

20th International CODATA Conference
Beijing, China



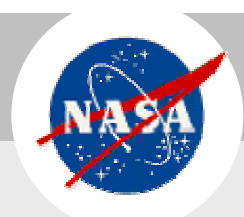
**Development of a Long-
Term Interdisciplinary
Data Archive with the
Columbia University
Library System**

24 October 2006

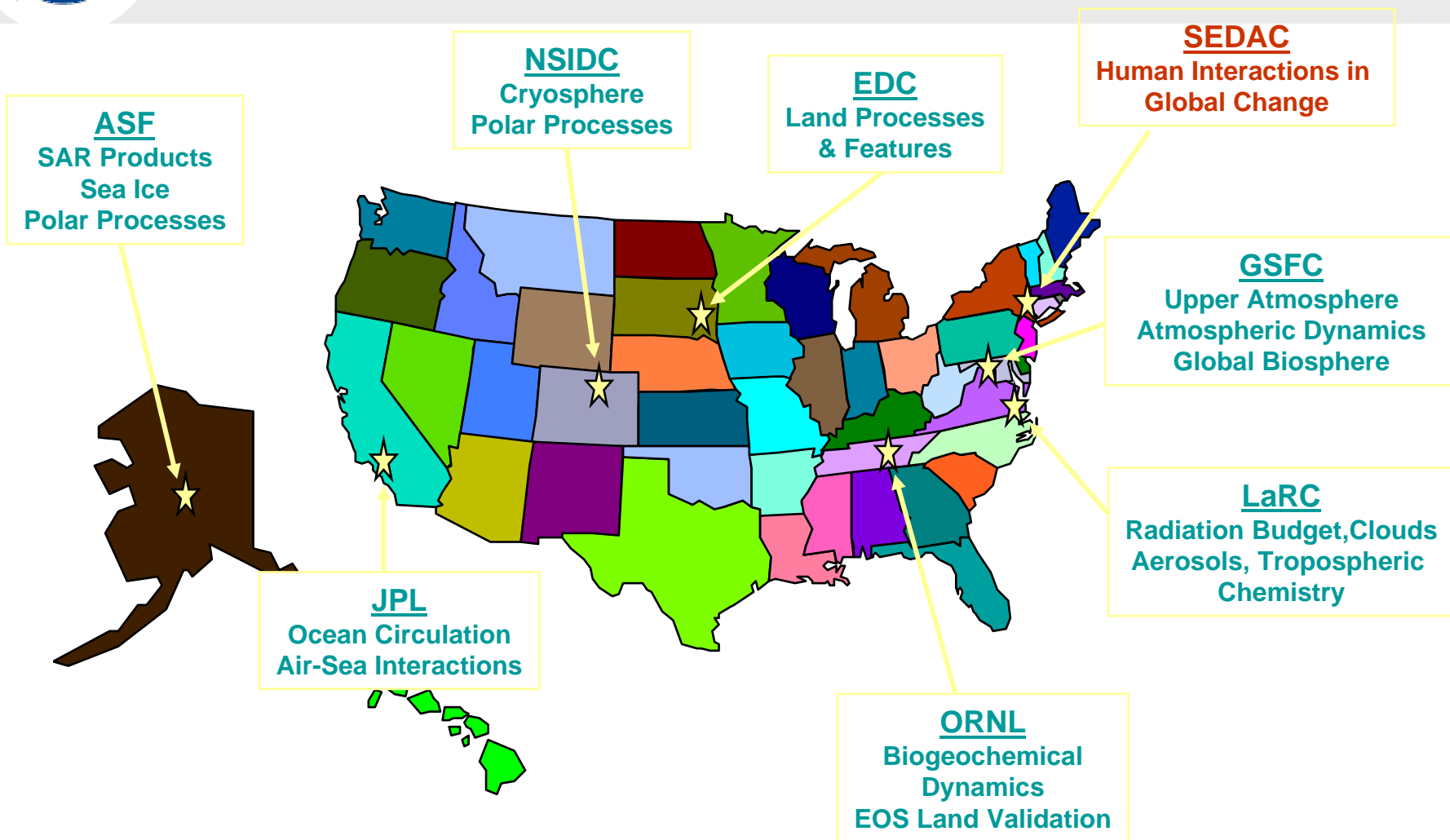
**Robert S. Chen, Robert R. Downs, and W. Christopher Lenhardt
CIESIN, Columbia University**



Columbia University
in the City of New York



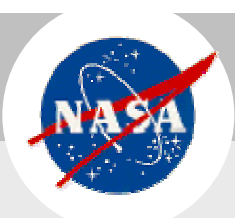
SEDAC is one of 8 NASA Active Archives



SEDAC = Socioeconomic Data and Applications Center
Based at CIESIN, part of the Earth Institute of Columbia
University in New York



