



World Data Center for Human Interactions in the Environment

*Science Archives in the 21<sup>st</sup> Century  
University of Maryland, Adelphi MD*

# Government-University Collaboration in Long-Term Archiving of Scientific Data

Robert S. Chen

Director & Senior Research Scientist; SEDAC Manager; CODATA Secretary-General

Robert R. Downs

Senior Digital Archivist; SEDAC Archives Manager

W. Christopher Lenhardt

Associate Director, Information Services; Deputy Director, SEDAC

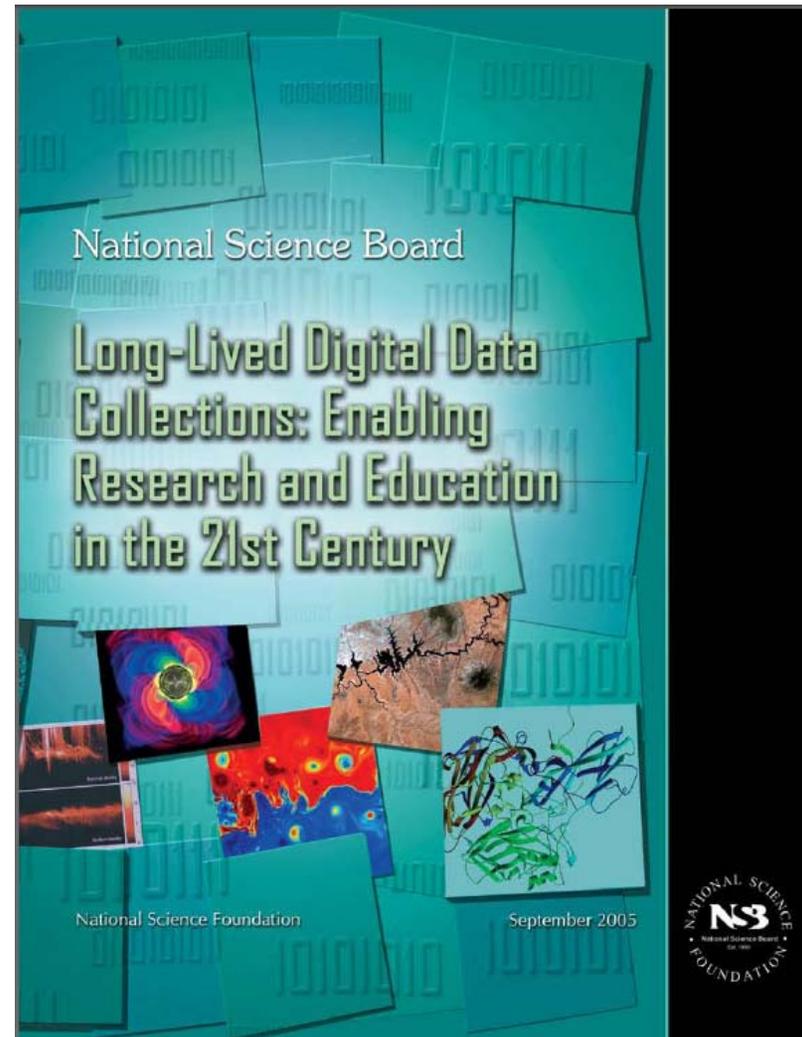
*Center for International Earth Science Information Network (CIRESIN)  
Socioeconomic Data and Applications Center (SEDAC)  
The Earth Institute at Columbia University*

