What Fran's Thinking About: Digital data: From here to eternity

Dr. Francine Berman

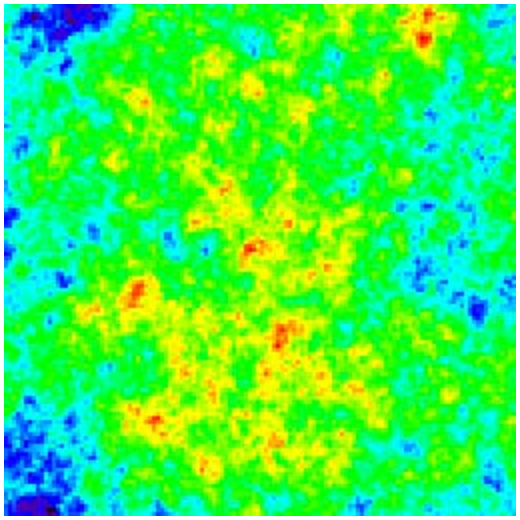
Director, San Diego Supercomputer Center

Professor and High Performance Computing Endowed Chair, UC San Diego

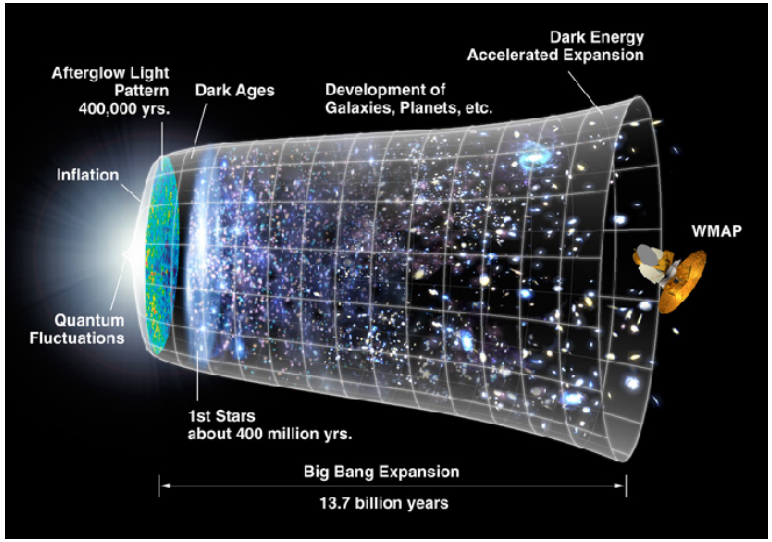
Our Digital World 1

- **The 2008 Cyber-election**
  - Fundraising via website
  - YouTube videos of the candidates and conventions
  - Blogs as vehicles for discussing issues
  - On-line organizing
- Digital data from historic 2008 cyber-election will be valuable for **decades+ to come**

Our Digital World 2



- **The First Billion Years After the Big Bang**
  - 400 TB of data produced from ENZO astrophysics simulations
  - Data will be mined and analyzed, of great value for **several years** after computation
- Simulation results illustrate growth of stars, galaxies, and galaxy clusters, dark matter, etc. after the Big Bang
- Large-scale simulations “refreshed” as resources become available



Images courtesy of Mike Norman

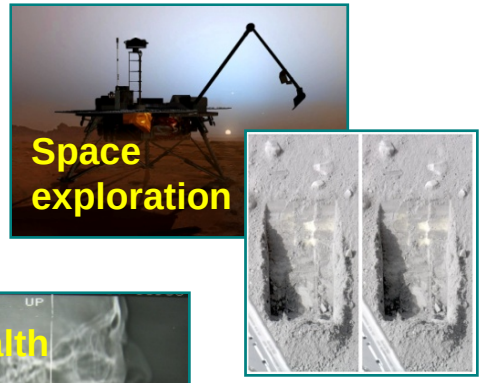
Our Digital World 3

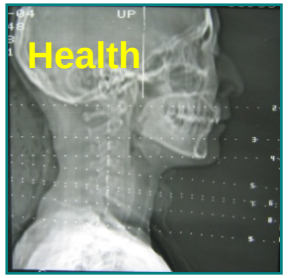
- **Family Photographs**

Some images courtesy of David Minor and the Library of Congress

Sustainable Preservation of Digital Materials is a Grand Challenge of the Information Age

