Data Stewardship and Reuse

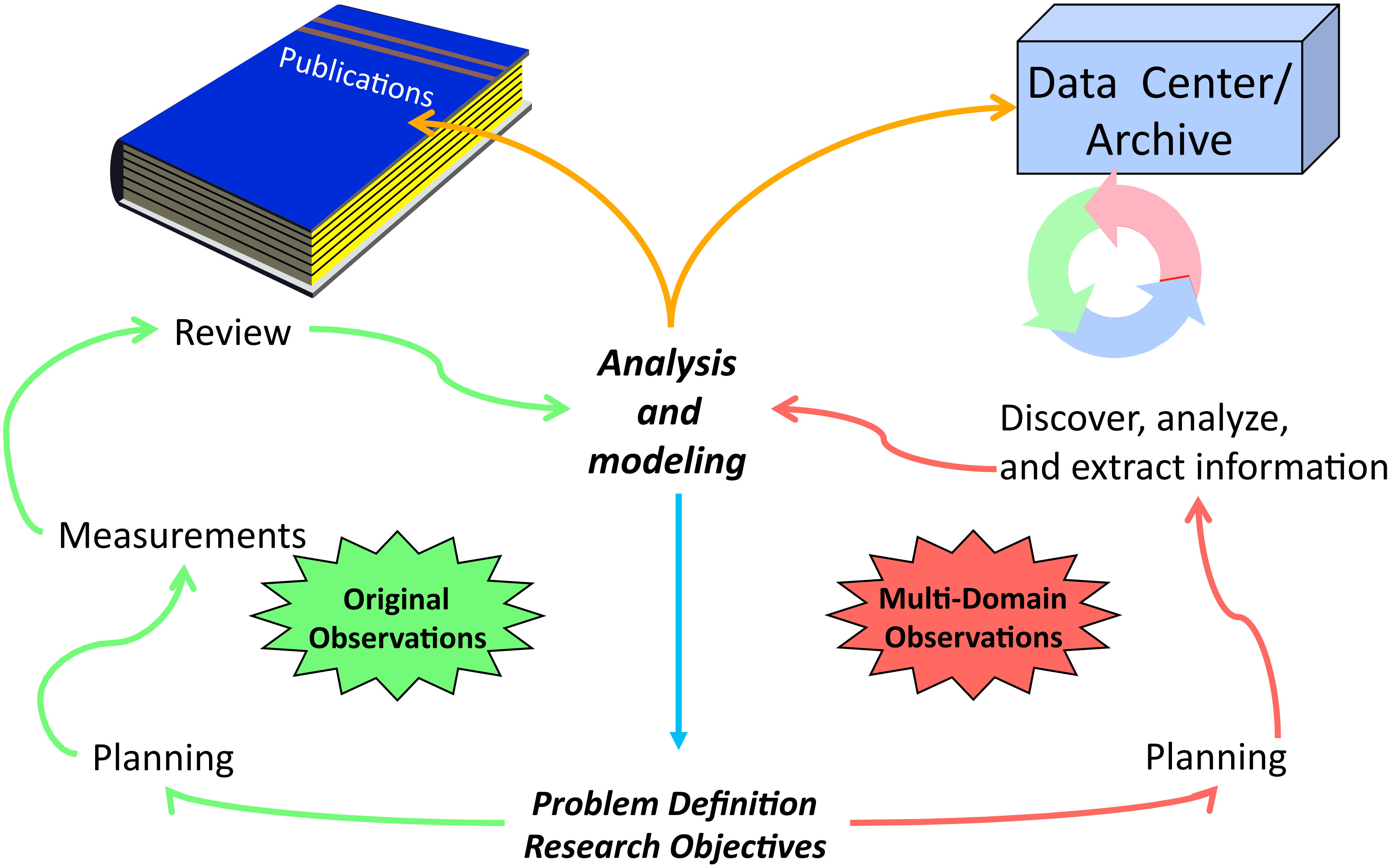
Bob Cook
Environmental Sciences Division
Oak Ridge National Laboratory
Email: cookrb@ornl.gov