Columbia University
in the City of New York



DAACs play a key role in the data system

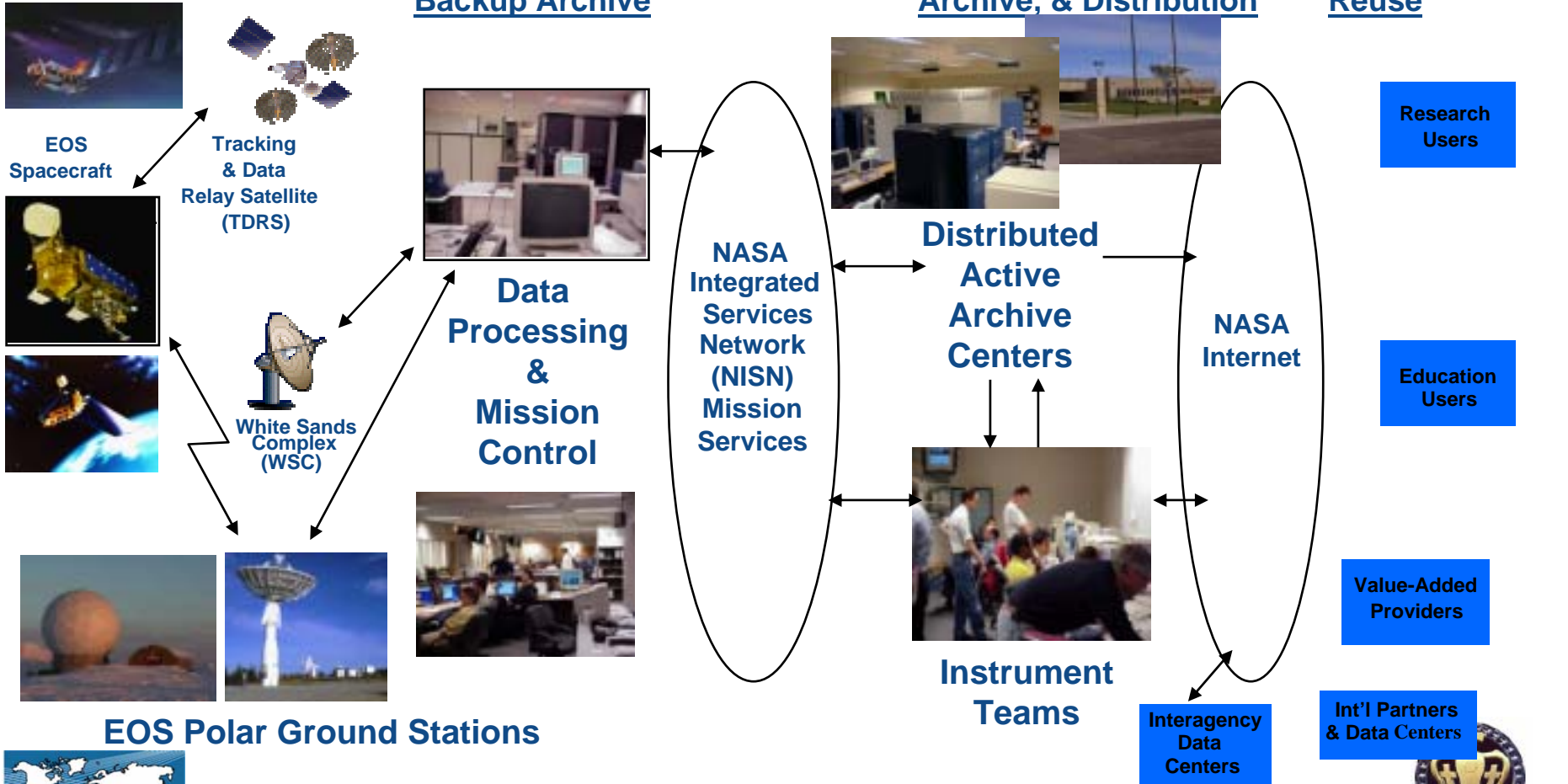
Data Acquisition

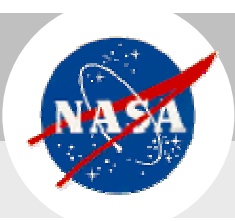
Flight Operations, Data Capture, Initial Processing, Backup Archive

Data Transport to DAACs

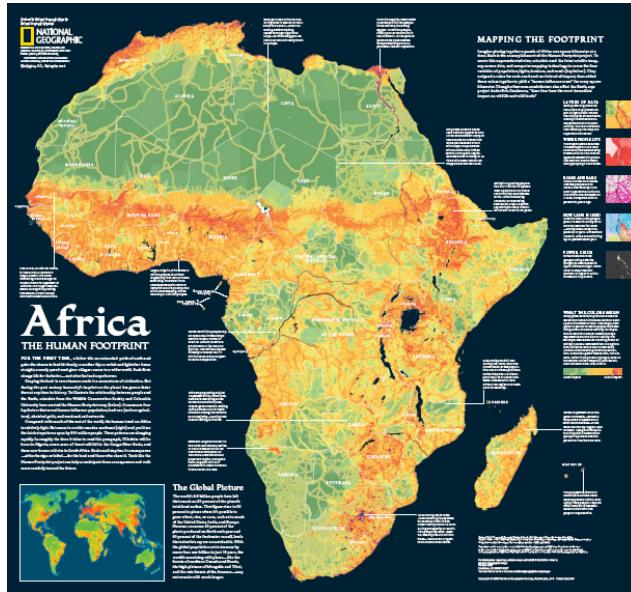
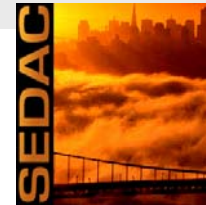
Science Data Processing, Info Mgmt, Data Archive, & Distribution