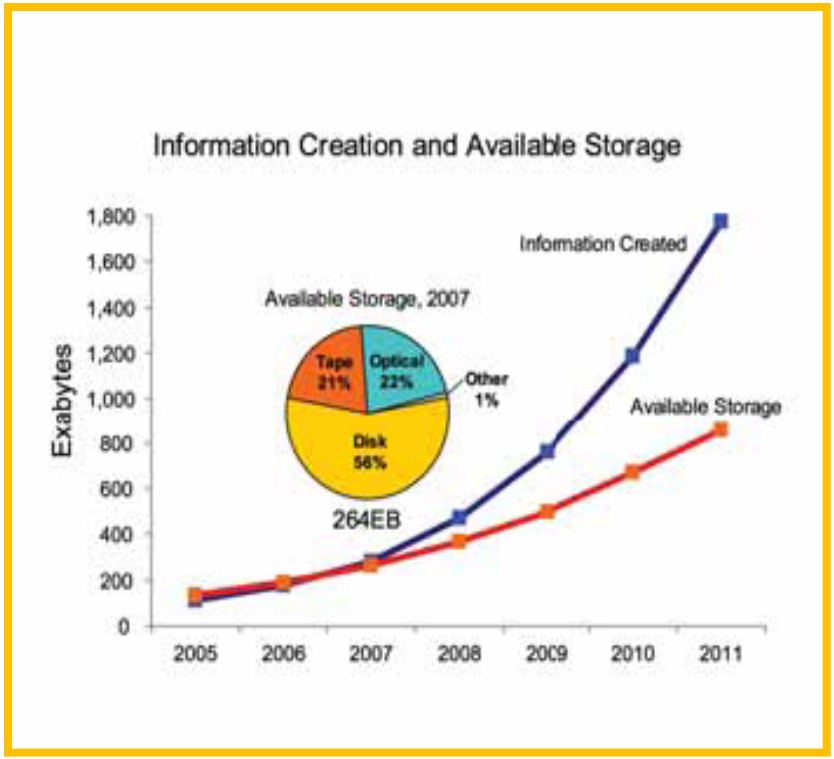
- **Digital preservation presents some of the greatest challenges for Cyberinfrastructure**
  - Greater emphasis on system reliability and security required
  - Smooth migration of digital materials from one generation of technology to the next critical
  - Indexing/organizing structures and associated meta-information must support current and support future search and use modes, policy and regulation, etc.
  - Data preservation efforts must be sustainable over the long-term (decades to centuries+)





Running Out of Room

- **Increasing requirements for data retention** in private, public, academic sectors render data infrastructure and preservation policies more critical. Many are currently “unfunded mandates”.
- ***However, even if we wanted to, we can’t save everything:*** **2007 was the “crossover year”** where the amount of digital information became greater than the amount of available storage



Source: *“The Diverse and Exploding Digital Universe” IDC Whitepaper, March 2008*

