

# Symposium on the Interagency Strategic Plan for Big Data: Focus on R&D

NAS Board on Research Data and Information  
October 23, 2014

## **Big Data R&D Strategy III**

**Ensure the long term sustainability, access, and development of high value data sets and data resources.**

# Big Data R&D Strategies: Theme III

## ENSURE THE LONG TERM SUSTAINABILITY, ACCESS, AND DEVELOPMENT OF HIGH VALUE DATA SETS AND DATA RESOURCES

Support research on designing new platforms that most effectively enable data sharing. Research what is most effective in existing data sharing systems and build on resources available today. Support research in both universal data interfaces for general use and tailored interfaces for domain specific use.

### Global Change Information System and the National Climate Assessment (USGCRP)



### USGS ScienceBase and the National Fish Habitat Partnership



Develop new technical approaches to interoperability to overcome major current bottlenecks in data analytics.

Problem: Lots of discrete downloads and repeated mashing up takes time and resources and can introduce inconsistency and error



One Solution: Build and release bigger, integrated databases

Invest in data science pilots for collaborative cyberinfrastructure for research communities' use and analysis of data.

PaleoBioDB - Real time live debate of the worldwide paleontology research community



Support cultural and social change, long-term sustainability, ethical use, and privacy considerations due to the transformational nature of Big Data.

Wyoming Landscape Conservation Initiative - synthesizing big data to understand the interplay between multiple societal drivers



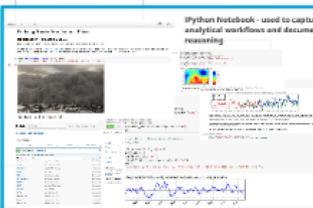
Accelerate the development of an "open source" data tools and techniques movement (including Interfaces, metadata, standards, platforms, etc.), analogous to the open source software movement that incorporates the needs of the public, researchers, industry, and the government

OGC Web Processing Service - standardized inputs and outputs for data processing



USGS Geo Data Portal  
Making massive amounts of disseminated climate model data accessible for research analysis through subjecting and summarizing Center for Integrated Data Analytics

Incentivize scientists to publish negative results, failed tests, and disproved hypotheses as a research artifact. Research how context can impact data analysis and how it can be stored as metadata to allow future analysis. Develop a workflow registration and repository to capture and share not only analytics, but analytical processes.



Support public private partnerships focused on developing challenges, improving access, ensuring security, and improving interoperability (e.g., seed a public private cloud infrastructure for research).

MQSIS and other earth observation data incorporated into Google Earth Engine sparking new pathways for research analysis (Matt Hansen and others)



1

Support research on designing new platforms that most effectively enable data sharing. Research what is most effective in existing data sharing systems and build on resources available today. Support research in both universal data interfaces for general use and tailored interfaces for domain specific use.

# Global Change Information System and the National Climate Assessment

## *U.S. Global Change Research Program*

**Description of evidence base**

The key message and supporting text summarizes extensive evidence documented in the Rural Communities Workshop Report.<sup>12</sup> Thirty one technical input reports on a wide range of topics were also received and reviewed as part of the Federal Register Notice solicitation for public input.

Evidence that the impacts of climate change are increasing is compelling and widespread. This evidence is based on historical records and observations and on global climate models, including those driven by B1 (substantial emissions reduction) and A2 (continued increases in global emissions) scenarios. This evidence is clearly summarized and persuasively referenced in the “Our Changing Climate” chapter of this Assessment and in the Scenarios developed for the NCA.<sup>13</sup>

