Information International Associates (IIa)

Scientific Data: Increasing Transparency and Reducing the Grey

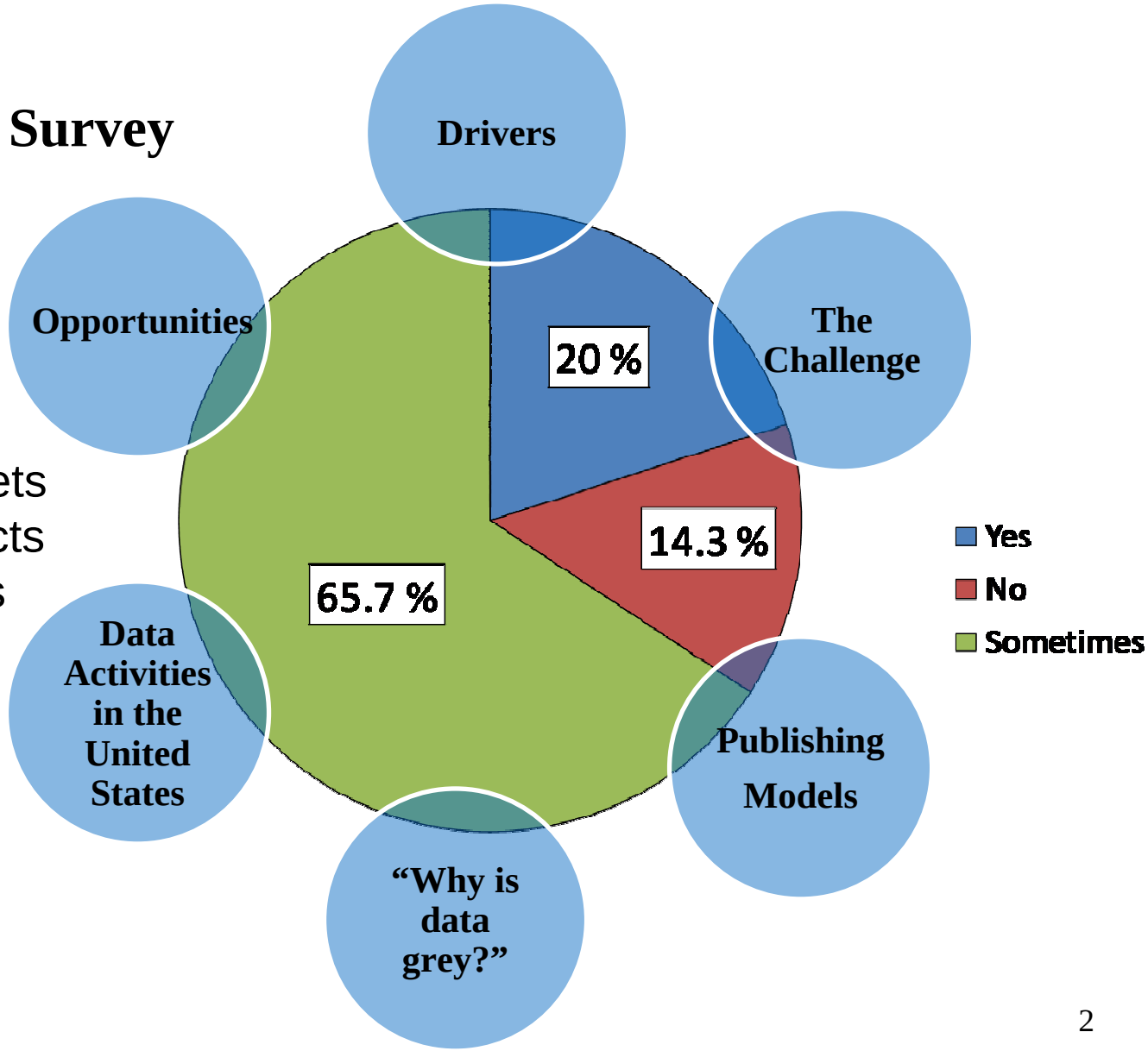
Bonnie C. Carroll, President
June Crowe, Technical Director, Intelligence Division