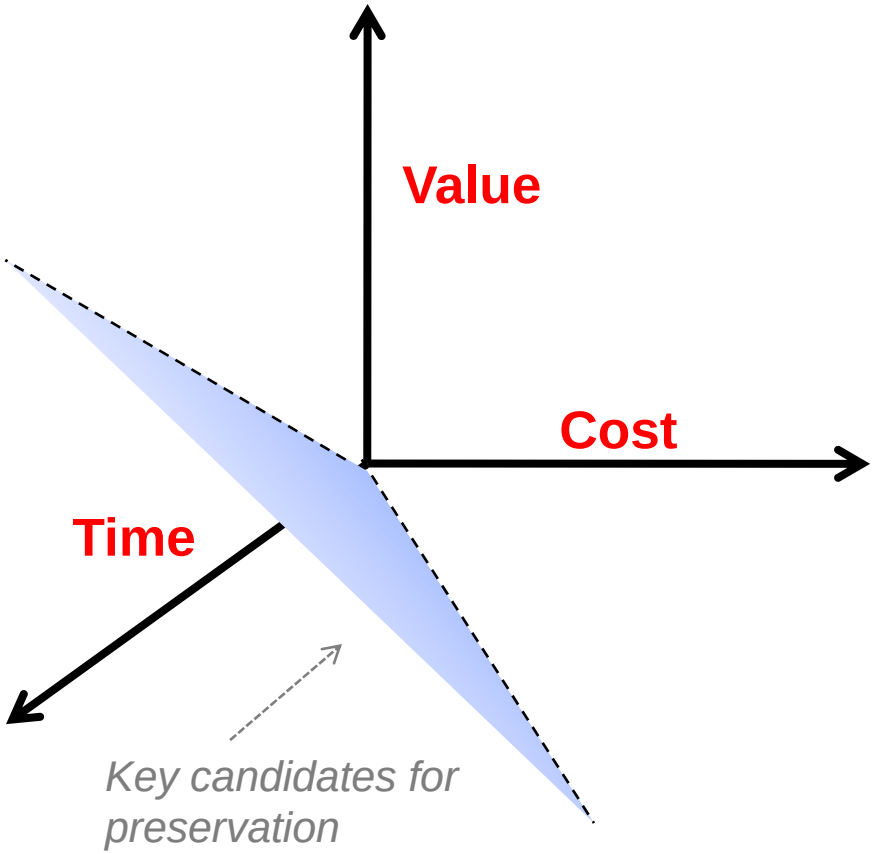
Data Preservation

- **Key Questions:**

  1) What should we save?
     – ***policy, regulation***

  2) How should we save it?
     – ***technology, best practice***

  3) Who should pay for it? -
     - ***economics***



Business Regulation
Requiring Data Preservation

Sarbanes-Oxley (Public Accounting Reform and Investor Protection Act of 2002)

- *Applies to all U.S. public company boards, management, and public accounting firms*
- **Includes electronic records** (correspondence, work papers, memoranda, etc.) that are created, sent, or received in connection with an audit or a review)
  - Section 103: "Board must require registered public accounting firms to "prepare, and maintain for a period of **not less than 7 years,** audit work papers, and other information related to any audit report, in sufficient detail to support the conclusions reached in that report."
  - Section 802: "any accountant who conducts an audit of an issuer of securities to which section 10(a) of the SEC …applies, shall maintain all audit or review work papers for a period of **5 years** from the end of the fiscal period in which the audit or review was concluded."

1. "Don't forget that **email and instant messaging are business records** …
4. Don't assume that the retention requirement …is …7 years. There are a lot of variables depending on the industry, type of organization and type of information. … **most lawyers that understand information retention agree that business records need to be kept indefinitely.**
10. Don't assume that just because you have access to archived information that you're going to be able to restore it within a reasonable amount of time…"

Kevin Beaver, "Thirteen Data Retention Mistakes to Avoid"
http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91_gci1186910,00.html

Health Regulation
Requiring Data Preservation

HIPAA (Health Insurance Portability and Accountability Act)

- *Applies to health information created or maintained by health care providers "who engage in certain **electronic transactions**, health plans, and health care clearinghouses" [www.hipaa.org]*
- Title II: Requires HHS to create rules and standards for the use and dissemination of health care information
- Healthcare providers must retain healthcare records for a period of **not less than 6 years**.

Table information partly based on "Data Retention – More Value, Less Filling", John Murphy, http://www.tdan.com/view-articles/5222

Increasing Policy and Regulation Affecting Research Community

- OMB requires that **federally funded research data,** supporting documentation, scientific notebooks, financial records, etc. be maintained **by the grantee for 3+ years**

- University libraries, federal agencies, institutional repositories ***not currently prepared*** to address the economic, technological, legal and social issues associated with widespread compliance of data retention policies

Crime and Punishment

Regulations	Retention Requirement	Penalty
HIPAA	Retain patient data for 6 years	$250K fine and up to 10 years in prison
Sarbanes-Oxley	Auditors must retain relevant data for at least 7 years	Fines to $5M and 20 years in prison
Gramm-Leach-Baily	Ensure confidentiality of customer financial information	Up to $500K and 10 years in prison
SEC 17a	Broker data retention for 3-6 years. Some require longer retention	Variable based on violation
OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)	"a three year period is the minimum amount of time that research data should be kept by the grantee"	Penalty structure unclear, likely fines?

Table information partly based on "Data Retention – More Value, Less Filling", John Murphy, http://www.tdan.com/view-articles/5222

How Should We Save It?

Technology: Increasing activity around data storage and preservation technologies, programs, and services

- **Academic sector:**
  - *IRODS* (rule-based distributed data management), *Fedora* (digital object repository system), *D-Space* (digital asset management), *LOCKSS* (peer-to-peer digital preservation infrastructure), etc.
- **Private sector:** *Amazon, MS, Google, Apple, Flickr, Sun, etc.*
- **Public Agencies/Institutions:** Library of Congress, NARA, NSF, NIH, DOE, Museums, Libraries, universities, state governments, etc.