The dependency of rural communities on their natural resources has been demonstrated,<sup>14</sup> with a number of studies showing that climate change results in crop and livestock loss,<sup>15</sup> infrastructure damage to levees and roads,<sup>16</sup> shifts in agriculture practices,<sup>17</sup> and losses due to disasters.<sup>18</sup> A number of publications project these impacts to increase, with effects on the natural environment<sup>19,20,21</sup> and increased competition for water between agriculture and energy.<sup>22,23</sup> Studies have projected that tourism locations in the Everglades and Florida Keys are threatened.<sup>24</sup> Meanwhile, Maine’s tourism could increase,<sup>25</sup> which coincides with a projected northern shift in outdoor recreation.<sup>26</sup> Hunting, fishing, and bird watching will be affected by beach erosion and wetland loss,<sup>27</sup> and changing plant and animal habitats and inter-species relationships (see also Ch. 8: Ecosystems). Outdoor recreation and tourism in many areas in the U.S. are affected by early snowpack melt.<sup>28,29</sup>



Develop new technical approaches to interoperability to overcome major current bottlenecks in data analytics.

# Problem: Lots of discrete downloads and repeated mashing up takes time and resources and can introduce inconsistency and error



USGS Home  
Contact USGS  
Search USGS

## Ocean Biogeographic Information System USA

[CSAS&L](#) / [OBIS-USA Home](#) / [Data Search and Access](#) / [Search](#)

↑ [OBIS-USA Home](#)

[Data Search and Access](#)

[Data Use Agreement](#)

[Contributing Data](#)

[Biogeographic Tool](#)

[Dataset & Participant Listing](#)

OR

[Taxonomic Search](#)

OR

[Geography Search](#)

1. Select the map and arrow icon to the right
2. Draw a bounding box on the map for data results in that area

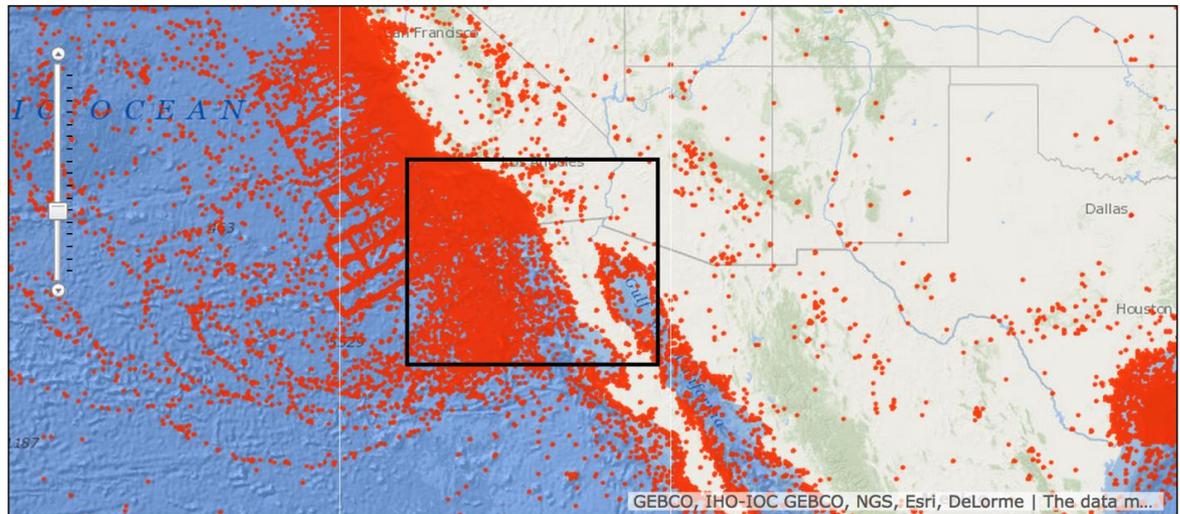


### OBIS-USA DATA DISCLAIMER

To access OBIS-USA data records you must agree to Use Constraints and cite the data as specified in the Citation Text for each dataset.

If you proceed, OBIS-USA will provide Use Constraints and Citation Text in the data you obtain.

By clicking any "Download" or "Customized Access" link you are indicating that you agree to observe the Use Constraints and cite the dataset using the Citation Text as provided by the metadata for each dataset used.