*25 April 2007*

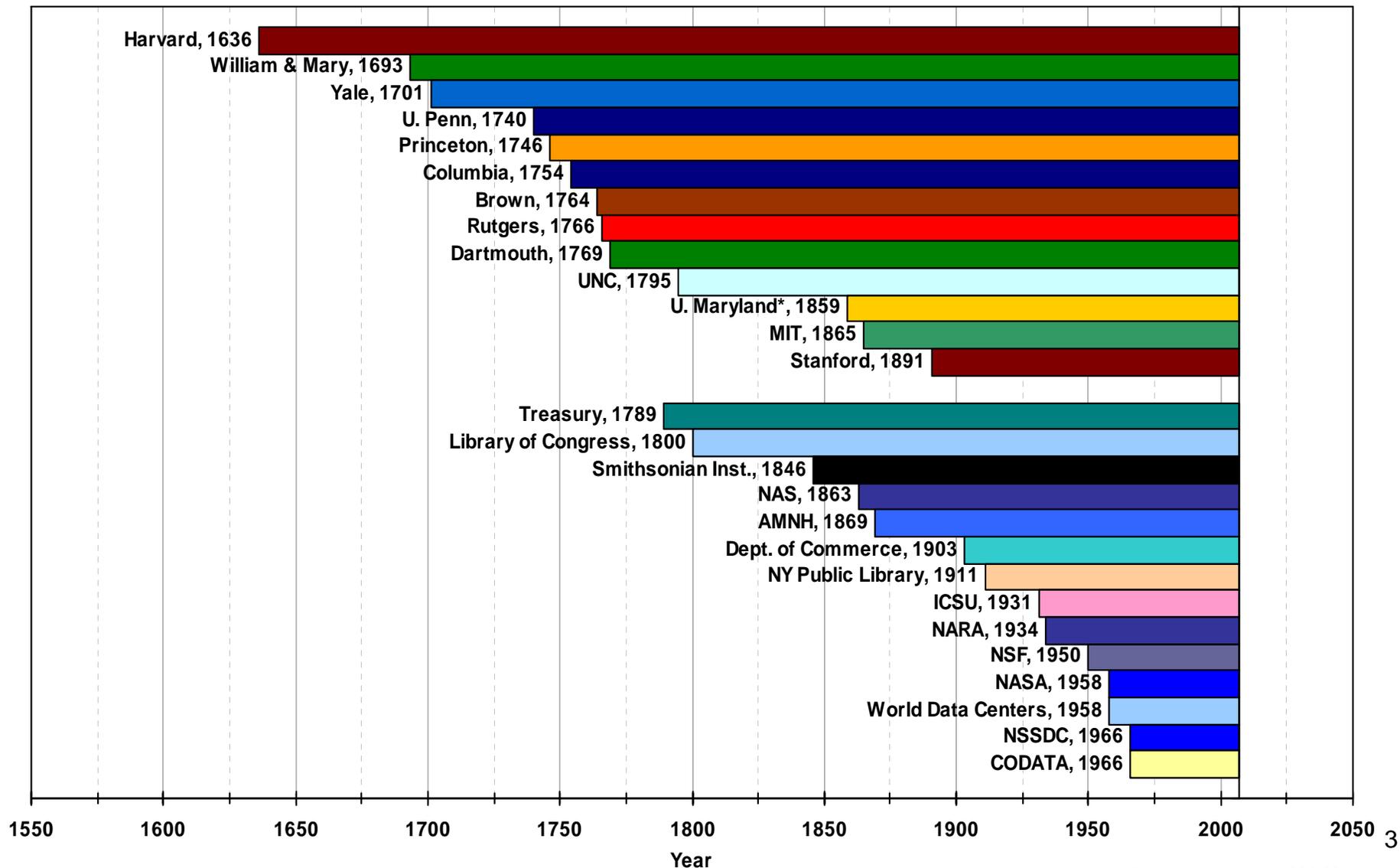


Columbia University  
in the City of New York

- “Digital data collections resemble large facility projects in terms of their extended lifetime; the need for stable, core support; the critical importance of effective project management in combination with domain expertise; the ability to energize and enable broad research and education communities; and the importance of partnerships, both national and international.” (p. 40)
- “Furthermore, unlike instrument-based facilities, data collections tend to increase in value the longer they are in operation, attracting ever-expanding groups of data users as the amount of data they include increases and spans greater periods of time.” (p. 41)



# Longevity of Selected Universities, Government Agencies, and Other Institutions



## *Advantages*

- Long-term commitment to knowledge generation, preservation, and dissemination
- Strong domain expertise, especially in data integration and analysis
- Close ties with both research and education communities
- Strong national & international partnerships
- Established expertise in non-digital preservation, access, and use (i.e., libraries, research collections, museums)
- Rapidly evolving digital infrastructure
- Emerging capabilities in digital data stewardship, knowledge management, and open access

