World Data Center for Human Interactions in the Environment

Conducting a Self-Assessment of a Long-Term Archive for Interdisciplinary Scientific Data as a Trustworthy Digital Repository

Robert R. Downs and Robert S. Chen
Center for International Earth Science Information Network (CIESIN)
The Earth Institute, Columbia University

Prepared for presentation to the
4th International Conference on Open Repositories (OR2009)
Atlanta, Georgia

May 19, 2009

Evolution of Scientific Practice

- Limited data sharing practices are being replaced by open data sharing practices
- Data management by individual scientists is being supplemented by scientific data repositories
- Data-driven science is becoming more prevalent within scientific communities

Data Centers and Libraries are Supporting Data-Driven Science

- Accessioning and preserving scientific data for use by current and future communities
- Preparing data and establishing services to meet evolving needs of user communities.
- Disseminating data products and services to support current uses and users

Digital Data Presents Challenges for Scientific Data Centers, Libraries, and Government Agencies

- Evolving technologies for obtaining and accessing data
- Impermanent media for storing data
- Evolving user populations and needs for using data and services
- Limiting access to sensitive, private, and protected data

Digital Repositories can help to meet challenges for managing and providing long-term stewardship for collections of scientific data

- Encapsulating data and related information as digital objects
- Assigning persistent identifiers and describing data sets and collections
- Documenting relationships between versions of data
- Enabling adoption of current standards and transformation to new standards

Content of Scientific Data Archives and Repositories Needs to be Trustworthy

- If scientific data cannot be trusted
  - Is the data useful to the scientific community?
  - Can the collection be trusted?
  - Is the repository trustworthy?

Potential Benefits of Ensuring Trustworthiness of Scientific Data Collections of the Digital Repository

- Data producers will want to submit data to the repository that holds trustworthy data
- Users will visit repository to find data
- Scientific, decision-making, and educational communities will use data and services of repository
- Authors will cite data used from repository

Scientific Data Repositories Need to Ensure Current and Future users that Data Can be Trusted

- Provenance information encapsulated with the data
- Clear distinction and relationships between versions
- Inclusion of all data files and documentation
- Data described for use
- References describe publications that used the data
- Clear rights for using the data
- Software and hardware requirements are specified

Data Depositors Need to Know that the Data Repository Can be Trusted to Manage the Data

- Are all the original files archived?
- Will the documentation be kept with the data?
- What level of preservation service is provided?
- Will the data producers receive attribution in any publications of new findings?

Will the Data Be Trustworthy for Future Communities of Users and Their Uses?

- Research
  - Can scientists replicate previous studies and use archived data to conduct longitudinal analysis?
- Decision-Making
  - Can policy makers determine the reliability and relevance of data produced from previous studies?
- Education
  - Can educators assign students to use the data from this repository for their educational activities?

Responsible Organizations Must Ensure that **Interdisciplinary** Data are Preserved for Future Use

- Can the data be easily used by scientists from other fields?
- Can the data be easily integrated with data from other fields?
- Can the data be prepared for use by non-scientists?

Standards for Scientific Data Preservation

- Consultative Committee for Space Data Systems. (2004). **Producer-Archive Interface Methodology Abstract.** (CCSDS 651.0-R-1) http://public.ccsds.org/publications/archive/651x0b1.pdf
- Consultative Committee for Space Data Systems. (2003). **Reference Model for an Open Archival Information System (OAIS).** Adopted as: Space data and information transfer systems - Open archival information system - Reference model (ISO 14721:2003) http://public.ccsds.org/publications/archive/650x0b1.pdf
- Consultative Committee for Space Data Systems. (forthcoming). **Metrics for Digital Repository Audit and Certification**. Draft Recommendation for Space Data System Standards. http://wiki.digitalrepositoryauditandcertification.org
- Online Computer Library Center (OCLC) and Research Libraries Group (RLG). **Data Dictionary for Preservation Metadata: Final Report of the PREMIS Working Group.** May 2005. http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Impetus for Self-Assessment

- Recommended by the SEDAC Long-Term Archive Board
- Identify improvements for repository management, policies, procedures, practices, and capabilities
- Repository development to be guided by standards

Assessment to Guide Digital Repository and Collections Development

- Staged Approach
  - Initial self-assessment of Long-Term Archive collection
  - Continuing assessment of the repository
- Self-Assessment Using TRAC
  - Trustworthy Repositories Audit & Certification: Criteria and Checklist. OCLC and CRL. 2007. http://bibpurl.oclc.org/web/16712
- Assessment is part of plan for Continuous Improvement
  - Sustainability
  - Capabilities
  - Management