Distribution, Access, Interoperability, Reuse

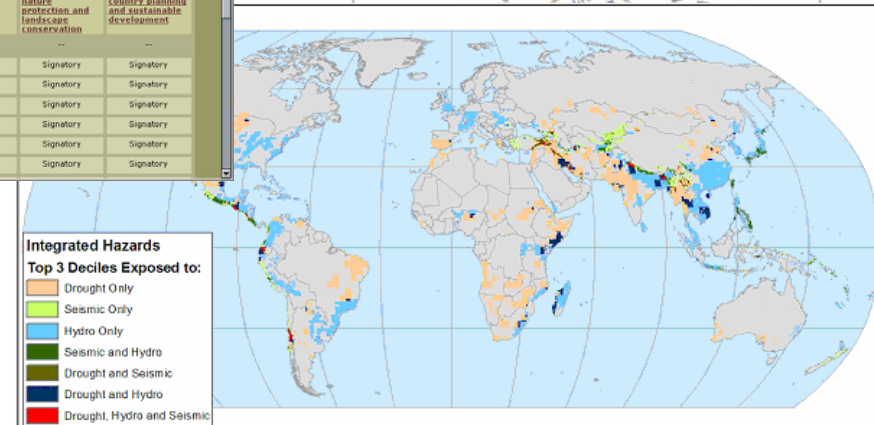
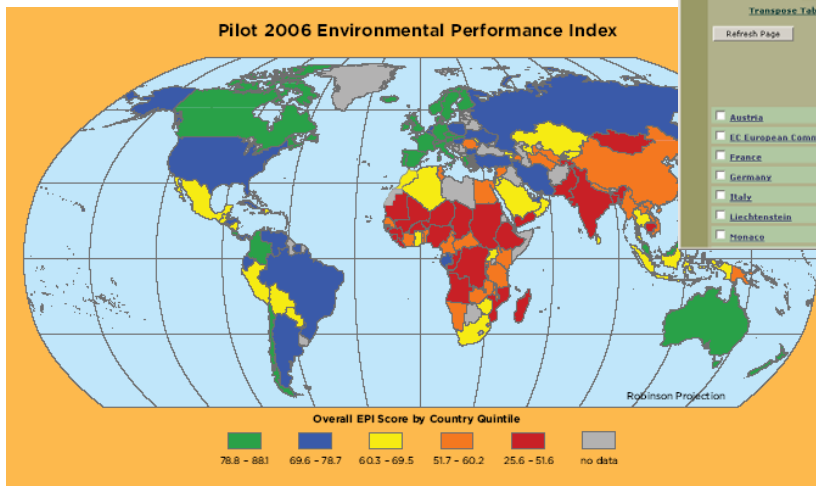
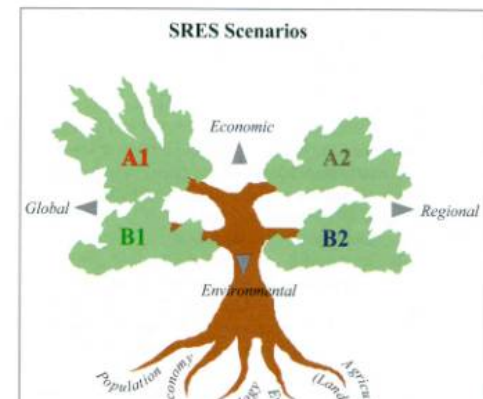