## *Disadvantages*

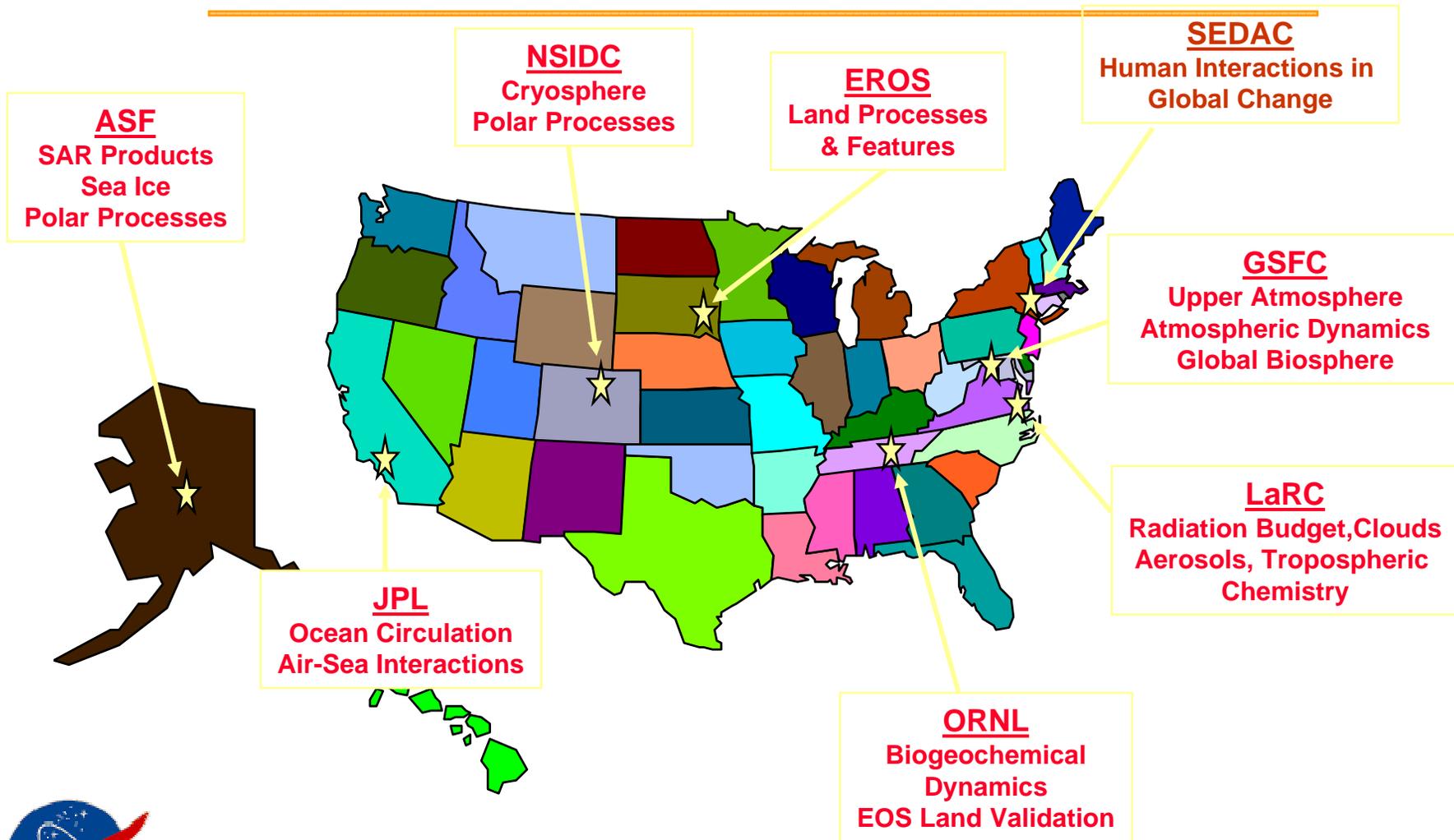
- Lagging in addressing overall digital data needs
- Often limitations in structure for large technical organizations not directly related to University operations
- Traditional departmental/disciplinary structures may encourage “stovepipe” databases, data collections, duplication
- Reluctant to include information technology costs in infrastructure (i.e., in indirect cost rate)
- Lack of communication between faculty, researchers, IT experts, librarians, data managers, etc.
- Tensions between open access principles and intellectual property concerns and proprietary approaches

## *Advantages*

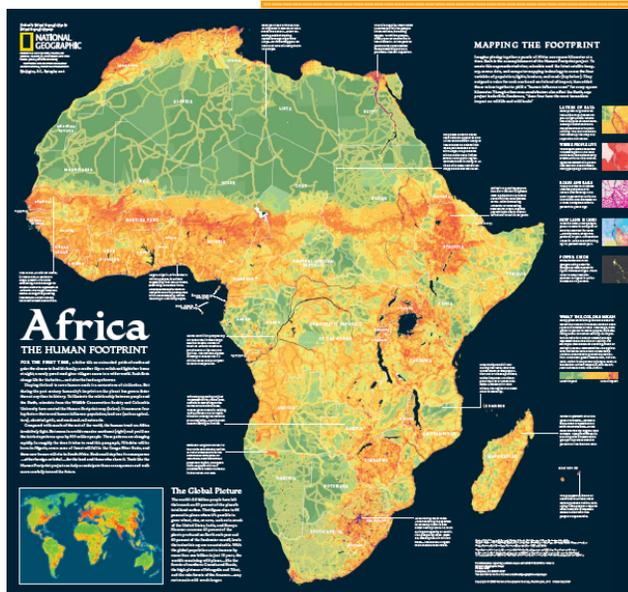
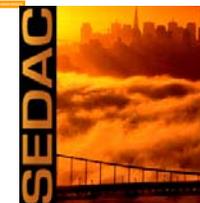
- Mission focus
- Strong domain expertise, especially in new instrumentation, large-scale observing systems, large-scale database development
- Close ties with user communities
- Strong national & international partnerships
- Established expertise in digital preservation, access, and use (existing data and research centers)
- Rapidly evolving digital infrastructure
- Leadership role in development of standards, interoperability
- Supportive of open access principles and policies

## *Disadvantages*

- Mission, organization can be significantly changed over time
- Mission doesn't necessarily prioritize knowledge preservation and access
- Budget, operations subject to annual fiscal fluctuations, problems (e.g., cost overruns in other programs!)
- Traditional organizational mandates and structures may encourage "stovepipe" databases, data collections, & duplication
- Information technology implementation constrained by procurement procedures and other policies
- Tendency towards outsourcing of functions; fewer government personnel with technical expertise



**SEDAC = Socioeconomic Data and Applications Center  
One of 8 Distributed Active Archive Centers (DAACs) in the  
NASA Earth Observing System Data & Information System (EOSDIS)**