Using TRAC to Conduct Self-Assessment of Long-Term Archive Collection

- Compared Each TRAC Requirement to Capabilities
  - Mission, policies, plans, procedures
  - Technological infrastructure and documentation
- Organization of TRAC Document
  - Organizational Infrastructure
    - Organizational aspects of repository sustainability
  - Digital Object Management and Technologies
    - Policies, procedures, and capabilities to manage objects
  - Technical Infrastructure and Security
    - Management of systems and standards

Benefits of Conducting Initial Self-Assessment of LTA Collection

- Identified relevant policies and procedures
- Improved policies, procedures, and practices
- Began preparing for self-assessment of repository capabilities for managing multiple collections
- Recommendations for initiatives to improve archival trustworthiness

Recommended Initiatives to Improve Archival Trustworthiness

1. Strategy for collaborative organizational sustainability
2. Model for supporting submission of scientific data to the repository
3. Plan for facilitating intra-organizational transfer of data between collaborating repositories

1.
Strategy for Collaborative Organizational Sustainability

Need: Organizational Sustainability for a Trustworthy Repository

- Projects and short-term funding cannot ensure long-term data stewardship
- Sustainable entity is needed to accept responsibility for managing and operating the repository in the future.
- Organizational capabilities are needed to manage a repository and its content in the future.
- Institutional commitment to preserve scientific data is needed for allocation of future resources
- Past commitments and record for preserving knowledge provide an indication of future performance.

Approach: Strategy for Collaborative Organizational Sustainability

- An organizational strategy relies on future organizational capabilities for long-term care of organizational assets
- The organization can consider the strategy along with other strategic plans and organizational initiatives, both competing and complementary
- Strategy drives the development of long-term, medium-term, and short-term plans to meet organizational objectives
- Organizational sustainability is independent of technological infrastructure
- Collaboration between Columbia University Libraries, the Earth Institute of Columbia University, and the NASA Socioeconomic Data and Applications Center

Overview of Strategy for Collaborative Organizational Sustainability

- Scope:
  - Long-Term Archive Collection
- Governance:
  - Long-Term Archive Board
- Management and Staff:
  - Systems development and archival operations
- Contingency Plan:
  - If Funding lapses

SEDAC Long-Term Archive Mission Statement

"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."

Source: SEDAC Long-Term Archive Implementation Plan (Draft revised 2008)

LTA Governance and Management

- LTA Board
  - Approves the Selection Criteria for Submission of SEDAC Data to the LTA
  - Approves the submission of recommended data or information resources that meet the LTA selection criteria
- LTA Manager
  - Reports to the LTA Board
  - Responsible for development and operations of LTA systems, staff and procedures, to ensure the stewardship of LTA holdings
- LTA Staff
  - Report to the LTA Manager
  - Accession and maintain LTA holdings in accordance with LTA procedures
  - SEDAC Engineering staff support infrastructure development and maintenance

Adapted from SEDAC Long-Term Archive Implementation Plan (Draft revised 2008)

Contingency Plan for the SEDAC Long-Term Archive Governance and Management

- Current LTA Board membership:
  - represented by the SEDAC Project Scientist, the SEDAC Systems Engineer, and the SEDAC LTA Manager,
  - two representatives designated by the Earth Institute, and
  - two representatives designated by the Columbia University Library.
- In the event of a lapse in SEDAC funding:
  - Libraries will replace chair and one of the SEDAC members
  - CIESIN will name the other SEDAC member

  => Columbia University Libraries will have majority of members
  - Columbia University will appoint the Long-Term Archive Manager and other staff as needed

Adapted from SEDAC Long-Term Archive Implementation Plan (Draft revised 2008)