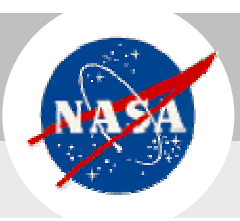


SEDAC supports a wide range of data



- 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

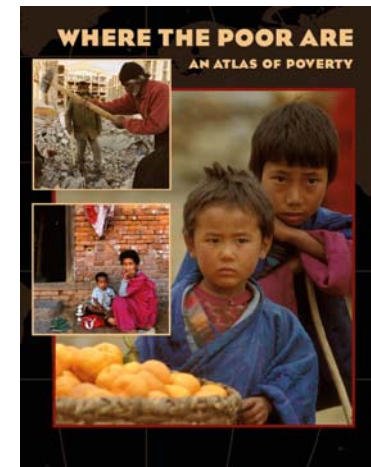
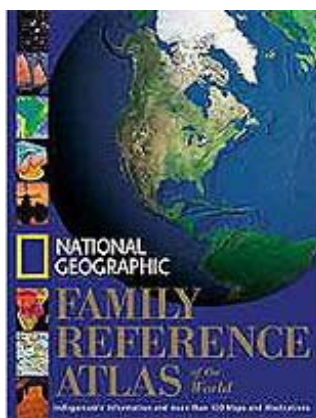
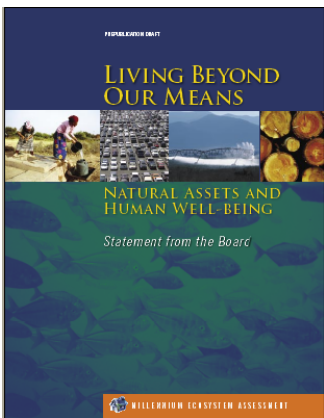
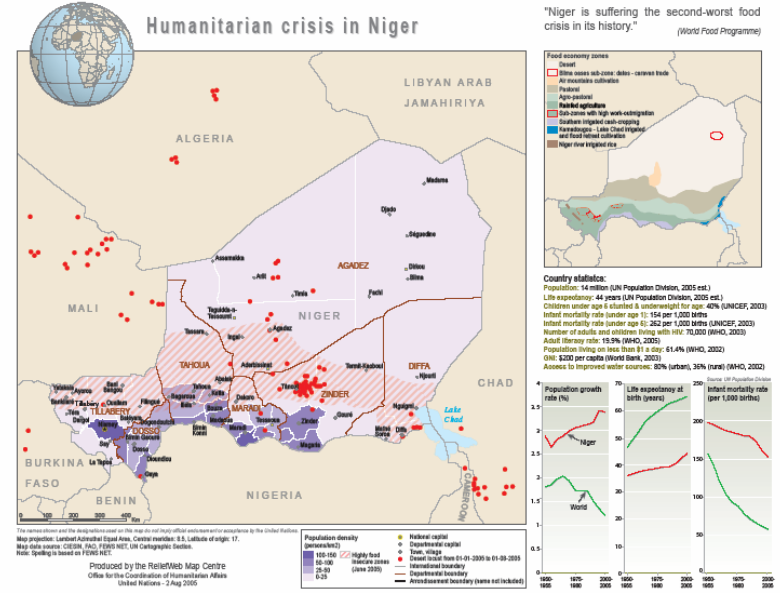


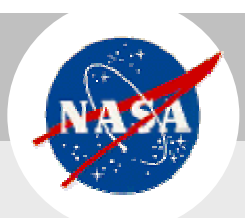


SEDAC users are diverse

Example Users:

- Millennium Ecosystem Assessment
- UN Millennium Project
- UN Geographic Information Support Team
- The World Bank
- *National Geographic*
- Earth & Sky
- *The Times Atlas*
- IPCC Fourth Assessment





Older SEDAC data need a long-term home

e.g., early versions of Gridded Population of the World

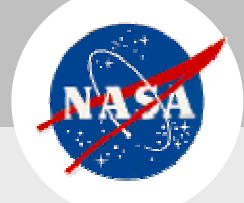


Version (pub)	GPW v1 (1995)	GPW v2 (2000)	GPW v3 (2005)
Estimates for	1994	1990, 1995	1990, 1995, 2000
Input units	19,000	127,000	~ 375,000