- Focus on human dimensions of environmental change
- Integration of social and Earth science data, especially with remote sensing
- Direct support to scientists, applied and operational users, decision makers, and policy communities

Treaty Multiple Status Matrix - netsc.page.6

ENTR Environmental Treaties and Resource Indicators

home | treaty locator | country explorer

**Status Matrix**

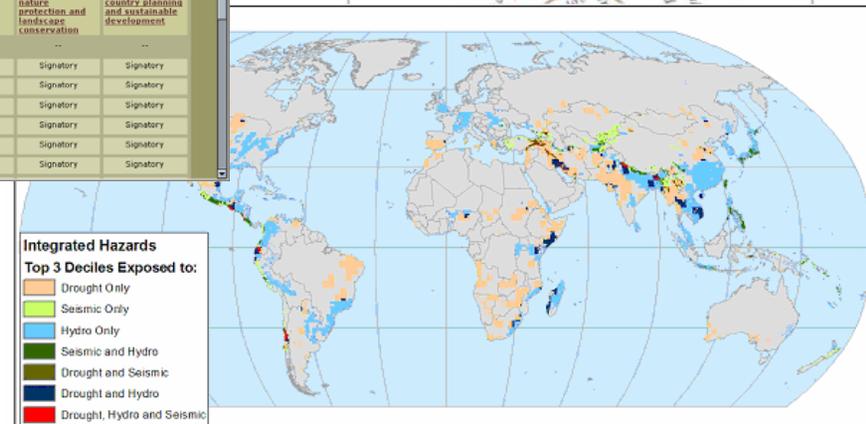
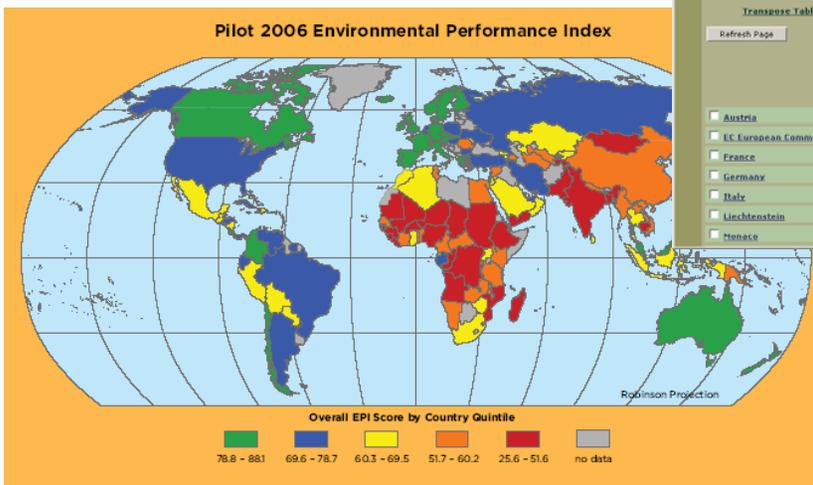
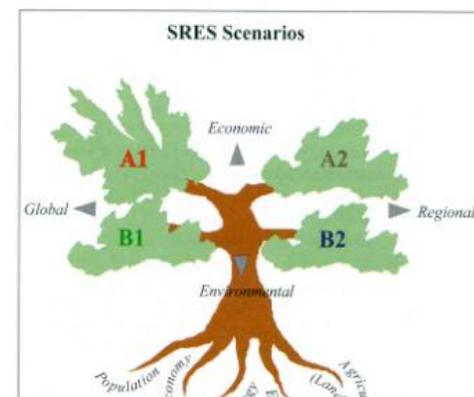
-Definitions for [Signatory](#) | [Party](#) | [Former Party](#)  
 -Remove items from either columns or rows by unchecking them, then click refresh page.  
 -Add items to cart by checking, then click refresh page or view cart.  
 -Clicking on an agreement or party will take you to a summary for that item.

[VIEW CART](#)

download this matrix | treaty results.set | country results.set

View in matrix: All Relevant Countries | Display this variable: Status

Transgress Table	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
Refresh Page	Convention concerning the Protection of Alps	Protocol for the implementation of the Alpine Convention in the field of mountain agriculture	Protocol for the implementation of the Alpine Convention in the field of nature protection and landscape conservation	Protocol for the implementation of the Alpine Convention in the field of town and country planning and sustainable development
<input type="checkbox"/> Austria	Party	...	...	...
<input type="checkbox"/> EC European Communities	Party	Signatory	Signatory	Signatory
<input type="checkbox"/> France	Party	Signatory	Signatory	Signatory
<input type="checkbox"/> Germany	Party	Signatory	Signatory	Signatory
<input type="checkbox"/> Italy	Signatory	Signatory	Signatory	Signatory
<input type="checkbox"/> Liechtenstein	Party	Signatory	Signatory	Signatory
<input type="checkbox"/> Monaco	...	Signatory	Signatory	Signatory

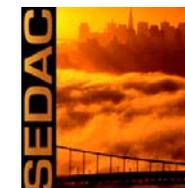




Version (pub)	GPW v1 (1995)	GPW v2 (2000)	GPW v3 (2005)
Estimates for	1994	1990, 1995	1990, 1995, 2000
Input units	19,000	127,000	~ 375,000