Credits: Tauntingpanda, Anita363, Stonebird, NeilsPhotography, Rick Smit, Jschinker

Topics

- Introduction
- Protection
- Archiving
- Sharing & Reuse

Cycles of Research – An Information View



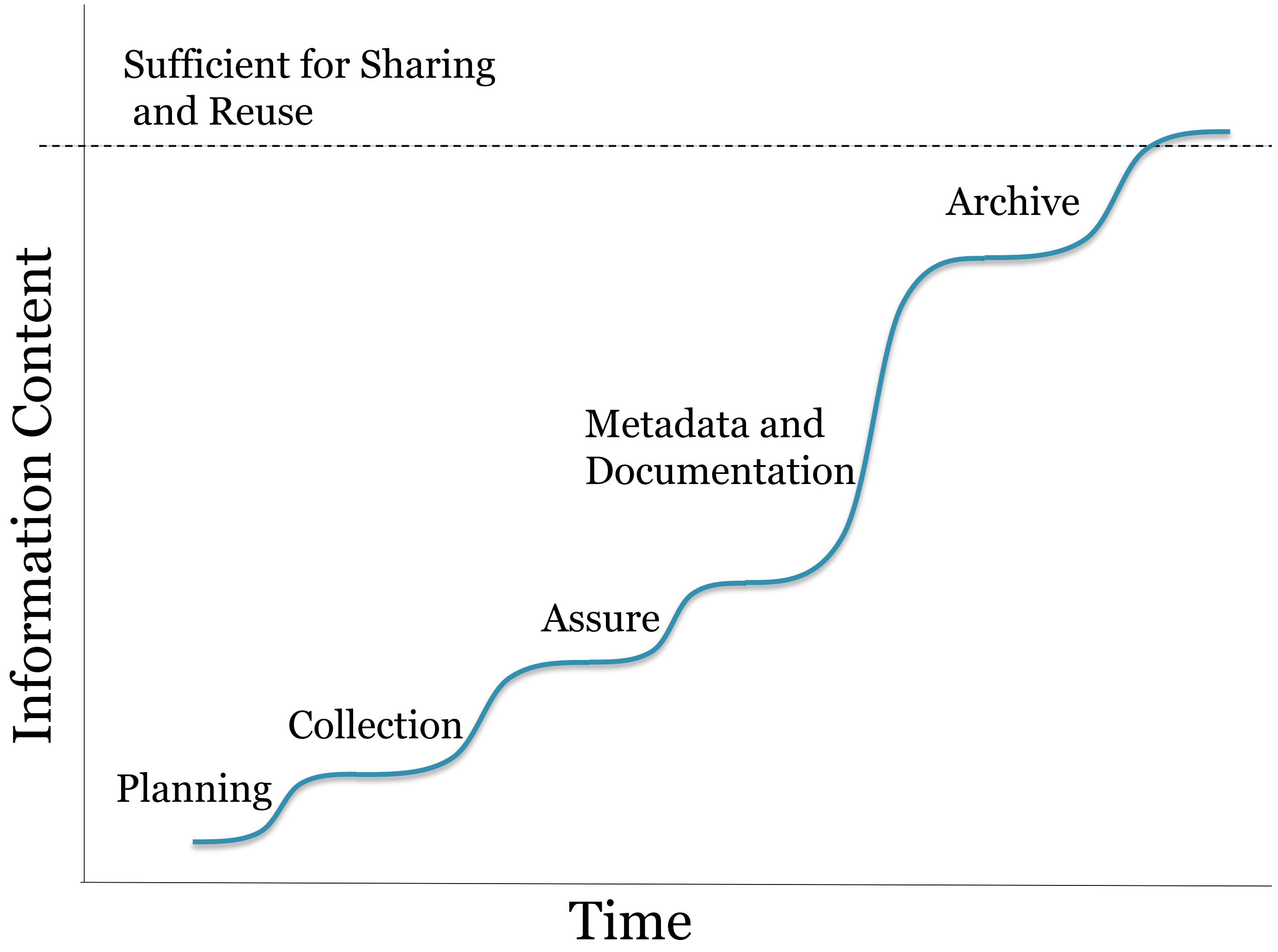
Courtesy of Raymond McCord, ORNL

The 20-Year Rule (NRC 1991)

