

# *Long-lived Data Collections*



## **The National Science Board**

**Dr. Chris Greer**

*Executive Secretary*

*NSB Long-lived Data Collections Task Force*

***A briefing to the:  
Biodiversity and Ecosystems Informatics Work Group  
August 25, 2004  
NSF, Arlington, VA***

# About the National Science Board

**National Science Board was established by the NSF Act of 1950:**

**“...to promote the progress of science; to advance the national health, prosperity, and welfare; and secure the national defense.”**

**The National Science Board provides:**

- **Oversight and policy-making** for NSF, and
- **Advice to the President and the Congress** on matters of national science and engineering policy.

# **National Science Board Long-lived Data Collections Task Force**

## **Charge:**

**“... delineate the policy issues relevant to the National Science Foundation and its style and culture of supporting the collection and curation of research data and make recommendations for the National Science Board and the community to consider.”**

# Central role of data collections in research and education

- **Provide a primary mechanism for scientific output**
- **Provide opportunities to broaden participation**

# LLDC Workshops

- **Workshop I: Nov. 18, 2003**

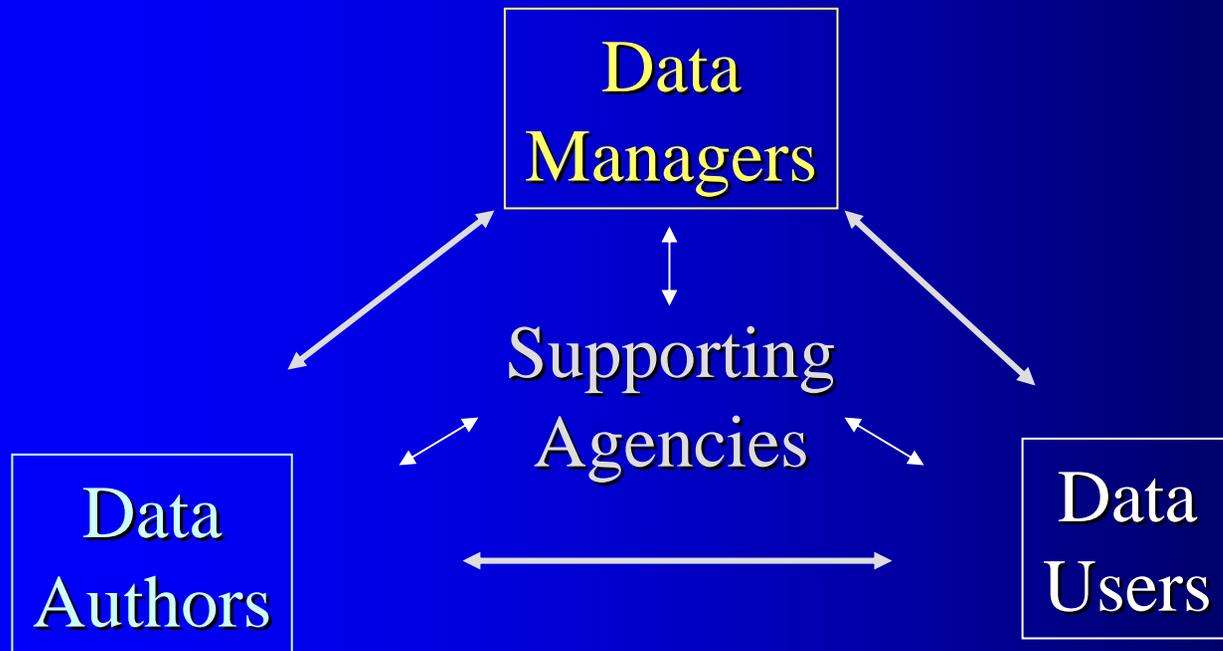
**NSF, DOD, DOE, EPA, NASA, NIH,  
NIST, NOAA, USGS**

- **Workshop II: Mar. 23, 2004**

**Universities, Research Consortia,  
Industry, Non-profit organizations,  
Federal Labs (FFRDCs)**

# Summary of issues raised at workshops

1. Policy should be informed by a clear vision of the needs and responsibilities of the participants in the “data collections universe”



# Summary of issues raised at workshops

2. The phrase “data collections” refers to a dynamic, heterogeneous community system
- **Research Collections:** Project level
  - **Resource Collections:** Community level
  - **Reference Collections:** Global

*Informed policy should recognize and build on the existing structure of the DC universe*

# Summary of issues raised at workshops

3. *How should crucial data collections, particularly long-lived data collections, be supported?*

**In what forms?**

**Centralized vs Distributed Models**

**By what means?**

**Direct vs Indirect Support Models**

# Summary of issues raised at workshops

4. *Should a comprehensive 'data plan' be a part of every proposal that would generate significant data sets?*

**Examples of data plan components:**

**Data management**

**Applicable standards**

**Access and Archiving provisions**

*Provides opportunity for peer-review, development and implementation of community standards*

# Summary of issues raised at workshops

5. *What would be the costs of implementing a 'data plan' requirement and how would these costs be covered?*
- Direct Costs Model
  - Indirect Costs Model
  - Structured, Combined Model

# Summary of issues raised at workshops

6. The activities of a data collection often go well beyond the collection and distribution of data.

**Examples: Curation, quality assessment/control**  
**Peer review**  
**Author attribution/credit**  
**Standards development and implementation**

*Which data collections should be expected to take on these 'community-proxy' responsibilities and how should the costs be supported?*

# Summary of issues raised at workshops

7. *What are the needs and opportunities for training scientists and educators for a digital research and education environment?*

**K through gray**

Pipeline and retraining issues

**New career paths**

Data scientists

# Overview of Issues Raised at NSB-sponsored Workshops

- **Develop policy that embraces the structure of the ‘collections universe’**
- **Delineate responsibility/authority/needs**
- **Evaluate data management plan options**
- **Evaluate community-proxy functions**
- **Provide for education/training/work force needs**