More than 180 citations of GPW versions 1 and 2

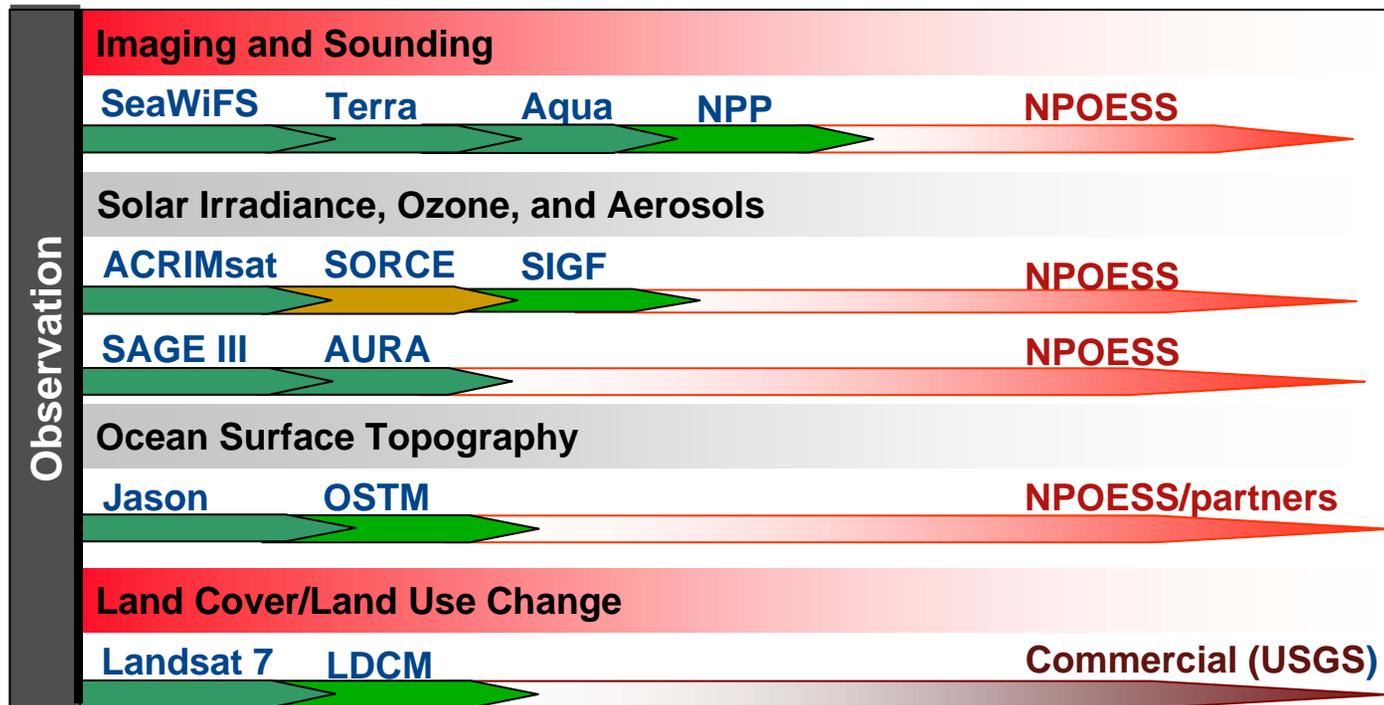
<http://sedac.ciesin.columbia.edu/gpw/>



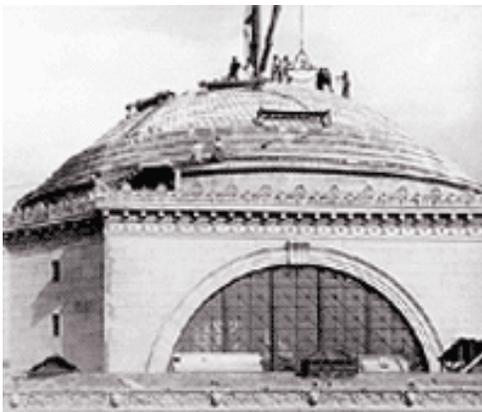
# DAACs Do Not Have a Long-Term Charge



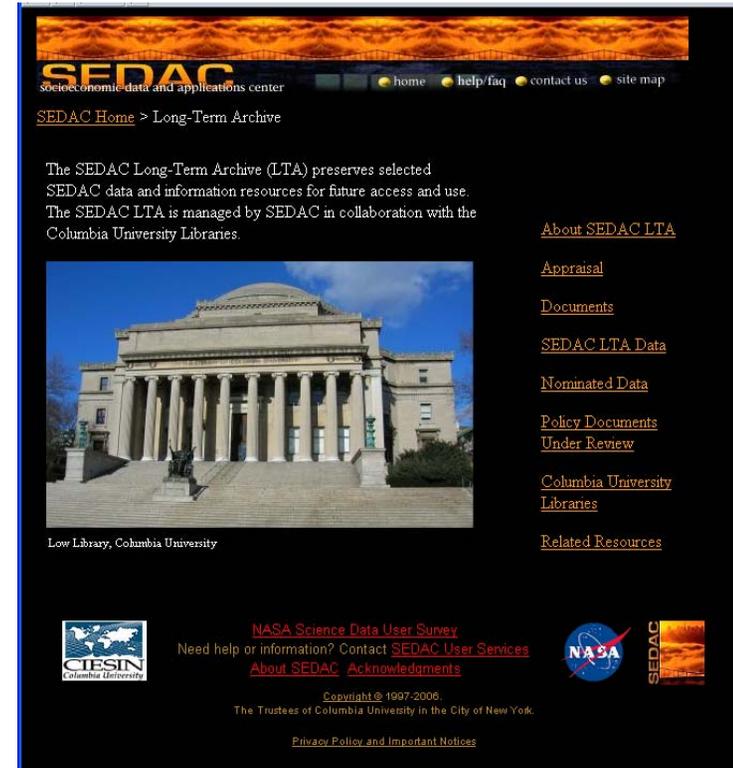
- NASA as a research agency is supposed to transition observations to NOAA, an operational agency
  - Earth Observing System program could end around 2015
  - SEDAC is on a five-year contract; could be terminated before then.
- What happens to SEDAC's data and information resources if SEDAC disappears??



- After SEDAC's relocation to Columbia in 1998, CIESIN began to explore the University Library as a suitable long-term home for a SEDAC long-term archive (LTA)
- **SEDAC LTA Mission:**
  - The SEDAC Long-Term Archive acquires, preserves, and maintains the content of selected high-quality data, data products, documentation, and services relevant to human dimensions of global change in a digital form to support the discovery, access, and use of archived resources by scientific, educational, and decision-making communities for at least the next 50 years.