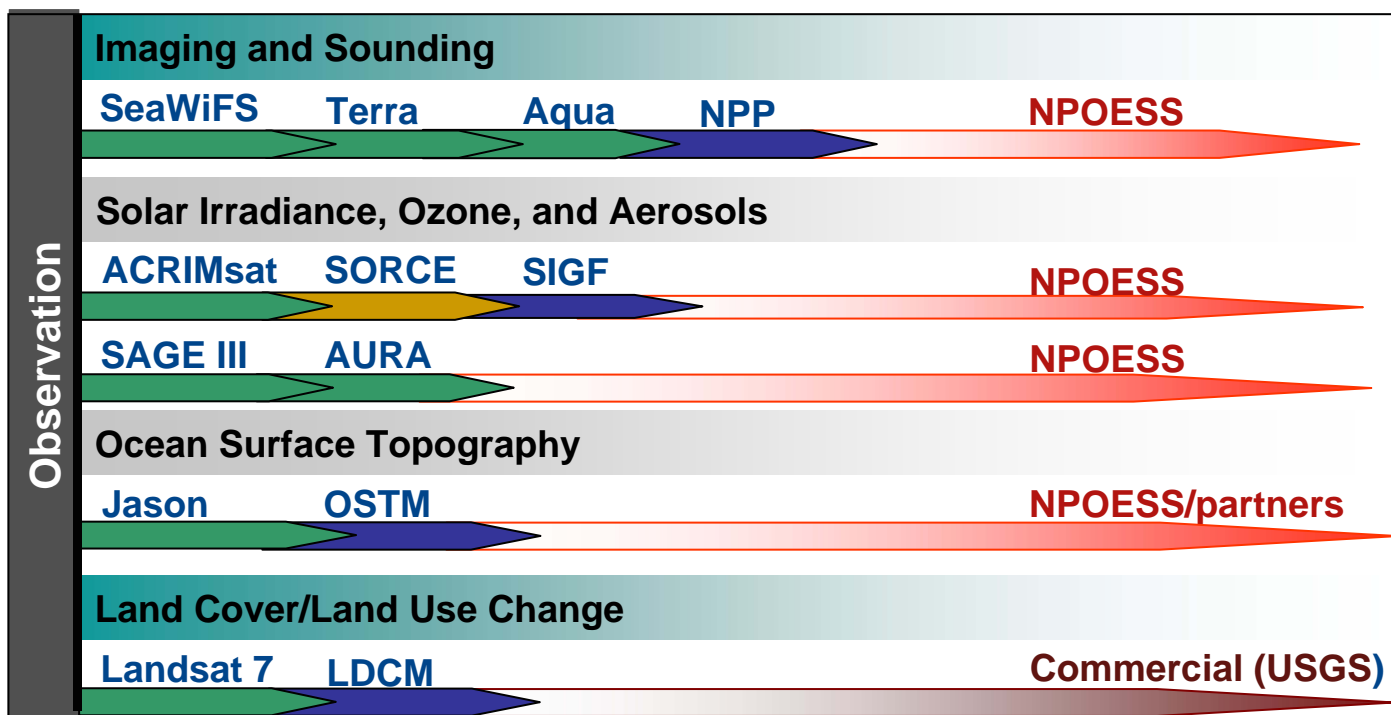
<http://sedac.ciesin.columbia.edu/gpw/>

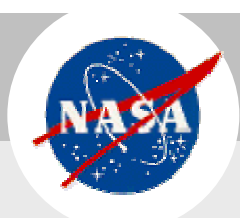




DAACs do not have a long-term charge

- 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??





SEDAC LTA at Columbia University

- Columbia University established in 1754 (before the U.S. government!)
- Library potentially a suitable long-term home for SEDAC long-term archive (LTA)



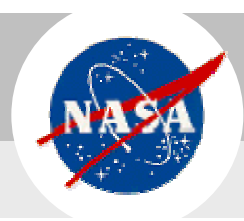
Columbia's first campus



**Low Memorial Library
circa 1897**



Low Memorial Library today



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.

SEDAC
socioeconomic data and applications center

[home](#) [help/faq](#) [contact us](#) [site map](#)

[SEDAC Home](#) > Long-Term Archive

The SEDAC Long-Term Archive (LTA) preserves selected SEDAC data and information resources for future access and use. The SEDAC LTA is managed by SEDAC in collaboration with the Columbia University Libraries.

[About SEDAC LTA](#)

[Appraisal](#)

[Documents](#)

[SEDAC LTA Data](#)

[Nominated Data](#)

[Policy Documents Under Review](#)

[Columbia University Libraries](#)

[Related Resources](#)

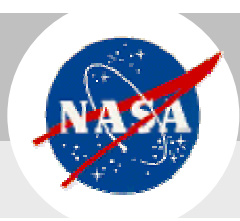
Low Library, Columbia University

[NASA Science Data User Survey](#)
Need help or information? Contact [SEDAC User Services](#)
[About SEDAC](#) [Acknowledgments](#)

Copyright © 1997-2006.
The Trustees of Columbia University in the City of New York.

[Privacy Policy and Important Notices](#)





SEDAC LTA Organizational Structure

■ SEDAC LTA Board

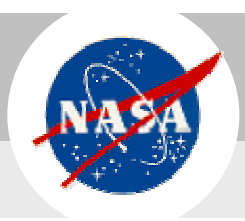
- Responsible for approving mission, goals, and strategic plans
- Responsible for approving appraisal criteria
- Appraises and selects data for accession

■ SEDAC LTA Manager

- Reports to the LTA Board
- Responsible for development and operations of LTA systems, including staff and procedures, to ensure data stewardship
- If SEDAC operations are discontinued, university appoints LTA Manager

■ SEDAC LTA Staff

- Report to LTA Manager
- Responsible for accessioning and maintaining LTA holdings in accordance with LTA procedures
- If SEDAC operations are discontinued, University appoints LTA staff members



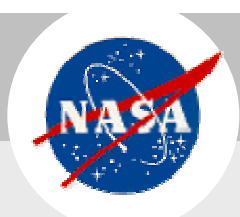
SEDAC LTA Board

■ 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



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)

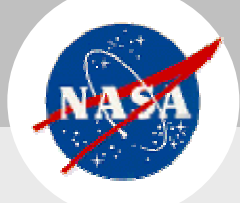
- received in format meeting technical criteria for the Service Level designated for the resource