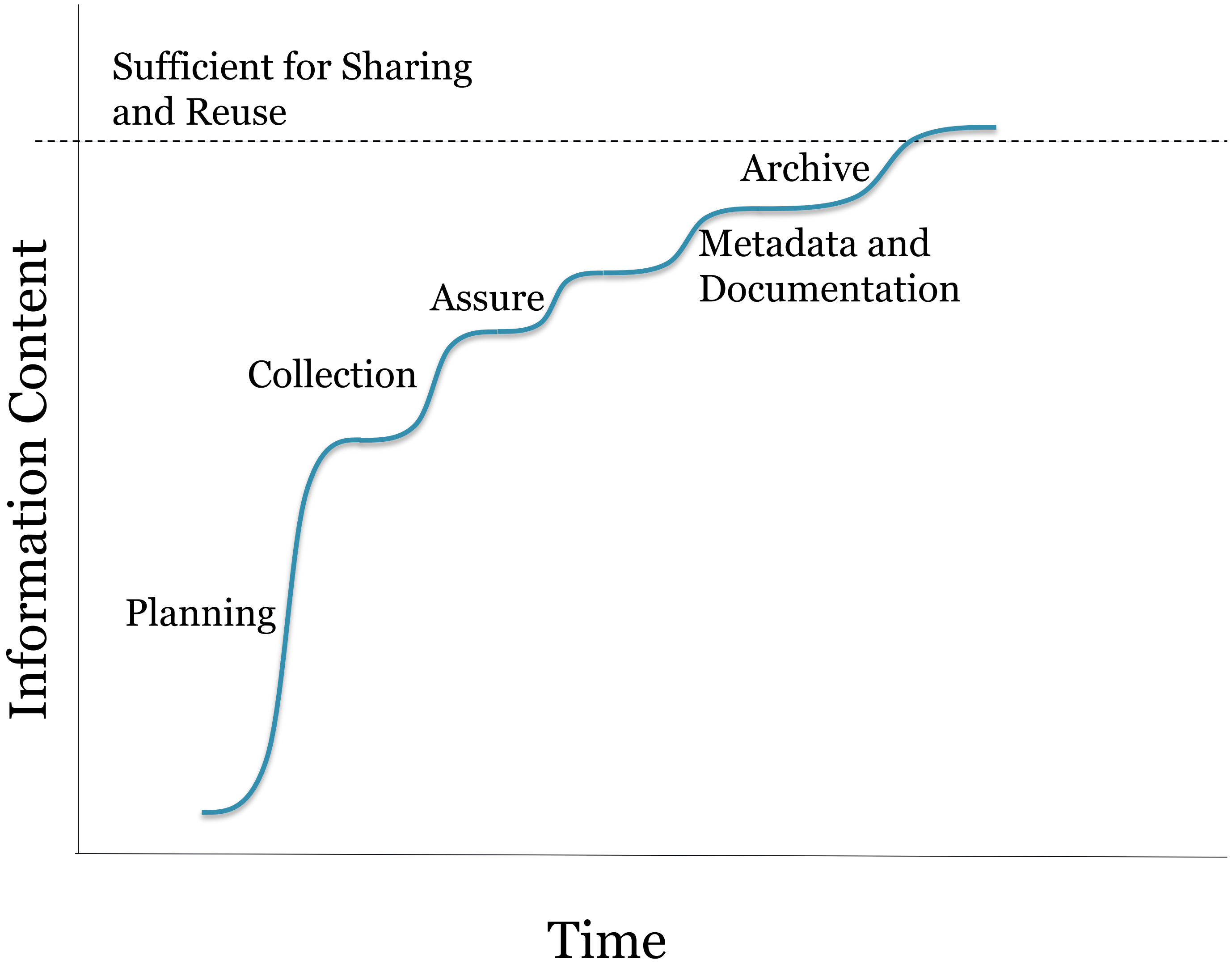
The metadata accompanying a data set should be written for a user 20 years into the future--*what does that investigator need to know to use the data?*