**Low Memorial Library  
circa 1897**

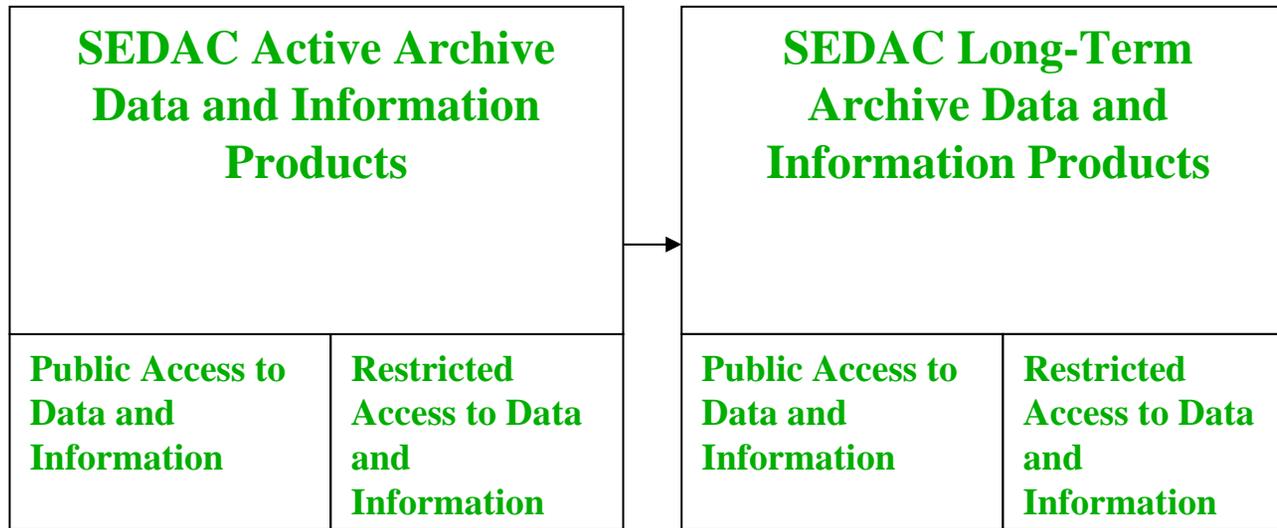


<http://sedac.ciesin.columbia.edu/ita>



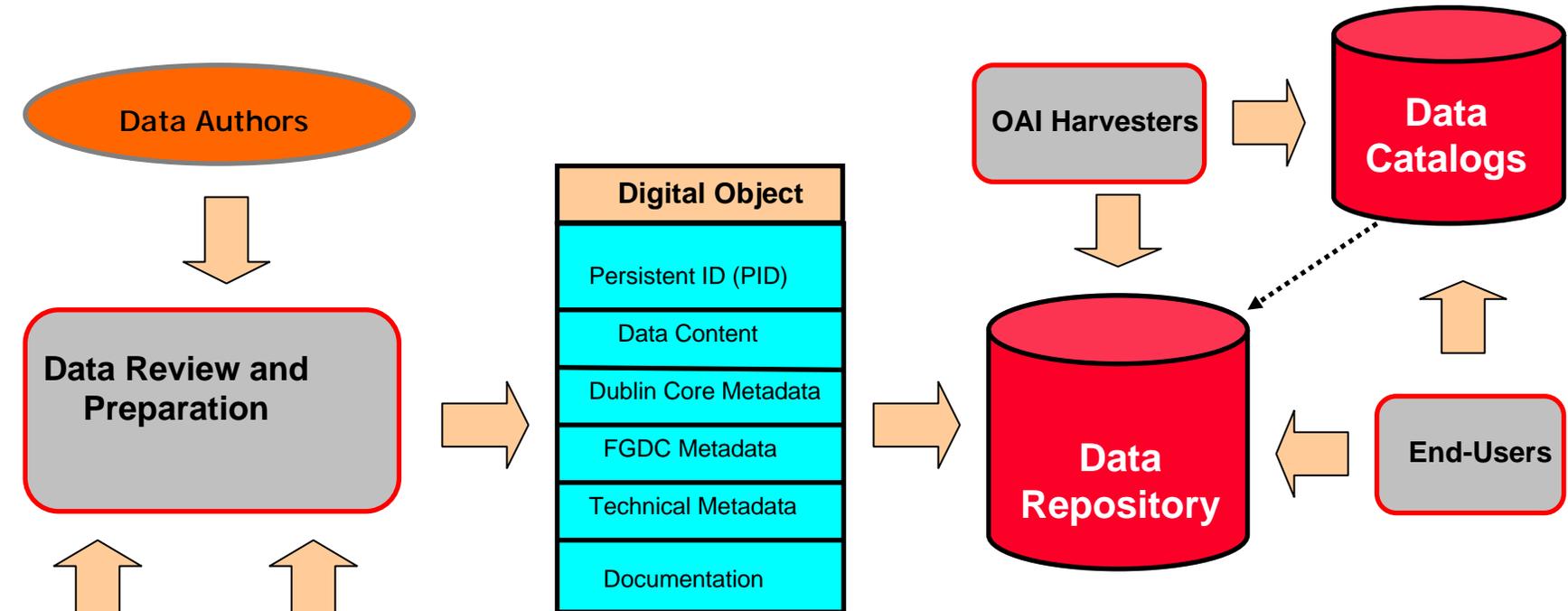
- **LTA Board established with representation from SEDAC, the Earth Institute, and the Columbia University Libraries:**
  - SEDAC Project Scientist
  - SEDAC Systems Engineer
  - SEDAC Archives Manager (serves as Chair)
  - Two representatives designated by Earth Institute
  - Two representatives designated by Columbia University Libraries
- **If SEDAC discontinues operations at Columbia University**
  - CIESIN will designate a replacement for one SEDAC position
  - Columbia University Library will appoint replacements for the other two positions, including the chair

## SEDAC Digital Object Repository



**Active Archive** is for near-term dissemination with high levels of service. Primary users are discipline-specific scientists.

**Long-Term Archive** is for the 50 – 100 year preservation time-frame with different expectations for levels of service.