However, there is no technology magic bullet …

Preserving digital data 100+ years will involve

- Tens-hundreds of new generations of technologies
- Thousands+ of new data standards and formats
- Millions+ of new valued collections
- Billions+ of potential users with as yet unknown information needs and workflows

Librarians' Perspective: People, Planning, Protections Critical Focus for the Preservation Environment

A Sample View of the Library of Congress Stewardship Network

Humanities and Sciences

- NETWORK ENVIRONMENT
- NETWORK SERVICES
- ORGANIZATIONAL FUNCTIONS
- ROLES
- NETWORK MANAGEMENT

Figure courtesy of Martha Anderson, Library of Congress

Research and Education User's Perspective: Key Questions Focus on Outcomes rather than Technology

How do I make sure that my data will be there when I want it?

How can I combine my data with my colleague's data?

How should I organize my data?

How should I display my data?

What are the trends and what is the noise in my data?

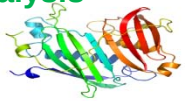
My data is confidential; how do I make sure that it is seen/used only by the right people?

How can I make my data accessible to my collaborators?

CI Perspective: Key Questions Focus on Integration, Capability, Usability, Reliability, Interoperability

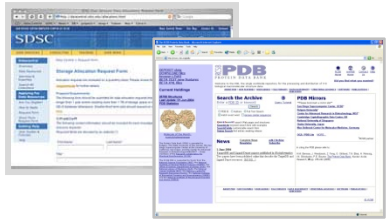
modeling

analysis



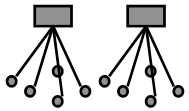
simulation

visualization

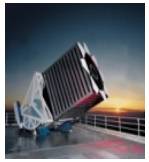


File systems,
Database systems,
Collection Management
Data Integration, etc.

instruments



Sensor-nets



computers

Many Data Sources

Data Access

Data Manipulation

Data Management

Data Storage

Integrated Cyberinfrastructure

SERVICES

- *Database selection and schema design*
- *Portal creation and collection publication*
- *Data analysis*
- *Data mining*
- *Data hosting*
- *Preservation services*
- *Domain-specific tools*
  - *Biology Workbench*
  - *Montage (astronomy mosaicking)*
  - *Kepler (Workflow management)*
- *Data visualization*
- *Data anonymization, etc.*

Current Best Practices in Digital Preservation

- **Replication** – make multiple copies and store some off-site
- **Heterogeneity** – more bio-diverse solutions tolerate greater error
- Associate **metadata** with data to aid access, management, search
- **Plan ahead** for smooth transition of data to new generations of media
- Align necessary level of **"trust"** with **reliability, infrastructure**
- Include **data costs** as part of the IT bill
- Pay attention to **security**
- Know the appropriate **regulations, policies, and penalties** that pertain to your data

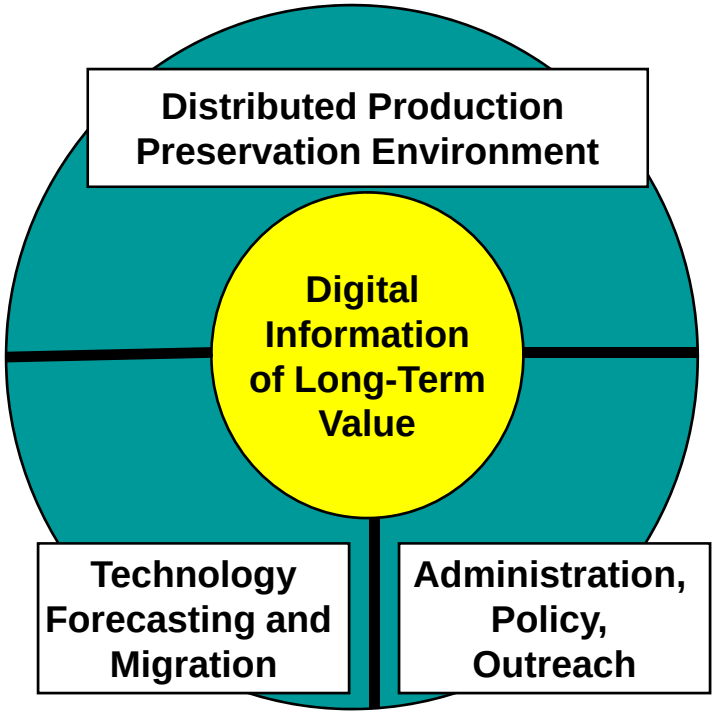
Why are 3 copies used as best practice?