December 7, 2010

Scientific Data Landscape

CENDI Workshop Survey

% with Sufficient Knowledge of Data Sets When Planning Projects & Research Programs

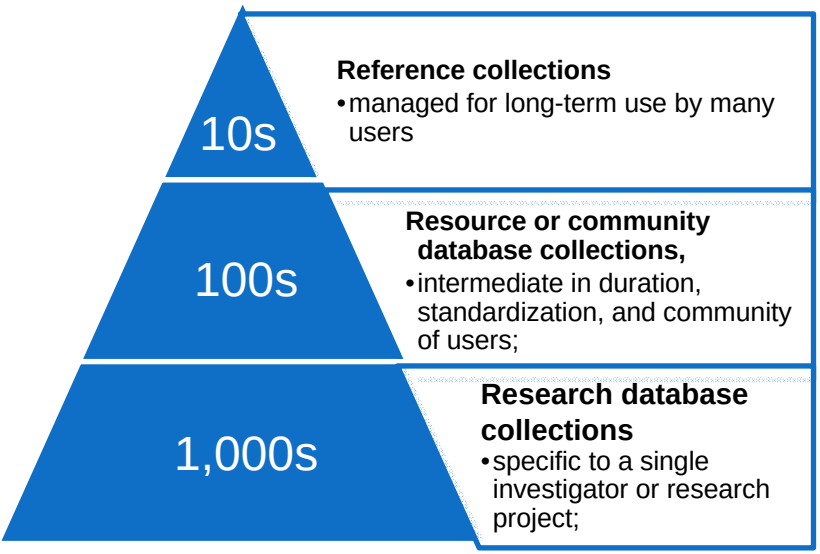


The Drivers: Why Scientific Data?

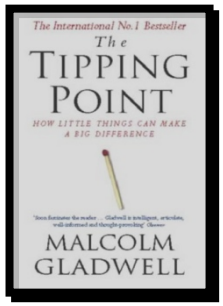
Empowered by an array of new digital technologies, science in the 21st century will be conducted in a fully digital world. In this world, the power of digital information to catalyze progress is limited only by the power of the human mind. **Data are not consumed by the ideas and innovations they spark but are an endless fuel for creativity.** A few bits, well found, can drive a giant leap of creativity. The power of a data set is amplified by ingenuity through applications unimagined by the authors and distant from the original field.

- Technology is enabling -- Push
- Data Intensive Science is the future -- Pull

What's the Challenge?
Data Complexity: Heterogeneity and Volume

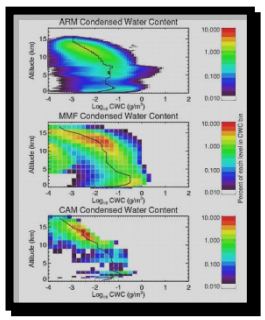


Kilo	**10^3**
Mega	**10^6**
Giga	**10^9**
Tera	**10^{12}**
Peta	**10^{15}**
Exa	**10^{18}**
Yatta	**10^{24}**



