CIESIN
Columbia University

The Earth Institute
AT COLUMBIA UNIVERSITY

WDC
World Data Center for Human Interactions in the Environment

Managing Risks to Scientific Data

Robert R. Downs, PhD
Center for International Earth Science Information Network (CIESIN),
Columbia University

Prepared for Presentation to the
NYU/IBM Workshop on
Managing Data Risk:
Acquisition, Processing, Retention and Governance

New York University
New York, NY
April 24, 2009

CIESIN
Columbia University

The Earth Institute
AT COLUMBIA UNIVERSITY

WDC
World Data Center for Human Interactions in the Environment

The Center for International Earth Science Information Network (CIESIN): An Interdisciplinary Scientific Data Center Dedicated to Reducing Risks to Scientific Data

- Creating, acquiring, archiving, developing, and disseminating interdisciplinary scientific data products to improve understanding on human interactions in the environment.
- Conducting research on scientific data management, data stewardship, and digital preservation
- Collaborating with the data science community to continually improve data management practices.
- Recommending data management and digital preservation practices to data managers and science teams.

Observed Risks to Scientific Data

- Various risks to data were observed in a study of data management practices within organizations that use geospatial data and related electronic records.
- Lessons for Data Management
  - Risks to scientific data must be managed to ensure that the data can be used for the purposes for which they were created.
  - Not addressing the risks to scientific data can reduce the potential for using the data, now or in the future, and can pose additional risks for stakeholders involved in the creation, dissemination, stewardship, and use of the data.
- Scenarios of risks to data management characterize risks and offer techniques for mitigating such risks.

Identifying Risks to Scientific Data

- Legality: Has permission been obtained for the intended use of the data?
- Integrity: Are the data complete and correct?
- Interpretability: Can the data be understood by the intended audience?
- Accessibility: Are the data viewable with hardware and software available to the intended audience?
- Discoverability: Are the data discoverable by the intended audience?
- Security: Are the data protected from loss, theft, or tampering?
- Confidentiality: Are confidential aspects of the data identified and protected from unauthorized access?
- Recoverability: Will the data be available in nn years or if a disaster occurs?
- Sustainability: Can the organization address the risks to the data in the future?

Scenarios of Risks to Scientific Data: Legality

- Data and permissions are obtained for use in an analytical project.
- Years later, the data are recognized as being valuable for another project that would build on the previous results, but permissions for the new use cannot be obtained due to changes in the policies of the data provider.
- Risk Mitigation:
  - Obtaining permissions during data acquisition for a broad set of potential uses would reduce the need to renegotiate permissions for any new uses.
  - Encouraging and using open licenses and permissions statements that are attached to data and data products will reduce confusion about intellectual property rights.

Scenarios of Risks to Scientific Data: Integrity

- A data set is obtained and the files are stored in a directory used by various projects.
- Years later, users of the data set report that files appear to be missing.
- Data Risk Mitigation:
  - A master copy of the data set would enable restoration of the original data.
  - An inventory of the data set, including checksums, would enable verification of data integrity.

Scenarios of Risks to Scientific Data: Interpretability

- A data set is created without any documentation and is used by the creator to produce various results.
- After the creator has left the organization, the data cannot be used for a new longitudinal analysis since the variables and assumptions are not described.
- Data Risk Mitigation:
  - Provide guidance describing procedures for staff to follow for the documentation of data and data products.
  - Conduct reviews of data products produced to ensure that the data are properly documented to enable subsequent use.

Scenarios of Risks to Scientific Data: Accessibility

- Data are created using proprietary software and stored on currently available portable media.
- Years later, when the data are needed, no drive can read the media. After the data are migrated to current media, the data cannot be rendered by available software.
- Data Risk Mitigation:
  - Establish a media migration plan and a data conversion plan based on an assessment of the future need for the data.
  - Establish a technology watch to identify old and new media and data formats.

Scenarios of Risks to Scientific Data: Discoverability

- After obtaining a free data set from a data provider to produce maps, the map creator does not store the data or record the title of the data set that was used to create the maps.
- Years later, when the map creator requests the original data from the data provider, the data cannot be identified as a result of the availability of several newer versions representing the same location. Without knowing the exact title of the data set and the date when the data were collected, the original data cannot be determined.
- Data Risk Mitigation:
  - Archive the data that have been used to create data products.
  - Document data products with information describing source data used in their creation.

Scenarios of Risks to Scientific Data: Security

- After creation, data are written to tapes stored on an available shelf near the tape drives.
- Later, the tapes are mistakenly rewritten since they are improperly labeled and located near other tapes that have been designated for reuse.
- Data Risk Mitigation:
  - Establish safe locations (offsite and onsite) for long-term storage of any media containing data.
  - Establish procedures for labeling media used to store data.

Scenarios of Risks to Scientific Data: Confidentiality

- Data are created that identify locations of vulnerable wildlife populations
- The data are accessed by local groups to identify hunting locations.
- Data Risk Mitigation:
  - Protect data containing confidential information by determining who may access the data and only allowing designated users to have access.

Examples of Risks to Scientific Data: Sustainability

- Data is created as part of a short-term project, which submits final reports and reassigns personnel upon completion of the project.
- Later, the data are recognized as a resource for a new project, but are no longer available.
- Data Risk Mitigation:
  - Include plan for preparing and submitting data and related products to an archive as part of the project activities.
  - Ensure that the data are prepared and submitted for archiving during the project.