- Approach comes from Lamport, Shostak, and Pease's solution to the *Byzantine General's Problem*
  - Method for agreement on a battle plan for a group of Byzantine generals communicating only by messenger
  - Analogous to reliable computer systems with malfunctioning components
- Solution: When generals can send unforgeable signed messages to one another, the minimum number required for agreement is 3.

Preservation Data Grid

The Chronopolis Model

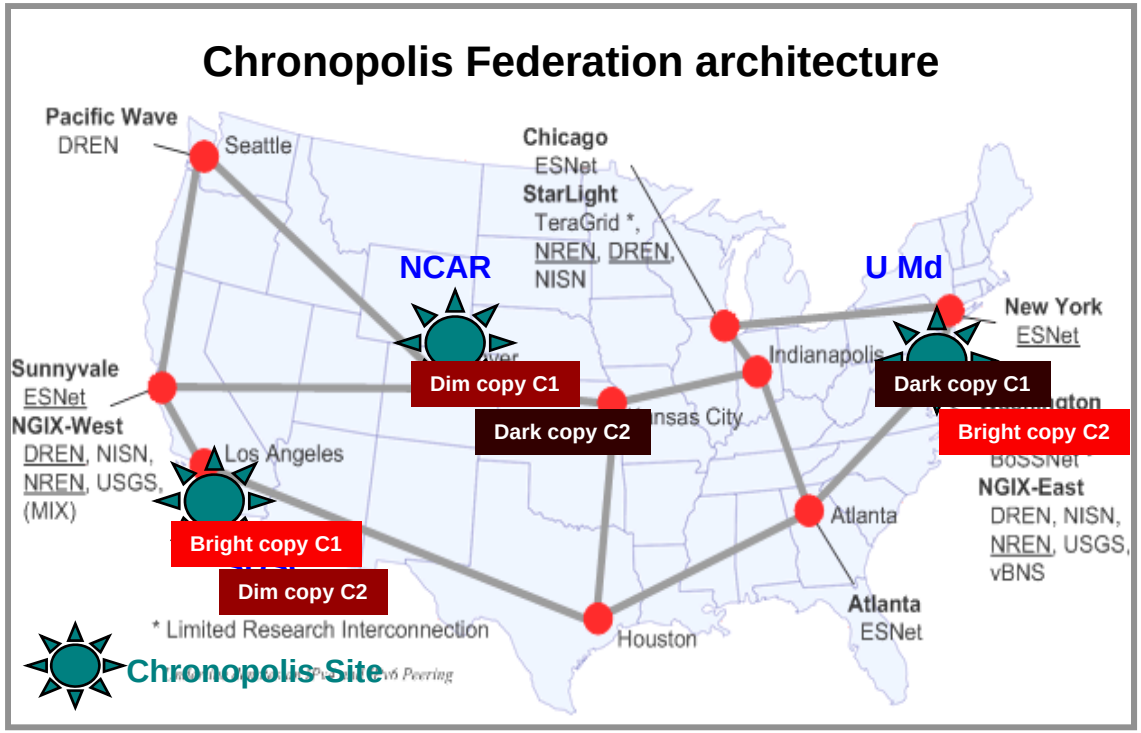
- Geographically distributed preservation data grid
  - supports long-term management, stewardship of, and access to digital collections
- Focus: technological, human, and policy infrastructure for preservation and life cycle management through multiple technology generations
- Chronopolis “users” = data stewards



http://chronopolis.sdsc.edu/

Data Replication and Distribution

- **Focus on supporting multiple, geographically distributed copies of preservation collections:**
  - **"Bright copy"** – Chronopolis site supports ingestion, collection management, user access
  - **"Dim copy"** – Chronopolis site supports remote replica of bright copy and supports user access
  - **"Dark copy"** – Chronopolis site supports reference copy that may be used for disaster recovery but no user access
- **Each site may play different roles for different collections**



- **Project Partners:** SDSC, UCSD Libraries, NCAR, University of Maryland
- **Sponsoring Agency:** Library of Congress
- **Data Partners:** ICPSR, CDL, Library of Congress, NCAR, NVO

Technology and Preservation: Hard Questions

Formalizing / quantifying "trust", reliability, etc.