Prepare the data and metadata / documentation for a user who is unfamiliar with the details of your project, methods, and observations

Proper Curation Enables Data Reuse



Proper Curation Enables Data Reuse



Topics

Introduction

Protection

Archiving

Sharing & Reuse

Data Protection: Backups

Create back-up copies

- Ideally three copies

  original, one on-site (external), and one off-site

  (e.g., Dropbox, Carbonite, etc.)

- Frequency based on need / risk

Know that you can recover from a data loss

- Periodically test your ability to restore information

Data Protection: File transfers

Ensure that file transfers are without error

- Compare checksums before and after transfers

  Example tools to generate checksums

  http://www.pc-tools.net/win32/md5sums/

  http://corz.org/windows/software/checksum/

Topics

Introduction

Protection

Archiving

Sharing & Reuse

Data Center:
Stewardship and Archive Functions

- Acquisition
  - identify how best to serve the scientific community
  - establish how and when to receive data
- Ingest
  - perform QA checks
  - compile project-provided metadata
  - convert to archivable file formats
- Enhance (as requested)
  - convert to standard formats & units
  - aggregate files
- Metadata / Documentation
  - Prepare final metadata record and documentation
- Archive / Publish
  - generate citation
- Exploration and Distribution
  - provide tools to explore, access, and extract data for users worldwide
- Post-Project Data Support
  - serve as a buffer between end users and PIs
  - provide usage statistics
- Stewardship
  - provide long-term secure archiving of the data
  - security, disaster recovery
  - migration to new computer systems

Choosing a Data Archive

Institution vs. science discipline archive

- Keep discipline data together
- Resources ($)

Functionality

- Discovery and access
- Specialized data types (geospatial data, genetic sequences, etc.)

Requirements

- Data center's and project's requirements

Topics

- Introduction
- Protection
- Archiving
- Sharing & Reuse

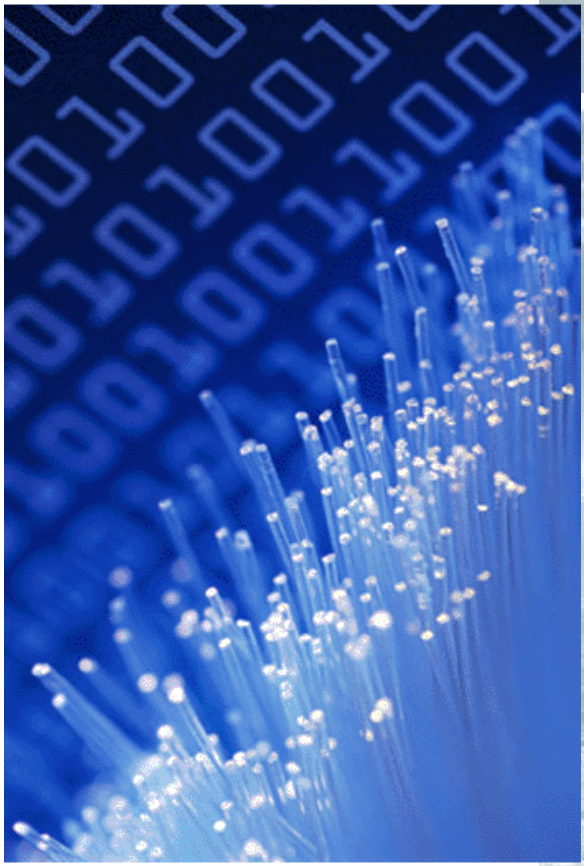
Data Sharing & Reuse: Policies

US Funding agencies: Open Access

- NASA: no period of exclusive use
- NSF: reasonable time, charge user ≤ marginal cost of providing data
- NOAA: short period of exclusive use for QC/QA