### Geographic Search Results:

122,819 Records Found / 25 Dataset(s) / 41,230 Observation Locations

[Dataset Summary](#) | [Analysis by Taxon](#) | [Observation Dates and Depth Profile](#)

122,819 Results Found: [Download All](#) [Access All](#)

Dataset(s)	Record(s)	Metadata	Download	Customized Access
MMS Low Altitude Survey for Seabirds, Southern California Bight	7650	<a href="#">View</a>	<a href="#">Download</a>	<a href="#">Customized Access</a>

## One Solution: Build and release bigger, integrated databases

3

Invest in data science pilots  
for collaborative  
cyberinfrastructure for  
research communities' use  
and analysis of data.

# PaleoBioDB - Real time live debate in the worldwide paleontology research community

PBDB Navigator Beta



Paleobiology Database Classic

Quick search

Full search

Download

About

Log in

## Community



Name	Institution
Melanie Hopkins	American Museum of Na
Jin Meng	American Museum of Na
Jocelyn Sessa	American Museum of Na
Achim Herrmann	Arizona State University
Lisa Churchill-Dickson	Augusta Maine
Phil Novack-Gottshall	Benedictine University
Bill Stein	Binghamton University
Vladimir Davydov	Boise State University
Scott Foss	Bureau of Land Manage
David Blackburn	California Academy of S
Nicole Bonuso	California State Universit
Chris Beard	Carnegie Museum of Na

Total: 131 collections including 148 occurrences

4

Support cultural and social change, long-term sustainability, ethical use, and privacy considerations due to the transformational nature of Big Data.

# Southwest Wyoming: Booming energy development competing with world class wildlife, agriculture, recreation, and other resources



## Wyoming Landscape Conservation Initiative

Conserving world-class wildlife resources. Facilitating responsible development.

Map Help Return to Integrated Assessment Overview

### Integrated Assessment

Theme: Multi-criteria Index

Opacity: 0 100



### Other Assessment Layers

Theme: Focal Ecosystems Mask

### Reference Layers

Theme: Land Ownership

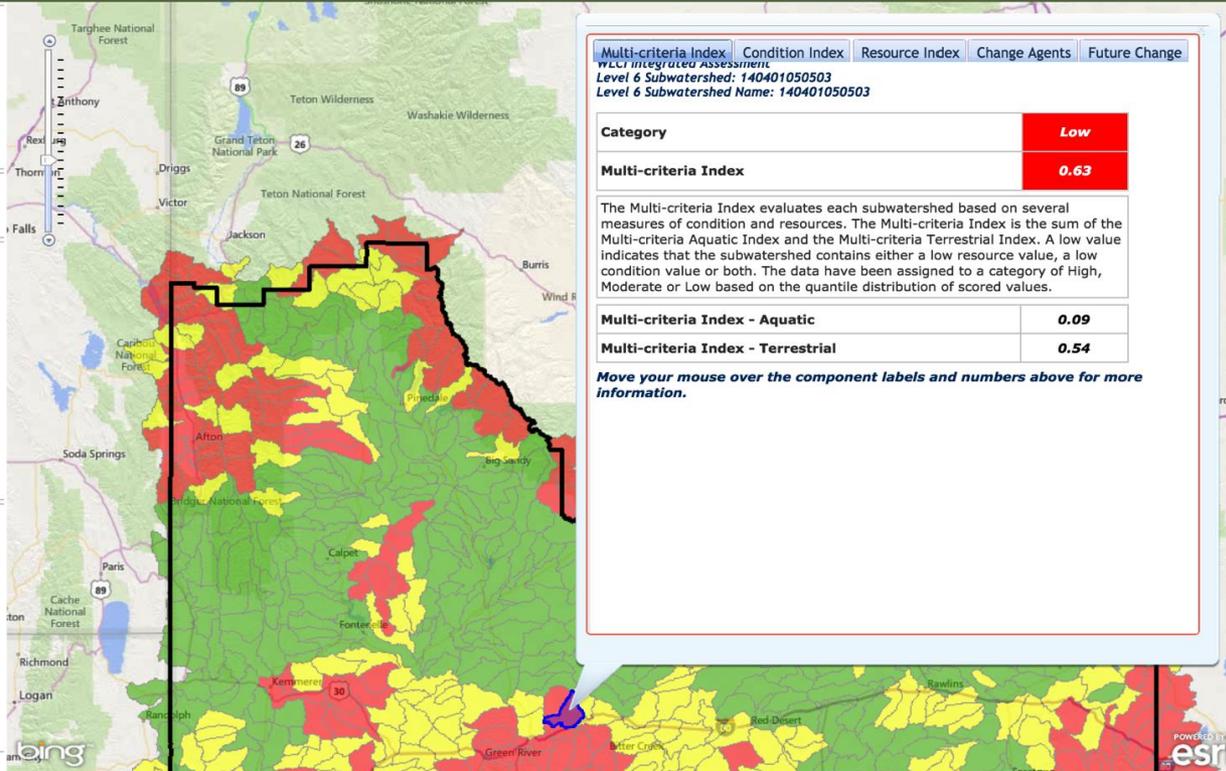
- Bureau of Indian Affairs
- Bureau of Land Management
- Bureau of Reclamation
- Fish and Wildlife Service
- Forest Service
- National Park Service
- Private
- State
- Water