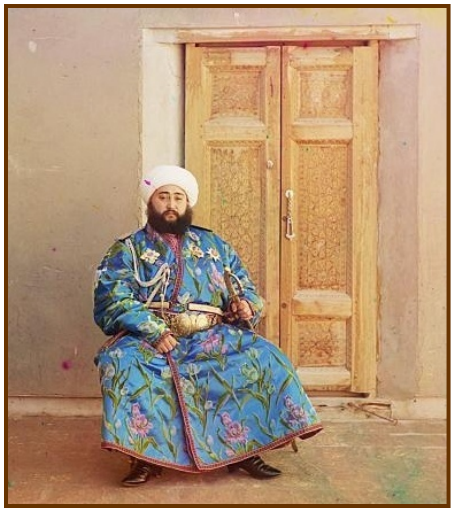
- What is the "gold", "silver", "bronze" standard for data reliability?
- What is the best approach for configuration / costing of data cyberinfrastructure at each level of reliability?

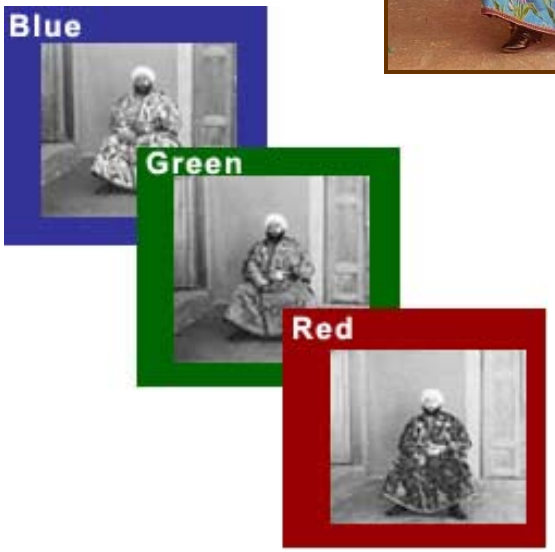
Risk management

- How should risk of single failure (media damage or corruption, natural disaster, operator error, hacker, etc.) be avoided?
- How much and what kind of data loss can be mitigated?

Optimizing Use

- What ontologies, metadata, and other structures optimize current and future use-case scenarios and policy/regulation





Prokudin-Gorskii Photographs

(Library of Congress Prints and Photographs Division)

http://www.loc.gov/exhibits/empire/

Library of Congress Pilot Project images and information courtesy of David Minor and LC

Who Should Pay?

The "Free Rider" Non-Solution

- **Inadequate/unrealistic approach: "Let X do it" where X is:**
  - The Government
  - The Libraries
  - The Archivists
  - Google, Yahoo, Microsoft, etc.
  - Data users
  - Data owners
  - Data creators, etc.

- **Valued digital data is a "public good"**
- **Creative partnerships needed** to provide preservation solutions with
  - Trusted stewards
  - Feasible costs for users
  - Sustainable costs for infrastructure
  - Very low risk for data loss, etc.

Economic Sustainability Requires A Holistic Approach

The economists' perspective: ***Economic sustainability for digital preservation*** is the set of business, social, technological and policy mechanisms that

1. *Encourage the gathering of important information assets into digital preservation systems*
2. *Support the indefinite persistence of the digital preservation systems, securing access to and use of the information assets into the long term future.*

- Economically sustainable digital preservation requires
  - **Recognition** of the benefits of preservation from decision makers
  - **Appropriate incentives** to induce decision makers to act in the public interest
  - **Mechanisms** to secure ongoing allocation of resources to digital preservation activities
  - **Efficient use of limited preservation resources**
  - **Appropriate organization and governance** of digital preservation activities.

