a sunset over a cloudy sky with the text "Climate Mirror" and "an open project to mirror public climate datasets". Below this is an "About Us" section with the following text:

Climate Mirror is a distributed effort conducted by volunteers, in conjunction with efforts from institutions such as University of Pennsylvania, University of Toronto, and the Internet Archive, to mirror and back up U.S. Federal Climate Data. It started pre-emptively out of concerns based on President Trump's past anti-science statements, and has continued into his administration's time in office.



Libraries Add Value

- Make items available in contexts meaningful to academic institutions
- Pull things apart / push things together
- Fully represent the preservation context

The screenshot displays the NYU Spatial Data Repository interface. At the top, the NYU logo and 'Spatial Data Repository' are visible, along with 'Submit', 'History', and 'Login' links. The main content area shows the title '2012 New York City Real Estate Sales' with a location pin icon. Below the title, the 'Author(s)' are listed as 'GIS Lab, Newman Library, Baruch CUNY'. The 'Description' provides a detailed account of the data's origin, mentioning 'Detailed Annual Sales Reports by Borough' and the use of 'NYC Geoclient API'. It notes that the data is a point layer representing property sales locations in New York City. The 'Publisher' is 'Newman Library (Bernard M. Baruch College)', and the 'Collection' is 'NYC Geocoded Real Estate Sales'. The 'Place(s)' field lists various boroughs and counties in New York City and the United States. The 'Subject(s)' are 'Real property, Property, and Real estate business'. The 'Format(s)' is 'Shapefile', the 'Year' is '2012', and the 'Held by' is 'Baruch CUNY'. A 'Preservation record' is provided with the URL 'http://hdl.handle.net/2451/34675'. On the right side, there are 'Tools' including 'Email', 'Web services', 'Open in Carto', 'Documentation', and a 'Download Shapefile' button. Below the tools is a 'Data Relations' section with a 'Source Datasets' link for '2016 NYC Geocoded Real Estate Sales Geodatabase, Open Source Version'. At the bottom, there is a map showing the real estate sales data as red points over a map of New York City, with an 'Attribute' table and a 'Click on map to inspect values' instruction.

Individual institutions have different standards or make different choices regarding how they collect data



Test Case: Preserving *U.S. Forest Service* Data

- Represented in multiple places
 - Data.gov
 - FSGeodata Clearinghouse
- Frequently updated
- Released in a variety of formats
 - Geodatabases
 - Shapefiles
 - Web services
- Atomized in different ways



Multiple Contexts, Same Data?

USDA United States Department of Agriculture Forest Service

FSGeodata Clearinghouse

Clearinghouse Home Help Contact Us

Enterprise Data

- Downloadable Data
- Data Extract Tool
- Map Services

Maps

- FSTopo
- Standard Map Products
- Other Map Products

Raster Data

- Caribbean Island Land Cover
- Chugach Updates
- National Forest Type
- Forest Biomass
- RAVG
- Special Purpose FS Data

FSGeodata Clearinghouse

The USDA Forest Service Geodata Clearinghouse is an online collection of digital data related to forest resources. Through the Clearinghouse you can find datasets related to forests and grasslands, including boundaries and ownership, natural resources, roads and trails, as well as datasets related to State and private forested areas, including insect and disease threat and surface water importance. You can also find downloadable map products, raster data, and links to other sources of forest resource information.

What's new?

- FSTopo data** is now linked from the Maps portion of the left menu. The new FS Topo page contains links to a page about raster map images, FS Topo data available on the downloadable data page, and links to the elevation contours data for the conterminous USA from the **USGS National Map Viewer**.
- New National Monument boundary data** – Recently, new national monuments have been designated on Forest Service lands. Boundaries of the national monuments can be found on the Downloadable data page in the dataset titled: "National Forest Lands with Nationally Designated Management or Use Limitations". Information on how to obtain that data is found in this Frequently Asked question.

FSGeodata Clearinghouse

Search Data.Gov

DATA.GOV

DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

DATA CATALOG / Datasets Organizations

forest service

Order by: Relevance

Datasets ordered by Relevance

Publishers: US Forest Service, Department of Agriculture

You are searching in the list of datasets. Show results in entire Data.gov site.