Institution's Policy

- Intellectual property



Data Sharing & Reuse: Restrictions

Protection policies and procedures for legitimate / appropriate needs based on data type

- privacy,
- confidentiality,
- intellectual property, or
- other security needs

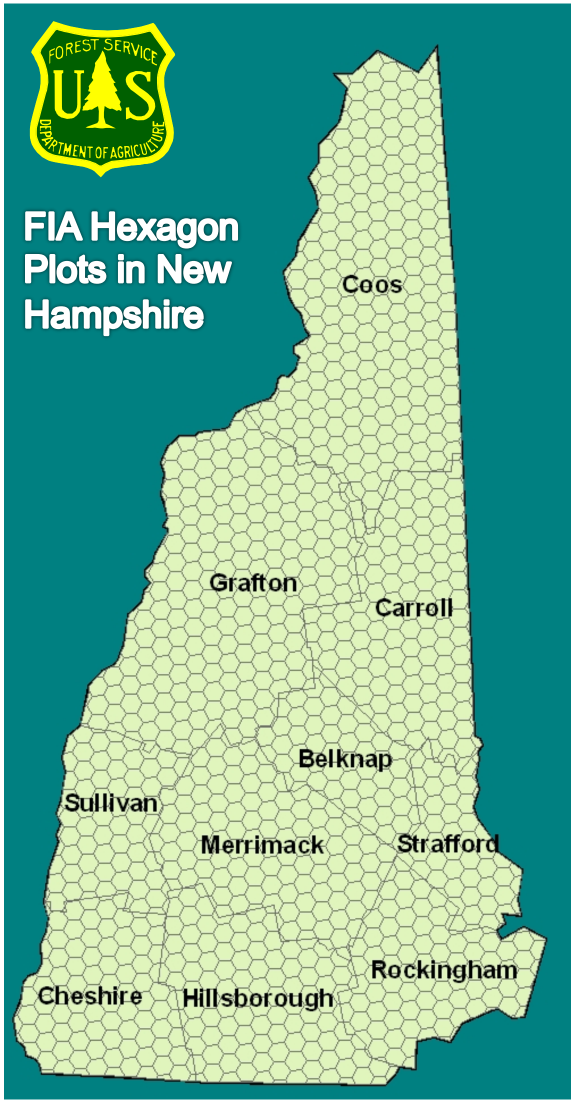
Examples of Restricted Information

US Forest Service Forest Inventory and Analysis data

- Specific location of forest sample plot within each hexagon is restricted

Threatened and Endangered species data

Personally identifiable information



Data Sharing & Reuse: Citation

- Practice analogous to journal article citations
- Enable readers to find data products themselves
  - Reproduce the results
  - Use data for new hypotheses, constructing or evaluating models
- Add to data author's CV
  - Citation indices for the data publication
- Data authors get credit for the data publication and subsequent citations
- Can be used to show funders the impact of their research programs on the advancement of science
- Shows the scientific impact of data centers' data holdings

Data Sharing & Reuse: Citation (cont)

Elements of a data product citation:

- Authors
- Year of publication
- Data product title
- Data center
- Persistent Identifiers
- Date accessed / version number

Examples:

Sidlauskas, B. 2007. Data from: Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: a case study from characiform fishes. Dryad Digital Repository. doi:10.5061/dryad.20

Turner, D.P., W.D. Ritts, and M. Gregory. 2006. BigFoot NPP Surfaces for North and South American Sites, 2002-2004. Data set. Available on-line [http://daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/750.

Benefits of Good Data Management Practices

Short-term

- Spend less time doing data management and more time doing research
- Easier to prepare and use data for yourself
- Collaborators can readily understand and use data files

Long-term (data publication)

- Scientists outside your project can find, understand, and use your data to address broad questions
- You get credit for archived data products and their use in other papers
- Sponsors protect their investment

Questions?