### Source Data

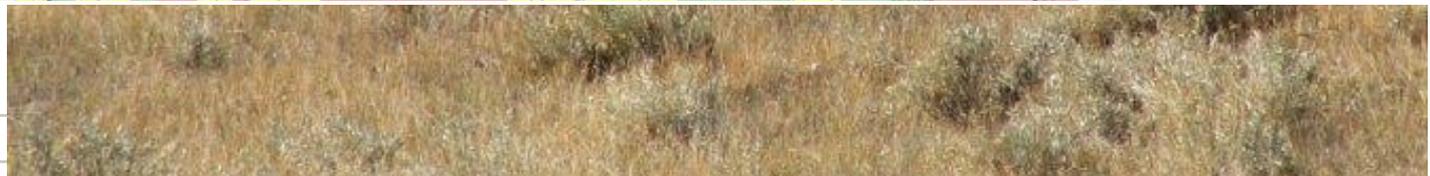
Theme: None

- Change Agents
- Condition
- Future Change
- Multi-criteria Index
- Other Data
- Resource Index

Low Value High Value



Map Aerial Terrain



5

Accelerate the development of an "open source" data tools and techniques movement (including interfaces, metadata, standards, platforms, etc.), analogous to the open source software movement that incorporates the needs of the public, researchers, industry, and the government

# OGC-Web Processing Service - standardized inputs and outputs for data processing



USGS Geo Data

This page is a cat  
core, the Geo Dat  
standa  
below.



Actual ET  
on the Sin  
from remc  
pre-define  
Gowda, P  
Journal of  
comprehe  
Environme

GitHub

This repository Search

Explore Features Enterprise Blog

Sign up

Sign in

USGS-CIDA / geo-data-portal

★ Star 1

Fork 10

2,269 commits

3 branches

13 releases

8 contributors

branch: master geo-data-portal / +

Merge pull request #142 from prusso-sse/master

isuftin authored 15 days ago

latest commit d51f1e4343

gdp-52n-wps-tests	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-common-utilities	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-core-processing	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-csw-client	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-data-access	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-derivative-processing	A couple changes as suggested by Ivan	a year ago
gdp-derivative-ui	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-process-wps	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-proxy	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-ui-landing	Fix for JIRA GDP-830	15 days ago
gdp-ui	[maven-release-plugin] prepare for next development iteration	2 months ago
gdp-utility-wps	[maven-release-plugin] prepare for next development iteration	2 months ago
geoserver	[maven-release-plugin] prepare for next development iteration	2 months ago
.gitignore	Merge commit	a year ago

<> Code

Issues 0

Pull Requests 0

Pulse

Graphs

HTTPS clone URL

https://github.com/

You can clone with HTTPS or Subversion.