2.
Model for Submission of Scientific Data to the Repository

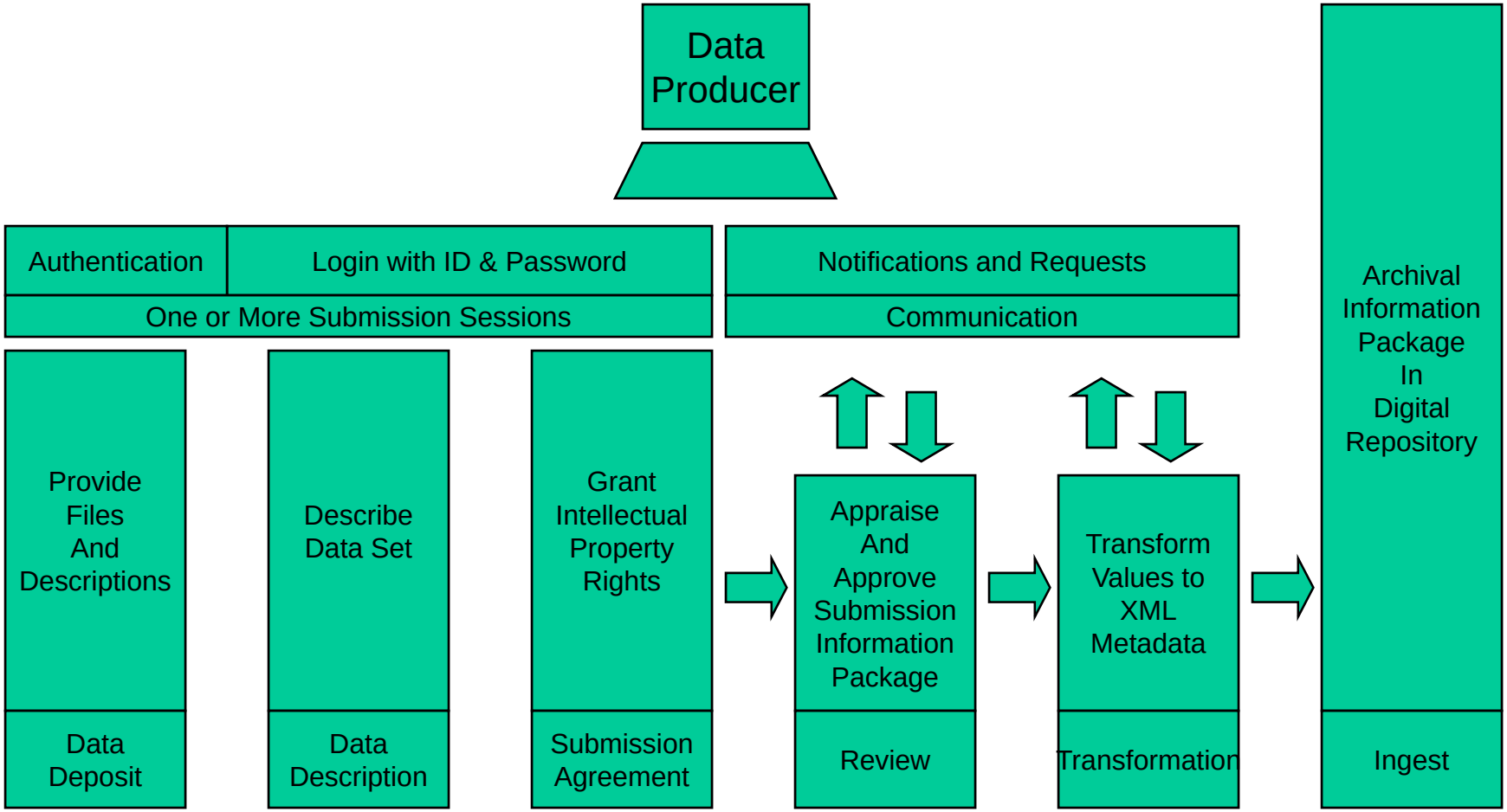
Need: Data Submission Capabilities

- Encourage data submission
  - Help data providers to deposit and describe data
- Enable acquisition
  - Opportunity to receive data and information
- Capture provenance information
  - Moving upstream to the original source
- Facilitate workflow
  - Systematically control pre-ingest processes

Approach: Model for Submission of Scientific Data to Repository

- Reflects requirements for a trustworthy repository described in TRAC document
- Visually represents submission and workflow interactions for pre-ingest services
- Informs design, development, and evaluation of the submission system

Model for Web-Based Data Submission and Workflow



Source: Downs and Chen. Creating a Trustworthy Digital Repository for a Long-Term Archive of Interdisciplinary Data: A Case Study. 21st International CODATA Conference, 5-8 October, 2008 Kyiv, Ukraine.