1 small novel = **1 MegaByte**

Printed materials in the Library of Congress = **10 TeraBytes**



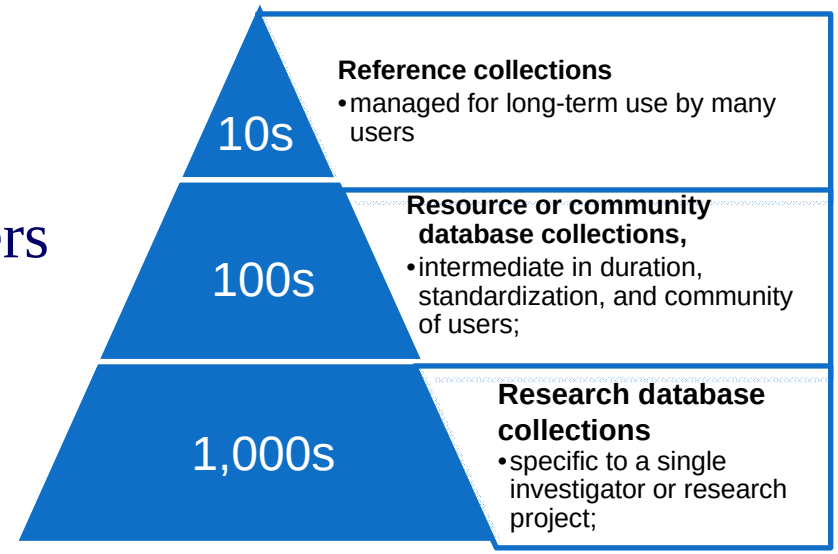
Atmospheric Radiation Measurement Program (ARM) Data Archive = **41 TeraBytes**

Large Hadon Collider (LHC) = **15 PetaBytes annually**

All worldwide information in one year = **5 ExaBytes**

Who are the Publishers of Scientific Data?

- Traditional Publishers
  - Professional Society (ESA, OSA, IUCr*)
  - Commercial (ProQuest – Deep Indexing)
- Repositories / clearinghouses / data archives
  - Dryad*
  - DataNet
- Information Analysis Centers
  - ORNL*
- Research Centers and Researchers
- Metadata Clearinghouses
  - Data Explorer*
  - Mercury*
  - Data.gov



** More to follow*

Key Thrusts that Increase Transparency

Top Down

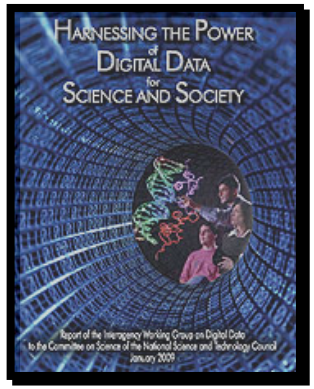
- Developing Data Policy
- Data Management Planning

Bottoms Up

- Data Citation
  - Communities of Practice (Crystalography, Ecology)
  - CODATA
- Persistent Identification of Data Sets
  - Data Cite

From the Side: Discovery tools are developing

- Semantic Web



Metadata
•Standards
•COPs

Interactive Science Publishing: A Joint OSA-NLM Project

- To evaluate the educational value of ISP used within actual scholarly journal articles
- To explore the problems of archiving this medium
- To develop an interactive software and curated database infrastructure “Interactive Science Publishing”
- To give authors the ability to submit their own databases and ISP-enables figures in actual peer-reviewed journal articles
- To give readers and editors the ability to view, analyze, and interact with source data published in conjunction with an article

through the left main bronchus and into the distal section of the trachea, acquiring a 3D scan of the airway lumen. As shown in the axial view of Fig. 3, the *a*OCT scan enabled quantification of the lumen diameters at the time of the bronchoscopy.

A strong correlation was observed between CT and *a*OCT estimates of airway lumen diameters. A representative site in the proximal left main bronchus was selected for the purposes of illustration, with the same anatomical site visually identified for comparison. Using CT, the airway diameter was estimated to be 17.8mm x 14.1mm (Fig. 2). In the *a*OCT scan, the diameter was measured as 17.3mm x 13.9mm. Note that with the CT scan, we have used the oblique (not axial) view, so as to orient the measurement perpendicular to the central axis of the airway.