Clone in Desktop

Download ZIP

arized

<> Code

Issues 43

Pull Requests 1

Wiki

Pulse

Graphs

HTTPS clone URL

https://github.com/

You can clone with HTTPS or Subversion.

Clone in Desktop

Download ZIP

## USGS Geo Data Portal

*Making massive amounts of downscaled climate model data accessible for research analysis through subsetting and summarization*

Center for Integrated Data Analytics

6

Incentivize scientists to publish negative results, failed tests, and disproved hypotheses as a research artifact.

Research how context can impact data analysis and how it can be stored as metadata to allow future analysis.

Develop a workflow registration and repository to capture and share not only analytics, but analytical processes.

# iPython Notebook - used to capture analytical workflows, document reasoning, and communicate science

nbviewer [FAQ](#) [IPython](#)

GitHub

This repository Search

[Explore](#) [Features](#) [Enterprise](#) [Blog](#)

Sign up

Sign in

 [rsignell-usgs / notebook](#)

★ Star 2

 Fork 5

In [31]:

```
ipython notebooks
```

73 commits

3 branches

0 releases

2 contributors



branch: master

[notebook](#) / +



Merge branch 'master' of github.com:rsignell-usgs/notebook



rsignell-usgs authored 7 days ago

latest commit [01a3b0e367](#)

 CSW	only ipynb and py files	2 months ago
 Cartopy	added Cartopy examples	6 months ago
 ERDDAP	made ERDDAP directory	7 days ago
 IRIS	added tests	7 months ago
 People	adding lots of notebooks	8 months ago
 Presentations	adding lots of notebooks	8 months ago
 ROMS	adding lots of notebooks	8 months ago
 SOS	updated ncSOS endpoints	2 months ago
 UGRID	added ESTOFS	8 months ago
 WFS	updated WFS output	5 months ago
 hsvpy-runtalk	only ipynb and py files	2 months ago
 secoora	only ipynb and py files	2 months ago
 system-test	only ipynb and py files	2 months ago

In [2]

<> Code

Issues 1

Pull Requests 0

Pulse

Graphs

HTTPS clone URL

<https://github.com/>

You can clone with [HTTPS](#) or [Subversion](#).

Clone in Desktop

Download ZIP

7

Support public private partnerships focused on developing challenges, improving access, ensuring security, and improving interoperability (e.g., seed a public private cloud infrastructure for research).

# MODIS and other earth observation data incorporated into Google Earth Engine sparking new pathways for research analytics (Matt Hansen and others)