3.
Plan for Facilitating Intra-Organizational Transfer of Data Between Collaborating Repositories

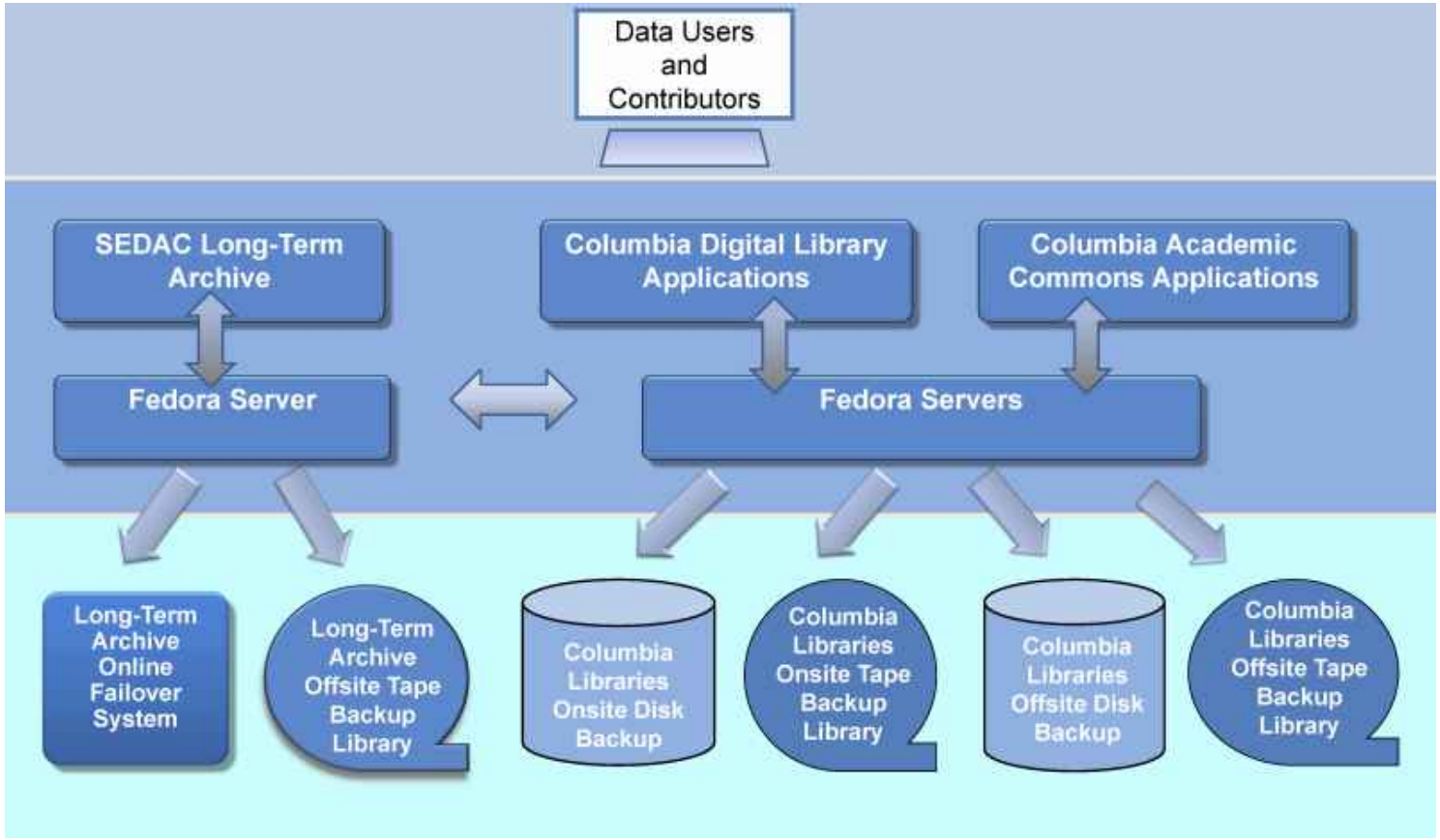
Need: Data Transfer Between Repositories

- Reduce risk associated with any single repository
- Independent systems have their own strengths and vulnerabilities
- Transfer between two repositories precedes possibility of future transfers between other repositories
- Provide capabilities for contingencies in LTA strategy
- Facilitate evolution of cyberinfrastructure

Approach: Plan to Facilitate Intra-Organizational Transfer of Data Between Repositories

- Establish parallel storage and processing capacity
  - Independently operated hardware and operating systems
- Operate repositories for hosting data
  - Operate Fedora repositories in both locations
- Conduct tests to transfer objects and collection
  - Transfer data objects and datastreams between repositories
  - Transfer LTA collection between repositories

Planned Intra-Organizational Transfer Capabilities



Adapted from: Downs, Chen, Cartolano, Bose (2008)/ Collaborative Establishment of a Long-Term Archive for Stewardship of Interdisciplinary Scientific Data. 2008 Fall AGU Meeting, San Francisco, CA, December 15-19, 2008. Eos Trans. AGU, 89(52), 2008 Fall Meet. Supplement, Paper Number: U23A-0047

Benefits of Planned Capabilities for Intra-Organizational Transfer

- Identify improvements to reflect changes in requirements and available resources
- Share knowledge about cyberinfrastructure development
- Share practices for data management and stewardship

Future Assessment

- Improve policies, plans, procedures, and capabilities for a trustworthy digital repository of interdisciplinary scientific data
- Repository self-assessment for managing multiple data collections using the Metrics for Digital Repository Audit and Certification (Draft Recommendation for Space Data System Standards).
- Repository audit using the Metrics for Digital Repository Audit and Certification (Draft Recommendation for Space Data System Standards).

The authors gratefully acknowledge support received from NASA for the operation of SEDAC under contract NNG08-HZ11C.