94 datasets found for "forest service"

U.S. Forest Service Surface Drinking Water Importance - Forests on the Edge

Department of Agriculture – A map service on the www depicting data that supports the publication RMRS-GTR-327, Private forests, housing growth, and America's water supply: A report from the...

ArcGIS Feature Service API XML ZIP

FS National Forest Dataset (US Forest Service Proclaimed Forests)

Department of Agriculture – A map service on the www depicting the boundaries encompassing the National Forest System (NFS) lands within the original proclaimed National Forests, along with...

ArcGIS Map Service ArcGIS Map Preview HTML API XML ZIP 2 more in dataset

Data.gov



Feature Classes		Abstract
<p>Activity Silviculture Timber Stand Improvement</p> <p>ESRI geodatabase (171MB) shape file (308MB)</p> <p>Date of last refresh: Aug 25, 2017</p>	<p>metadata map service</p>	<p>The SilvTSI (Silviculture Timber Stand Improvement) feature class represents activities associated with the following performance measure: Forest Vegetation Improved (Release, Weeding, and Cleaning, Precommercial Thinning, Pruning and Fertilization). The Activities data set portrays the areas where activities are accomplished... [see more]</p> <p style="text-align: right;">parent dataset: Activities</p>
<p>Activity SilvicultureReforestation</p> <p>ESRI geodatabase (239MB) shape file (420MB)</p> <p>Date of last refresh: Aug 25, 2017</p>	<p>metadata map service</p>	<p>The SilvReforestation feature class represents activities associated with the following performance measure: Forest Vegetation Establishment (Planting, Seeding, Site Preparation for Natural Regeneration and Certification of Natural Regeneration without Site Preparation). The Activities data set portrays the areas where... [see more]</p> <p style="text-align: right;">parent dataset: Activities</p>

Item on FSGeodata Clearinghouse

DATA TOPICS - IMPACT APPLICATIONS DEVELOPERS CONTACT

DATA CATALOG

[/ Datasets](#)
[Organizations](#)
?

[/ Department of Agriculture / US Forest Service, Department ...](#)

[Submit Data Story](#)
[Report Data Issue](#)

Forest Service

Publisher

US Forest Service, Department of Agriculture

Contact

Dave Green

Share on Social Sites

[Google+](#)
[Twitter](#)
[Facebook](#)

U.S. Forest Service Silviculture Reforestation

Metadata Updated: May 30, 2017

A map service on the www depicting the locations of activities within the Silviculture Reforestation Program. This map service portrays the area where activities accomplished as a part of the silviculture program of work, funded through the budget allocation process and reported through the Forest Service Activity Tracking System (FACTS) database and are part of the Performance Measures used to rate Agency performance in meeting the Department's Strategic Goals. It is important to note that this map service may not contain all accomplished activities; the spatial portion of the activity description is not currently enforced by FACTS and at this time some are optionally reported by Forest Service units. This map service only represents those activities associated with the performance measure Forest Vegetation Improved (Release, Weeding, and Cleaning, Precommercial Thinning, Pruning and Fertilization). As spatial data reporting is enforced by the application and acceptance of reporting increases for both tabular and spatial we hope to improve the quality and comprehensiveness of the data used for this layer in coming years.

Access & Use Information

[Public:](#) This dataset is intended for public access and use.
[License:](#) Creative Commons CCZero

Downloads & Resources

- API**
[Link to API](#)

[Visit page](#)
- ESRI geodatabase XML**
 S_USA.Activity_SilvReforestation.gdb.zip

[Download](#)
- Shapefile**
 S_USA.Activity_SilvReforestation.zip

[Download](#)
- Metadata**
 S_USA.Activity_SilvReforestation.xml

[Download](#)
- KML**
 Link to generate KML

[Visit page](#)

Same item on Data.gov



This strategy would involve:

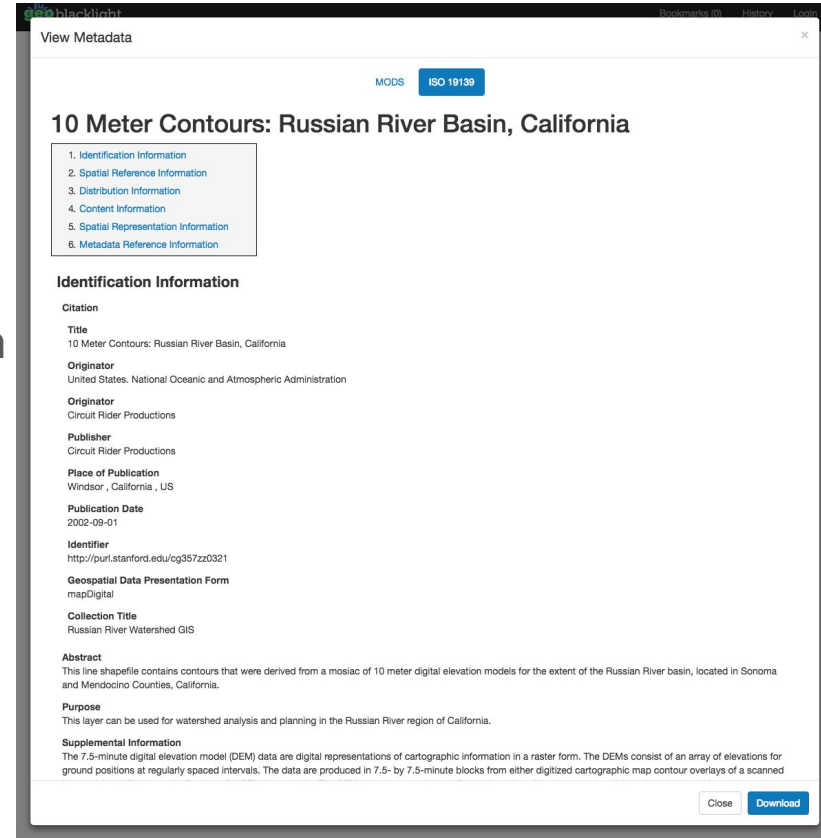
Creating **new** intellectual entities, but preserving the original context as accurately as possible

- a. Institutions preserve in accordance with their broader collection mandates
- b. Institutions capitalize on existing infrastructure
 - i. many academic libraries already have *some* approach to bit-level digital preservation



Alterations

- Augmentation of descriptions
- Cleaned and added subject terms and metadata fields for discovery
- Standardized syntax, and level at which we distribute the datasets
- Highlight and make original metadata more accessible
- Link back to government context
 - Data.gov ID (if one exists)
 - ID from any other catalog



The screenshot shows a web application window titled "View Metadata" for a dataset named "10 Meter Contours: Russian River Basin, California". The interface includes a search bar with "MODS" and "ISO 19139" selected. A table of contents lists six sections: Identification Information, Spatial Reference Information, Distribution Information, Content Information, Spatial Representation Information, and Metadata Reference Information. The "Identification Information" section is expanded, showing fields such as Title, Originator, Publisher, Place of Publication, Publication Date, Identifier, Geospatial Data Presentation Form, and Collection Title. An abstract and purpose statement are also visible, along with a supplemental information section at the bottom.

View Metadata

MODS ISO 19139

10 Meter Contours: Russian River Basin, California

1. Identification Information
2. Spatial Reference Information
3. Distribution Information
4. Content Information
5. Spatial Representation Information
6. Metadata Reference Information

Identification Information

Citation

Title
10 Meter Contours: Russian River Basin, California

Originator
United States. National Oceanic and Atmospheric Administration

Originator
Circuit Rider Productions

Publisher
Circuit Rider Productions

Place of Publication
Windsor, California, US

Publication Date
2002-09-01

Identifier
<http://purl.stanford.edu/cg357zz0321>

Geospatial Data Presentation Form
mapDigital

Collection Title
Russian River Watershed GIS

Abstract
This line shapefile contains contours that were derived from a mosaic of 10 meter digital elevation models for the extent of the Russian River basin, located in Sonoma and Mendocino Counties, California.

Purpose
This layer can be used for watershed analysis and planning in the Russian River region of California.

Supplemental Information
The 7.5-minute digital elevation model (DEM) data are digital representations of cartographic information in a raster form. The DEMs consist of an array of elevations for ground positions at regularly spaced intervals. The data are produced in 7.5- by 7.5-minute blocks from either digitized cartographic map contour overlays of a scanned

Close Download



But there are still problems...