■ Legality and Confidentiality

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

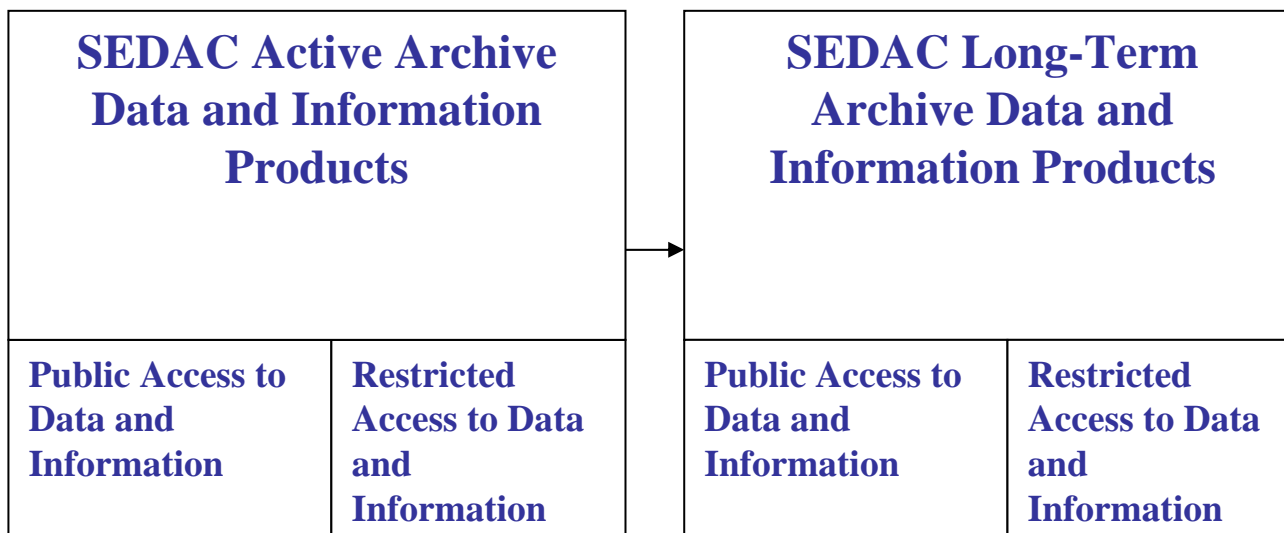
■ Non-Replicability

- data replication not feasible, excessively costly or prohibitive



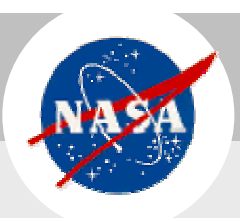
SEDAC Data Repository Organization

SEDAC Digital Object Repository

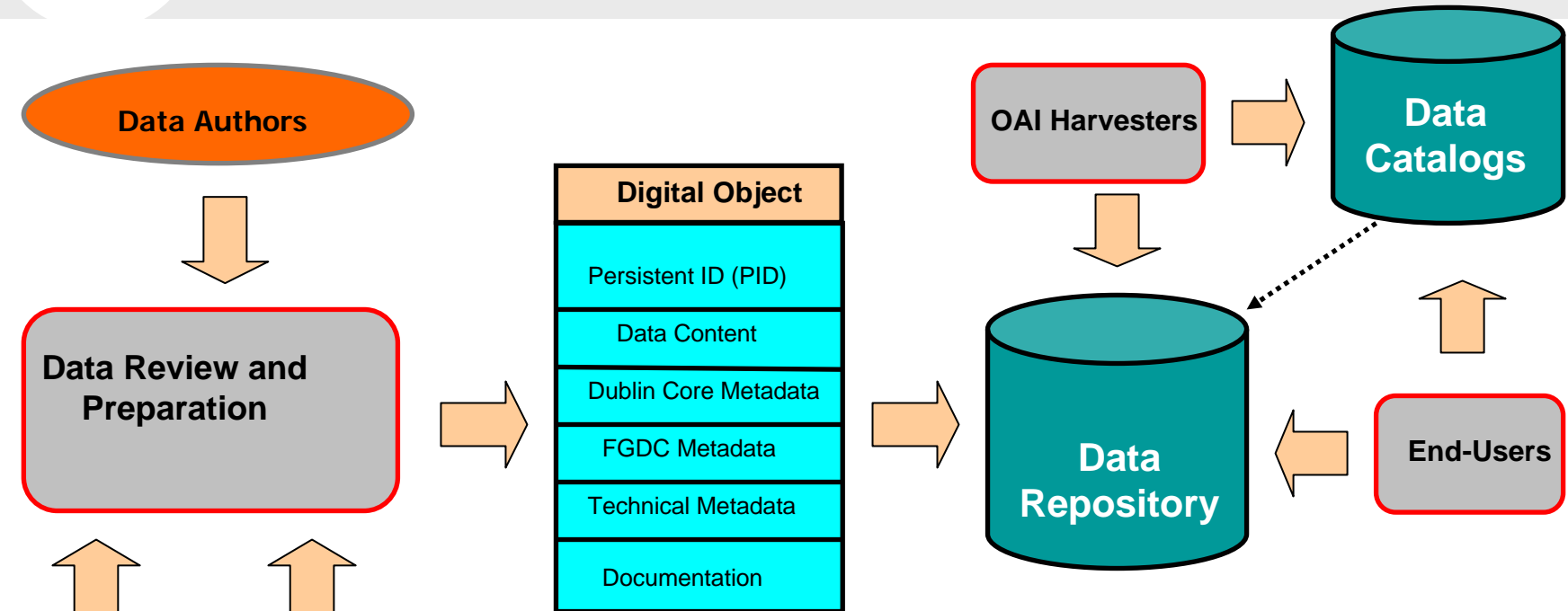


Active Archive is for near-term dissemination with high levels of service. Primary users are discipline-specific scientists.