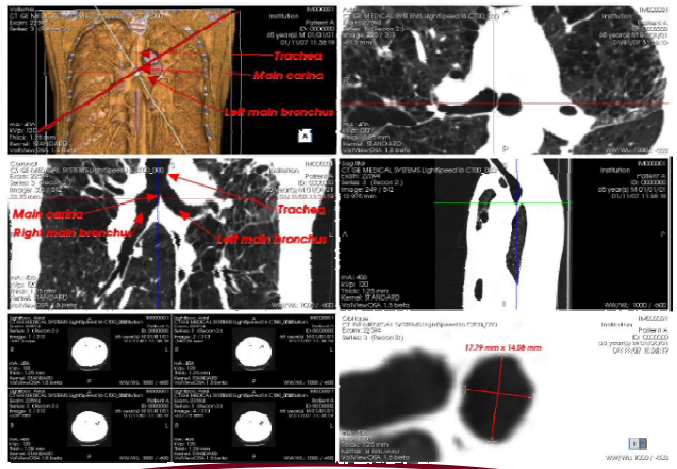


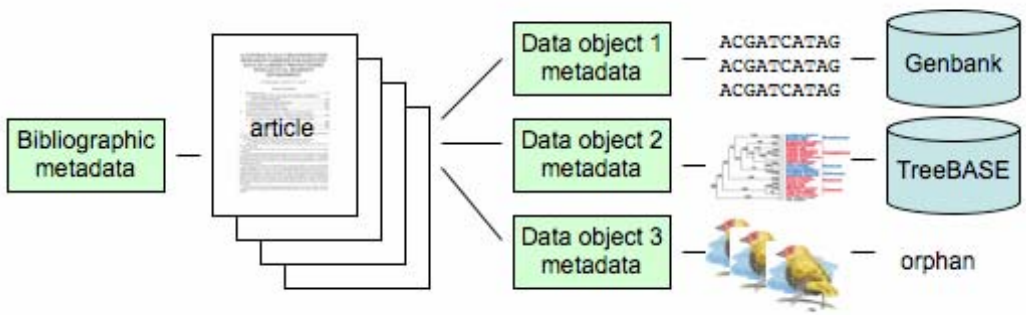
Fig. 2. Patient A. Chest CT depicting the lower airway (View 1). Top row (L-R): 3D view; Axial slice at the level of the main carina. Middle row (L-R): Coronal view; Sagittal view. Bottom row (L-R): Lightbox view; Oblique view measuring airway diameter.

Dryad's Repository

- What are the goals?
  - Preserve underlying data in a paper at time of publication.
  - Lower the burden of data sharing with a one-stop data-deposition via handshaking with specialized repositories.
  - Assign globally unique identifiers to datasets, thus enabling data citations.
  - Allow end-users to perform sophisticated searches over data.
  - Allow journals/societies to pool resources for one repository.
  - Enable bidirectional search and retrieval with data repositories from related disciplines.