Science

AAAS.ORG | FEEDBACK | HELP | LIBRARIANS

Google Maps

AAAS

NEWS

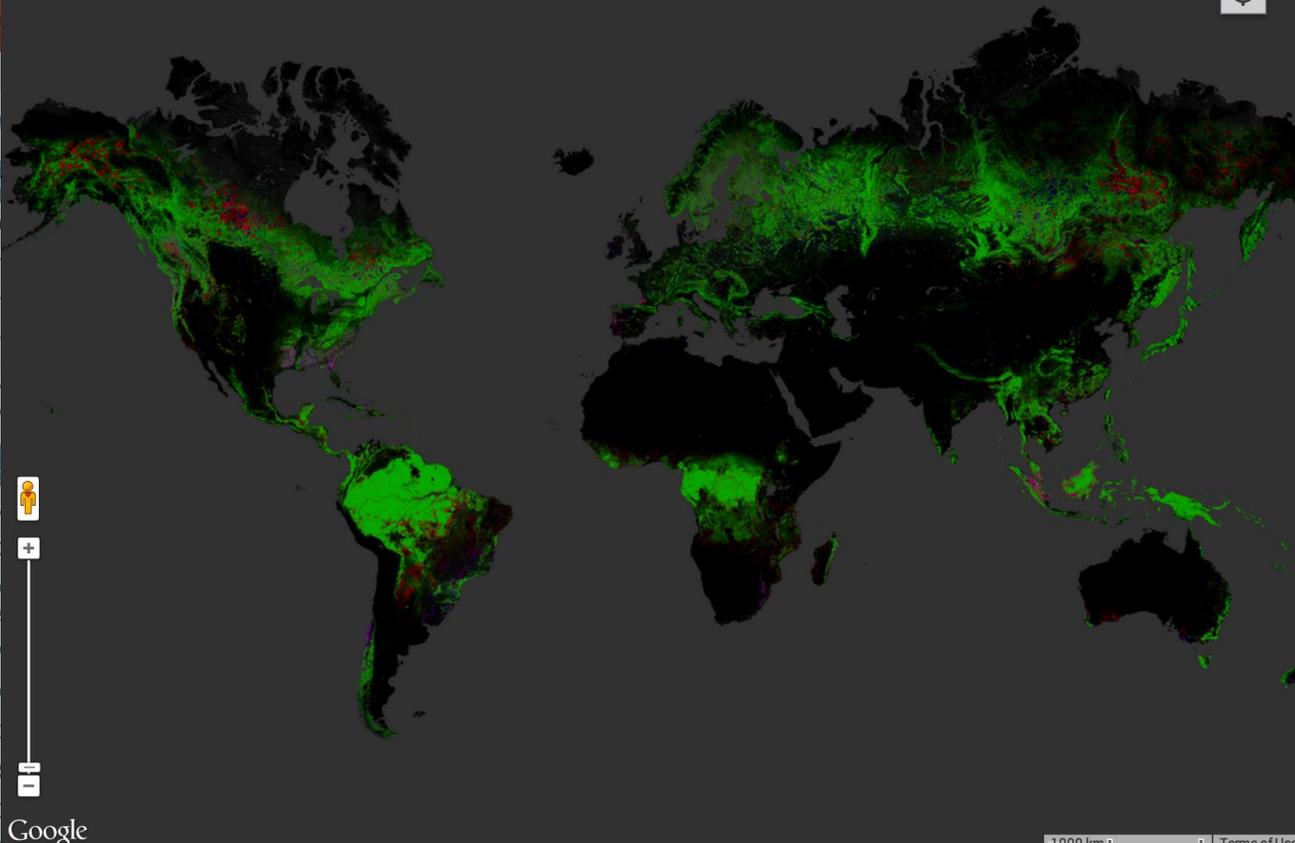
SCIENCE JOURNALS

CAREERS

Enter a location

Share

Map



## Global Forest Change

Published by Hansen, Potapov, Moore, Hancher et al.



UNIVERSITY OF MARYLAND  
DEPARTMENT OF GEOGRAPHICAL SCIENCES

Results from time-series analysis of 654,178 Landsat images in characterizing forest extent and change, 2000–2012.

Trees are defined as all vegetation taller than 5m in height and are expressed as a percentage per output grid cell as '2000 Percent Tree Cover'. 'Forest Loss' is defined as a stand-replacement disturbance, or a change from a forest to non-forest state. 'Forest Gain' is defined as the inverse of loss, or a non-forest to forest change entirely within the study period. 'Forest Loss Year' is a disaggregation of total 'Forest Loss' to annual time scales.

Reference 2000 and 2012 imagery are median observations from a set of quality assessment-passed growing season observations.

[Download the data.](#)

[Reset to default view](#)

Data Products

Loss/Extent/Gain (Red/Green/Blue)

- Legend
- Forest Loss 2000–2012
  - Forest Gain 2000–2012
  - Both Loss and Gain
  - Forest Extent

Background Imagery

Year 2000 Bands 5/4/3

Example Locations

Forestry and Tornado in Alabama

[Zoom to area](#)

The trail of destruction from the April 27 2011 Tornado in Birmingham is clearly visible in this

Published by Hansen, Potapov, Moore, Hancher et al. · Powered by Google Earth Engine · [Data Download](#) · [Help](#)

consistent and locally relevant record of for



# Thank you

**Sky Bristol**

USGS Core Science Systems

*Core Science Analytics, Synthesis, and Libraries*