- Many of the preservation challenges mentioned earlier remain:
 - How precisely can we tie these new entities back to their origins?
 - How can we actually **preserve** things if we're also adding data, and shifting contexts?
 - And what about the data and metadata changing constantly in the first place?



Checksum Sharing Proposal

At a *minimum*, we need a better way to collect and share claims made about **data**

- We can't trust the metadata we have access to
- We can't access all of the data we are told should exist
- Data appears in so many different contexts, that absent some central authority we can never know what has already been preserved versus what has never been captured



Checksum Sharing Proposal

At a *minimum*, we need a better way to collect and share claims made about **data**

- The only reference we can trust absolutely is the data itself
 - a particular dataset can be uniquely identified by its checksum
- Everything else needs to be contextualized as a *claim* about a piece of data with a checksum



Example

- NYU attempts to preserve a record from the USDA, originally found on Data.gov
- NYU downloads everything possible, and records the exact state of files and metadata as received
 - (*Bagit* specification useful here)
- There may not be a one-to-one correlation between structure of files on Data.gov and the NYU repository



Example (continued)

- ... so we share a claim that lets others know, that for *each individual file* taken from the USDA:
 - **NYU** has seen this file, which has checksum:
[sha256:8d93fd1be34...f343](#)
 - It came originally from: [data.gov:12345678](#)
 - NYU downloaded it on [2017-09-10T13:11:01](#)
 - And it now appears in:
 - [nyu.edu/98765432](#)
 - [handle.net/98765432](#)



```
{
  "id:sha256": "b4e5b7f8eda8b6f16ee94648da94fd003e768783ae0d20fa8b702450d2491406",
  "institution:name": "New York University",
  "institution:id": 3928,
  "claim:content-length-bytes": 508258784,
  "claim:institution-package-filename": "5_S_USA.RoadCore_FS/S_USA.RoadCore_FS.shp",
  "claim:harvest-data-gov-id": "b2c86e0e-826a-4d25-896b-cd8deb3b7f13",
  "claim:is-part-of": [
    "data.gov:b2c86e0e-826a-4d25-896b-cd8deb3b7f13",
    "https://data.fs.usda.gov/geodata/edw/edw\_resources/shp/S\_USA.RoadCore\_FS.zip"
  ],
  "claim:institution-retains-data": true,
  "claim:institution-data-part-of": [
    "http://hdl.handle.net/2451/36738"
  ],
  "claim:date-described": "2017-09-06T16:44:08-04:00"
}
```



What this infrastructure could enable

- A searchable graph of claims that relate files with any of the contexts in which they were originally discovered
- Ability to verify if a given file has been preserved by an academic institution
- Data is traceable even when presented in contexts other than mirrors
- We can find evidence of any data that was ever downloaded from a particular piece of metadata
 - i.e., search for all files derived from a given Data.gov ID
- A point from which to begin coordinating larger efforts



Contact

Andrew Battista

Librarian for Geospatial Information Systems
New York University
ab6137@nyu.edu

Stephen Balogh

Data Services Specialist
New York University
sgb334@nyu.edu

Slides from this talk will be available from the Library of Congress website. You can also contact us and we will share them with you directly.

Resources

Battista, A. & Balogh, S. (2017) “The Challenge of Rescuing Federal Data: Thoughts and Lessons.” blog post available at <https://data-services.hosting.nyu.edu/the-challenge-of-rescuing-federal-data-thoughts-and-lessons/>

“Rethinking Institutional Repository Strategies: Report of a CNI Executive Roundtable.” May, 2017. Available at <https://www.cni.org/wp-content/uploads/2017/05/CNI-rethinking-irs-exec-rndtbl.report.S17.v1.pdf/>

Communities

Data Refuge - <http://www.datarefuge.org>
Environmental Data and Government Initiative (EDGI) - <https://envirodatagov.org/>
End of Term Harvest - <http://eotarchive.cdlib.org/>
Climate Mirror - <http://climatemirror.org/>

Standards & Projects

Svalbard - <https://github.com/datproject/svalbard>