Partner Journals

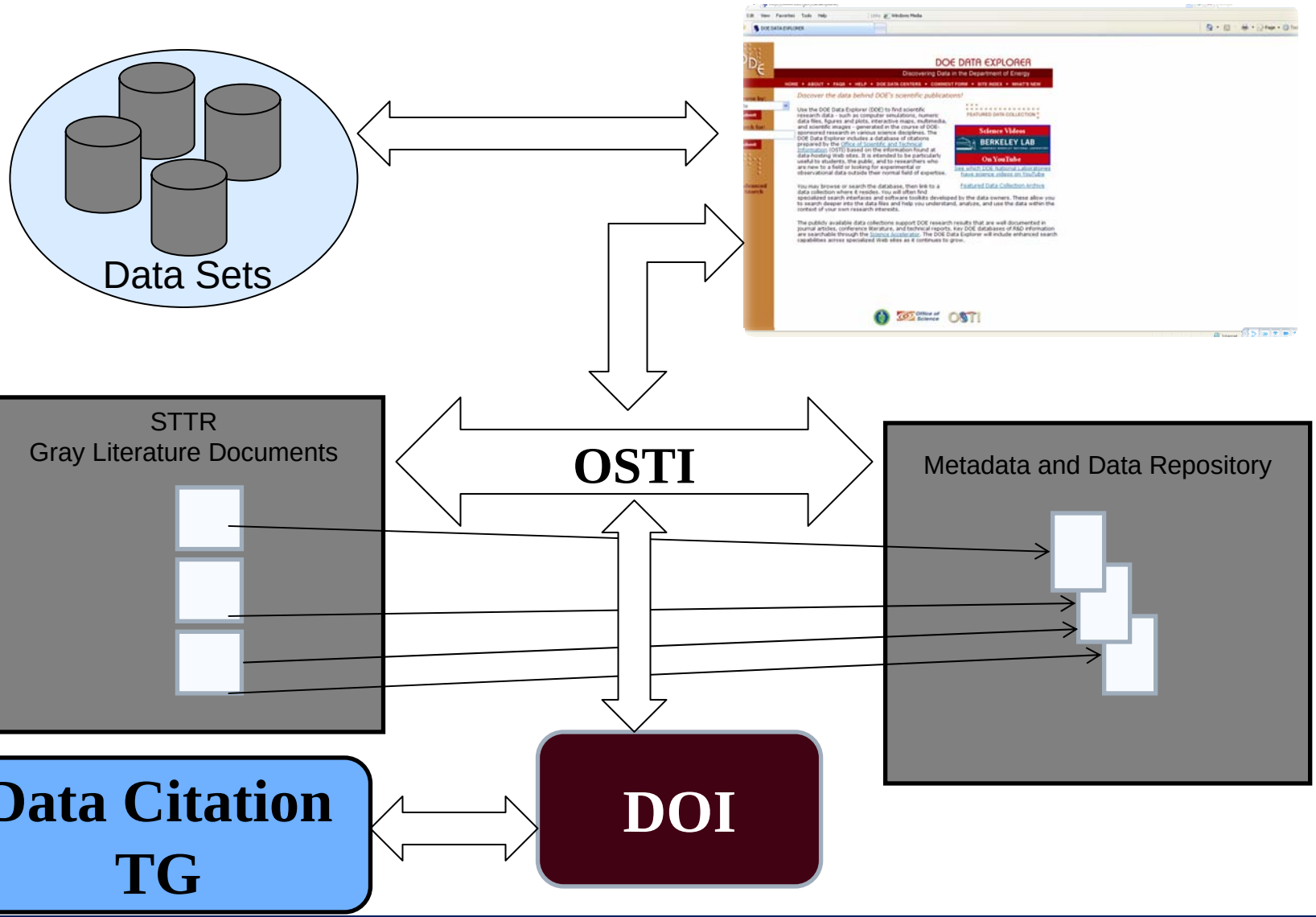
- Ecology
- Paleontology
- Population genetics
- Physiology
- Systematics
- Genomics



National Science Foundation
NESCent
UNC
&

North Carolina State University, University of New Mexico/LTER, Yale University, + partner journals and societies