**Data authors contribute data and related documentation**  
**Data is reviewed and prepared for ingest**  
**A Persistent Identifier (PID) is assigned by Handles server**  
**Technical metadata is validated using JHOVE server**  
**Digital object is ingested in data repository**  
**Open Archives Initiative (OAI) Harvesters get Metadata**  
**OAI Harvesters deposit metadata in data catalogs**  
**End-users discover data in data catalogs**  
**End-users access data from data repository**

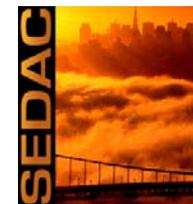
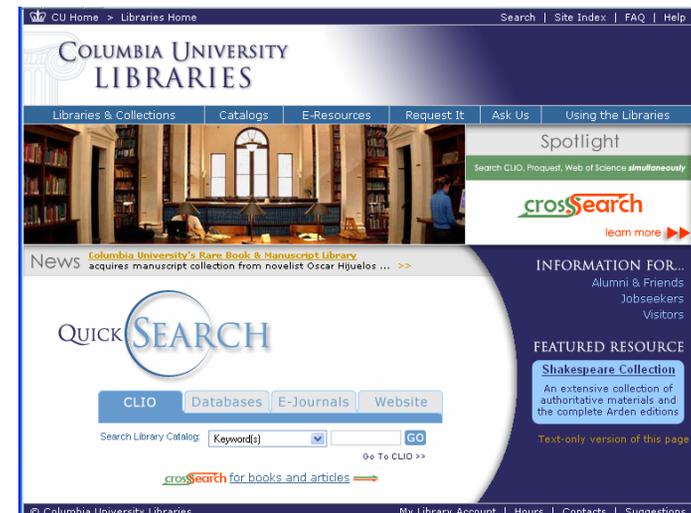
- Complete the initial LTA technical infrastructure
  - Installation of VITAL recently completed
  - Create Archival Information Packages (AIPs) and Dissemination Information Packages (DIPs)
  - Begin formal dissemination of accessioned LTA data
- Continue strategic planning with CU Libraries, Information Services, and the Earth Institute
- Conduct self-assessment using Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) and other reports
- Explore expansion of LTA to support other CIRESIN, Earth Institute, and Columbia University data resources
- Build on LTA as example of collaboration between the research community and academic libraries in long-term digital preservation