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

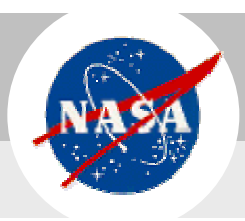


Use of Fedora to Implement LTA

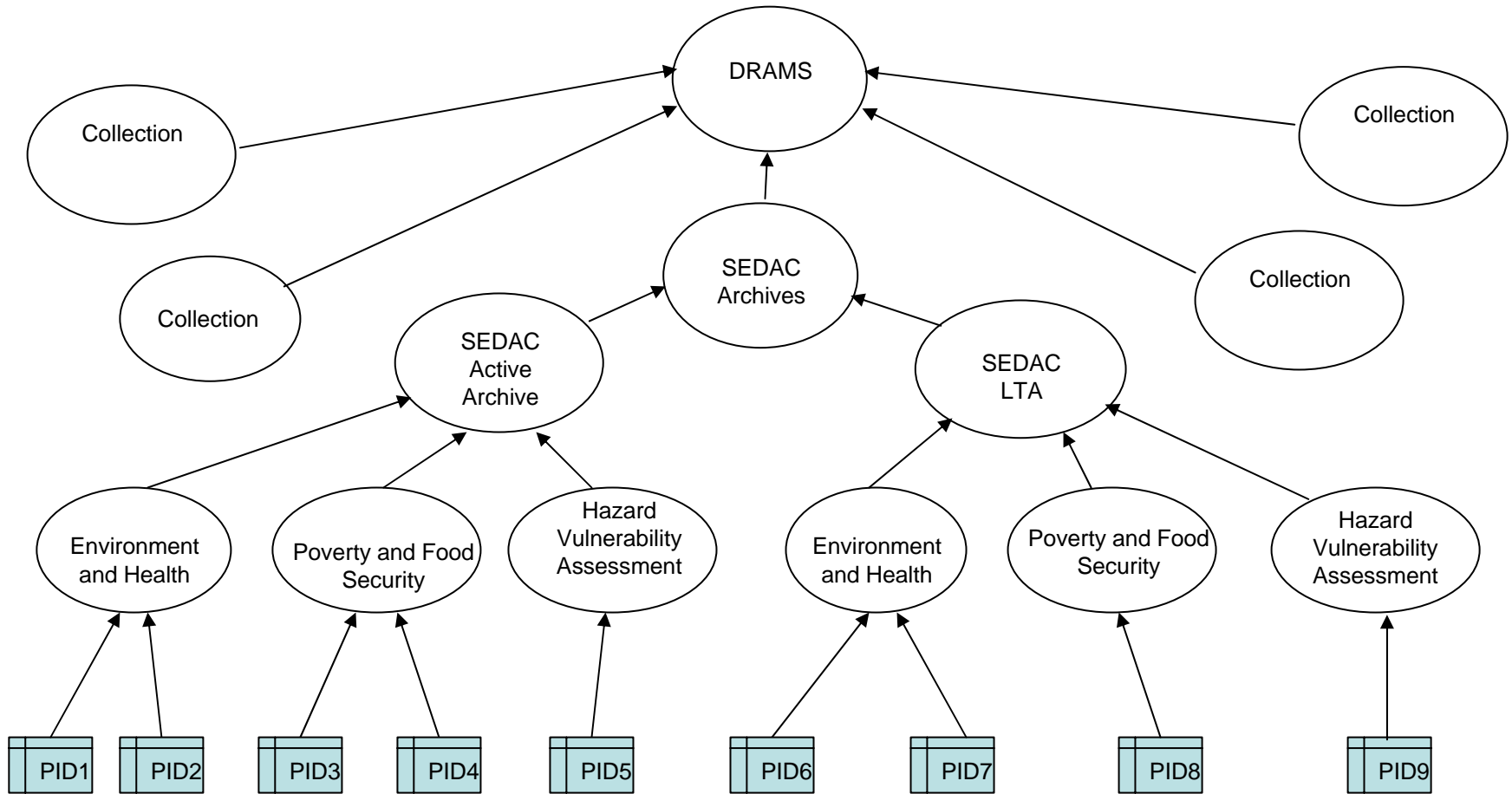


Data authors contribute data and related documentation
Data is reviewed and prepared for ingest
A Persistent Identifier (PID) is assigned by Handles server
Technical metadata is validated using JHOVE server
Digital object is ingested in data repository
Open Archives Initiative (OAI) Harvesters get Metadata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