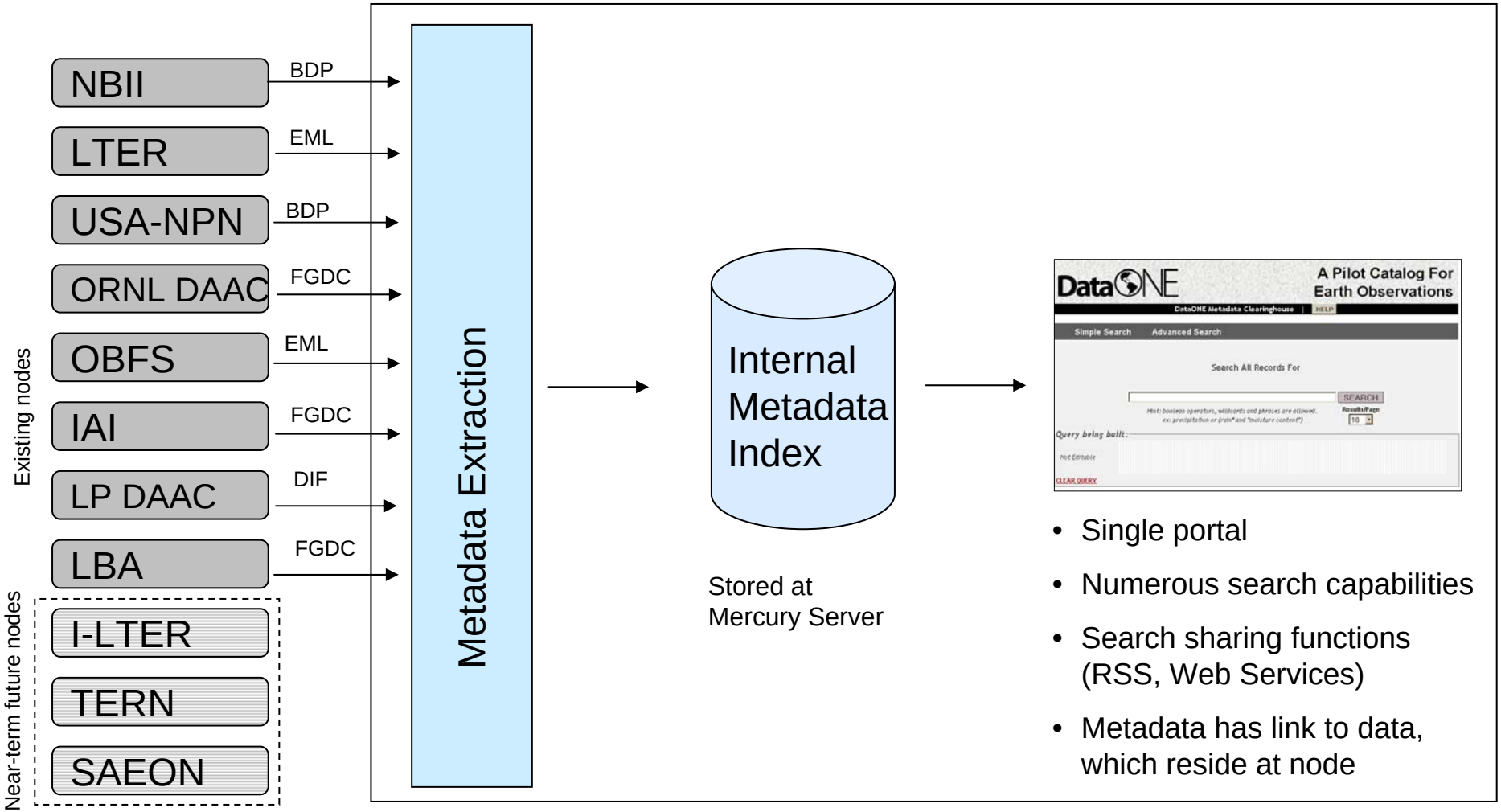
A Department of Energy Approach



Interagency Case Study
Oak Ridge National Laboratory

- Major Center for Environmental Scientific Data Management
  - responsible for archiving, managing, and distributing data
  - for enabling the distribution, use, and analysis of this data.
- Data Repositories
  - Atmospheric Radiation Measurement Archive
  - Carbon Dioxide Information and Analysis Center
  - ORNL Distributed Active Archive Center
- Mercury Metadata Repository