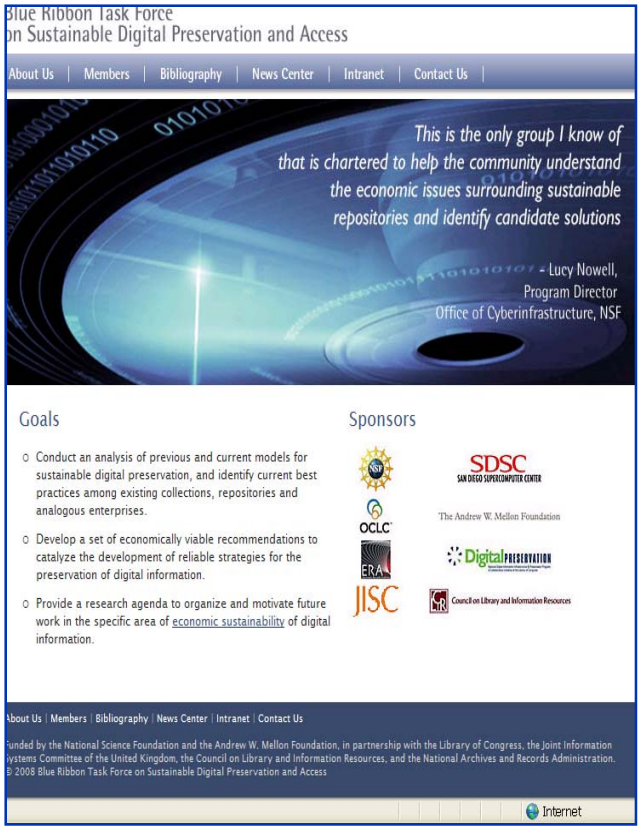
Many Economic Models Possible:

- ***Endowment***
  (data philanthropy)
- ***Institutional subsidy***
  (data welfare)
- ***Fee-based***
  - Membership /subscription
  - Ingestion fees
  - Access fees
  - Fee per use
- ***Advertising,*** etc.

Definition courtesy of Blue Ribbon Task Force on Sustainable Digital Preservation and Access

What Fran's Thinking About:
Blue Ribbon Task Force on Sustainable Digital Preservation and Access

- International **Blue Ribbon Task Force (BRTF-SDPA)** initiated to study issues of economic sustainability of digital preservation and access 2008-2009
- Supported by NSF, Library of Congress, Mellon Foundation, NARA, JISC, CLIR, member institutions



Task Force website
brtf.sdsc.edu

BRTF-SDPA CHARGE

1. To conduct a **comprehensive analysis of previous and current efforts** to develop and/or implement models for sustainable digital information preservation
2. To identify and **evaluate best practice** regarding sustainable digital preservation among existing collections, repositories, and analogous enterprises
3. To **make specific recommendations for actions** that will catalyze the development of sustainable resource strategies for the reliable preservation of digital information
4. Provide a **research agenda** to organize and motivate future work.

BRTF-SDPA

Participants:

- Paul Ayris, University College London
- Fran Berman, SDSC/UCSD
- Bob Chadduck, NARA Liaison
- Sayeed Choudhury, Johns Hopkins
- Elizabeth Cohen, AMPAS/Stanford
- Paul Courant, University of Michigan
- Lee Dirks, Microsoft
- Amy Friedlander, CLIR
- Chris Greer, NITRD Liaison
- Vijay Gurbaxani, UC Irvine
- Anita Jones, University of Virginia
- Ann Kerr, Consultant
- Brian Lavoie, OCLC
- Cliff Lynch, CNI
- Dan Rubinfeld, UC Berkeley
- Chris Rusbridge, DCC
- Roger Schonfeld, Ithaka
- Abby Smith, Consultant
- Anne Van Camp, Smithsonian

Deliverables

- **First Year (De**
  (positive, "w
  - Past and cur
    digital preser
  - Points of conv
  - What do we know
- **Second Year (Decem**
  (normative, "what should be")
  - General cost framework: key cost categories of digital preservation
  - Emodels/"scenarios": alternate ways of organizing digital preservation activities
  - Features, pros, cons, trade-offs, etc. of each model
  - List real world conditions for which each model is best suited.
    - "If your digital preservation context is X, we recommend you consider using model Y to organize your activities in a sustainable way."

Goal is to go *beyond* the "3 R's":

1. Data preservation is Really important
2. More Research is needed.
3. More Resources are needed.

Call to Action: *Focus Public Attention on the Need for Sustainable Digital Preservation and Cyberinfrastructure*

- Does your dry cleaner know what digital preservation is?
- Public discussion needed to focus attention on data preservation and cyberinfrastructure.
- When enough people think it's important, resources and opportunity will follow.

Thanks to Jack and Bernard for continuing this great tradition

Shameless plug:

Next Generation SDSC

mission, leadership, resources will be announced

October 14 *at our new building dedication.*

Please join us if you can!