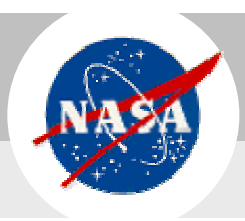




Digital Repository Collections Organization



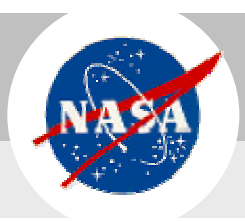
Each data object is assigned a unique Persistent Identifier (PID). Data objects are organized in Multiple collections and sub-collections within the Data Repository and Asset Management System (DRAMS).



Current LTA Infrastructure Initiatives

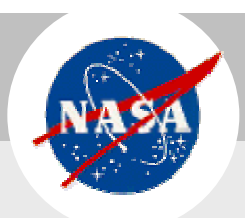
- Install VITAL digital library and asset management software based on Fedora Digital Object Repository Architecture
- Develop Data Repository and Asset Management System (DRAMS)
- Establish Data Preservation and Public Dissemination Services
- Import LTA Data from Fedora digital repository prototype to DRAMS
- Ingest, preserve, and disseminate data when approved for accession





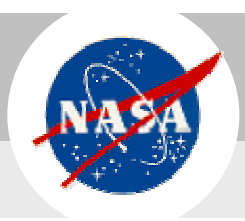
Next Steps for LTA

- Continue strategic planning with CU Libraries, Information Services, and Earth Institute
- Enhance the LTA technical infrastructure
- Disseminate accessioned LTA data
- Explore expansion of LTA to support other CIESIN, 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



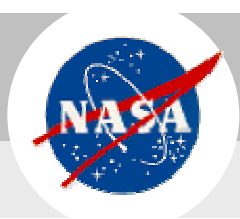
Summary: Benefits of Collaborative LTA

- 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



References

- **National Science Board (2005).** Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century. **National Science Foundation.** <http://www.nsf.gov/pubs/2005/nsb0540/>
- **Reference Model for an Open Archival Information System (OAIS).** Consultative Committee for Space Data Systems. Adopted as: Space data and information transfer systems - Open archival information system - Reference model (ISO 14721:2003). <http://www.ccsds.org/documents/650x0b1.pdf>
- **Producer-Archive Interface Methodology Abstract.** Consultative Committee for Space Data Systems (CCSDS 651.0-R-1) <http://ssdoo.gsfc.nasa.gov/nost/isoas/CCSDS-651.0-R-1-draft.pdf>
- **To Stand the Test of Time: Long-term Curation and Management of Large Data Sets in Science and Engineering (draft).** A report to the National Science Foundation from the Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, 26-27 September 2006, Arlington VA



Web Sites

SEDAC
socioeconomic data and applications center

home help/faq contact us site map

SEDAC Home > Long-Term Archive

The SEDAC Long-Term Archive (LTA) preserves selected SEDAC data and information resources for future access and use. The SEDAC LTA is managed by SEDAC in collaboration with the Columbia University Libraries.

[About SEDAC LTA](#)
[Appraisal](#)
[Documents](#)
[SEDAC LTA Data](#)
[Nominated Data](#)
[Policy Documents Under Review](#)
[Columbia University Libraries](#)
[Related Resources](#)

Low Library, Columbia University

[NASA Science Data User Survey](#)
Need help or information? Contact [SEDAC User Services](#)
[About SEDAC](#) [Acknowledgments](#)

Copyright © 1997-2006
The Trustees of Columbia University in the City of New York
[Privacy Policy and Important Notices](#)

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

CU Home > Libraries Home Search | Site Index | FAQ | Help

COLUMBIA UNIVERSITY LIBRARIES

Libraries & Collections Catalogs E-Resources Request It Ask Us Using the Libraries

Spotlight
Search CLIO, Proquest, Web of Science *simultaneously*

crossSearch
learn more >>

News [Columbia University's Rare Book & Manuscript Library](#) acquires manuscript collection from novelist Oscar Hijuelos ... >>

QUICK SEARCH

CLIO Databases E-Journals Website

Search Library Catalog: [Go To CLIO >>](#)

[crossSearch for books and articles](#) >>

INFORMATION FOR...
Alumni & Friends
Jobseekers
Visitors

FEATURED RESOURCE
Shakespeare Collection
An extensive collection of authoritative materials and the complete Arden editions
[Text-only version of this page](#)

© Columbia University Libraries My Library Account | Hours | Contacts | Suggestions

<http://www.columbia.edu/cu/lweb/>



Columbia University
in the City of New York