Mercury Metadata Clearinghouse Architecture

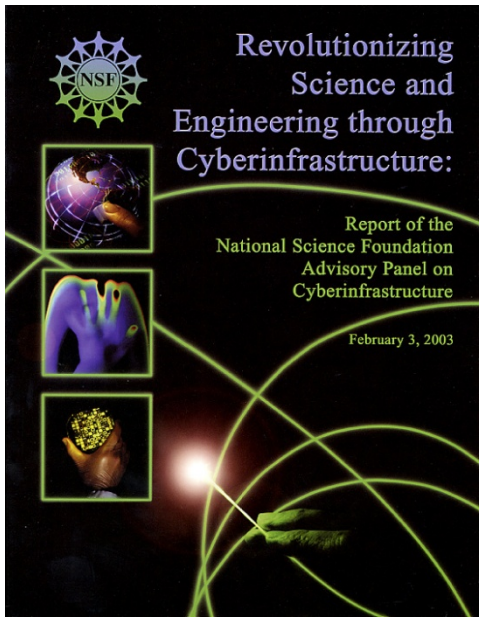


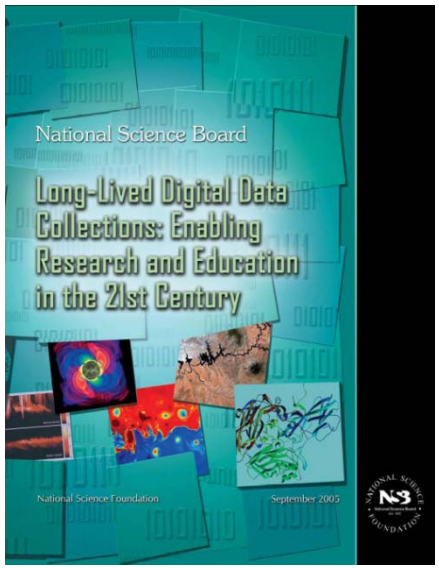
- Single portal
- Numerous search capabilities
- Search sharing functions (RSS, Web Services)
- Metadata has link to data, which reside at node

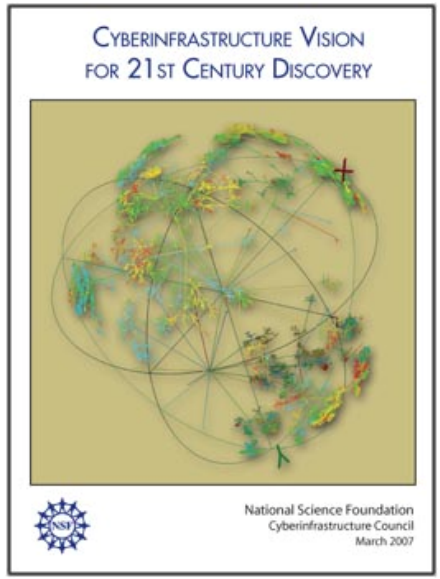
Making Data Sets More Transparent

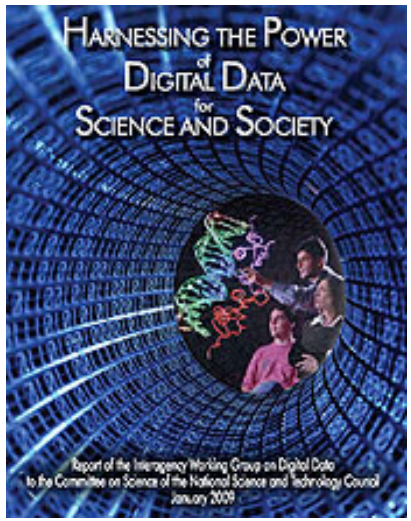
- Policy, Culture and Management
  - National Policy -government taxpayers funded projects should be accessible
  - Enhanced metadata
  - Journals supporting links to some published data sets
  - "People getting the message that data has to be accessible."
  - Increased involvement of libraries and lifecycle management of data
  - Younger generation post data as they go – expectation that data should be shared
- Technology Trends and Applications
  - Digital object management technology
  - Growth of scientific workflow software
  - Adaptation of "netcentric" way of doing business
  - Use of embedded links in publications
  - Increased number of portals serving data sets

Key Background to Where We Are Today*









* There are many reports that cover scientific data. These show a direct lineage to national policy

Information International Associates, Inc. (IIa)

104 Union Valley Road

Oak Ridge, TN 37831

865-481-0388 (Main Office)

865-481-0390 (Fax)

http://www.iiaweb.com

Bonnie Carroll
President
bcarroll@iiaweb.com
(865) 298-1220

June Crowe
Tech. Dir. Intelligence Div.
jcrowe@iiaweb.com
(865) 298-1268