- Columbia University community has >250 years of experience in preserving knowledge for future generations
- Fosters organizational learning on digital preservation
- Interdepartmental effort enhances LTA sustainability
- Columbia University Libraries contribute perspectives on supporting diverse users and uses
- Earth Institute contributes perspectives on science community needs
- SEDAC contributes data life cycle perspectives on data management, preservation, and dissemination
- Interdisciplinary scientific communities share experiences on developments to improve data archiving



Columbia University  
in the City of New York



- Data management co-located with domain expertise, research and educational users
- Longevity of universities combined with resources of government
- Universities somewhat insulated from day-to-day fluctuations in budget, government “fire drills”
- Government can help universities to jumpstart capabilities and infrastructure in digital data infrastructure
- Ongoing government involvement and review to:
  - prevent formation of “dead” archives
  - ensure use of current technology and standards
  - promote full and open access

## Selection Criteria for LTA Data Appraisal

### – Scientific or Historical Value

- citation, research, and educational use as published in refereed scientific publications/reports from recognized committee of scientists

### – Potential Usability and Use

- evidence of usability, usefulness, and sufficient usage by the community interested in human dimensions of the environment. Adequate evidence indicate potential for future use justifies costs of long-term archiving

### – Uniqueness of Data (non-redundant stewardship)

- not being preserved in any form in another archive and is at risk of loss if not accessioned into the Long-Term Archive

### – Relevance to LTA Mission

- currently endorsed or approved by community interested in human interactions in the environment. For the short-term, relevance includes content germane to SEDAC mission and SEDAC strategic plan

### – Documented for Accessibility

- completeness and correctness of documentation to facilitate future discovery, access, and use

### – Technological Accessibility (feasibility)

- received in format meeting technical criteria for the Service Level designated for the resource

### – Legality and Confidentiality

- unrestricted permissions for preservation and future dissemination. No information that is confidential or prohibited from dissemination

### – Non-Replicability

- data replication not feasible, excessively costly or prohibitive

## Standards and Best Practices

- Metadata Standards:
  - Identification: Generating Unique IDs, Also Implementing Handles
  - Dublin Core (ISO 15836-2003)
  - FGDC CSDGM -> ISO 19115
  - Implementing JHOVE to extract technical metadata
  - Reviewing Requirements to Adopt PREMIS and GML
- Best Practices
  - Longevity: Contingency Plans for Continuation of LTA Board and LTA Management Community-Based Transparency: Selection and Appraisal for Accession by Interdisciplinary LTA Board
  - Platform Independence: Implementing Fedora, an Open Source, JAVA-Based, Modular Digital Repository Platform to Reduce Proprietary Dependencies
  - Encoded Object Encapsulation, Metadata, and Object Relationships: XML
  - Redundancy: Offsite Archiving of Security Masters, Implementing Synchronized Failover System
  - Integrity Validation: Generation of SHA-1 Fixity Signatures on Ingest
  - Media Reliability: Moving from Portable Media (CDs and DVDs) to Digital Repository Infrastructure
  - Assurance: Reviewing Requirements with CCSDS WG Forming to Propose ISO Standard for Certification of Trusted Digital Repositories (TDR)

## Flexible Extensible Digital Object Repository Architecture (Fedora)

- Open Source Software
- Content Management
  - Unique Persistent Identifiers for Each Digital Object
  - Each Digital Object contains Content and Metadata Datastreams to create an Open Archival Information System (OAIS) Compliant Archival Information Package (AIP)
  - Changes to Digital Objects are Stored as New Versions
  - Enables Ingest and Management of Various Data Types
- Web-Based Content Dissemination
  - Search Supported by Resource Indexing
  - Objects Assigned Behavior Definition and Dissemination Methods
- Ingest, Store, and Export in Extensible Markup Language (XML)
- Collection and Object Relationship Management Using Resource Description Framework (RDF) Graphs
- Supports Open Archival Initiative Protocol Metadata Harvesting (OAI/PMH)

## VTLS Information Technology for Advanced Learning (VITAL): Added-Value Features for SEDAC

- VITAL Modules Integrated with Fedora
  - VITAL Advanced Server
  - VITAL Access Portal
    - Web-Based Discovery and Access for Public
      - Browsing, and Simple, Advanced, and Full-Text Search Capabilities
    - Web-Based Administration and Content Manager Client for Staff
  - VITAL Batch Ingest Utility
  - VTLS Automated Loading and Electronic-submission Tool (VALET)
    - Enables Workflow, Author Submission, Cataloging, and Review
  - Lightweight Directory Access Protocol (LDAP) Authentication Server
  - JStore Harvard Object Validation Environment (JHOVE)
  - Handle System Server
    - Assigns Unique Identifiers that are Resolved to URLs
  - Search Retrieve Web / URL (SRW/SRU) and Z39.5 Services
- VTLS Services
  - Installation and Testing
  - Training
  - Support 24/7/365
  - Maintenance
  - Fedora Upgrades
  - VITAL Releases and Enhancements

