Overview on Sustainable Digital Preservation and Access

eScience Seminar, Max Planck Society, June 20, 2008

Sayeed Choudhury

Johns Hopkins University

Presentation Outline

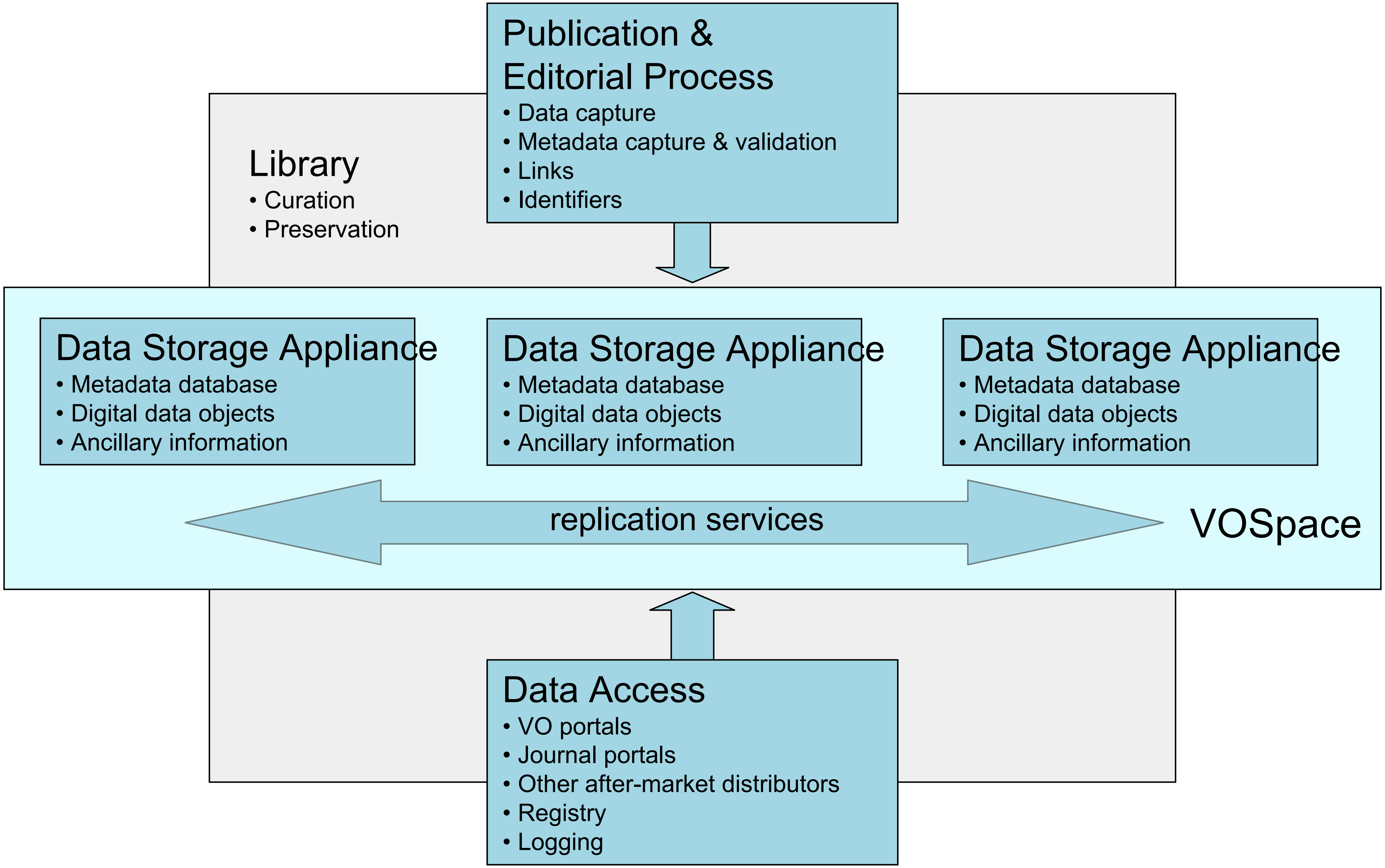
- Specific case study at Johns Hopkins University involving the Virtual Observatory
- Blue Ribbon Task Force on Sustainable Digital Preservation and Access
- Elements of DRAFT Task Force Report
- Report observations

The Virtual Observatory

- The Virtual Observatory enables new science by greatly enhancing access to data and computing resources. The VO makes it easy to locate, retrieve, and analyze data from archives and catalogs worldwide
- The VO is about *data discovery, access, and integration*
- The VO is NOT a huge centralized data repository
- The VO provides standard protocols for obtaining data from *distributed collections*
- The VO is *national* (US NVO) and *international* (IVOA)

Data Preservation Problem

- Research communities publish peer-reviewed journal papers that describe highly processed data
- Long-term preservation and curation systems for digital journal content, *including the digital data presented only graphically*, are not currently in place
- The research cannot be verified and the results cannot be easily compared to other data in order to broaden impact
- Public funds invested in scientific research do not have maximum return on investment. Essential legacy datasets may be lost



Publication & Editorial Process

- Data capture
- Metadata capture & validation
- Links
- Identifiers

Library

- Curation
- Preservation

Data Storage Appliance

- Metadata database
- Digital data objects
- Ancillary information

Data Storage Appliance

- Metadata database
- Digital data objects
- Ancillary information

Data Storage Appliance

- Metadata database
- Digital data objects
- Ancillary information

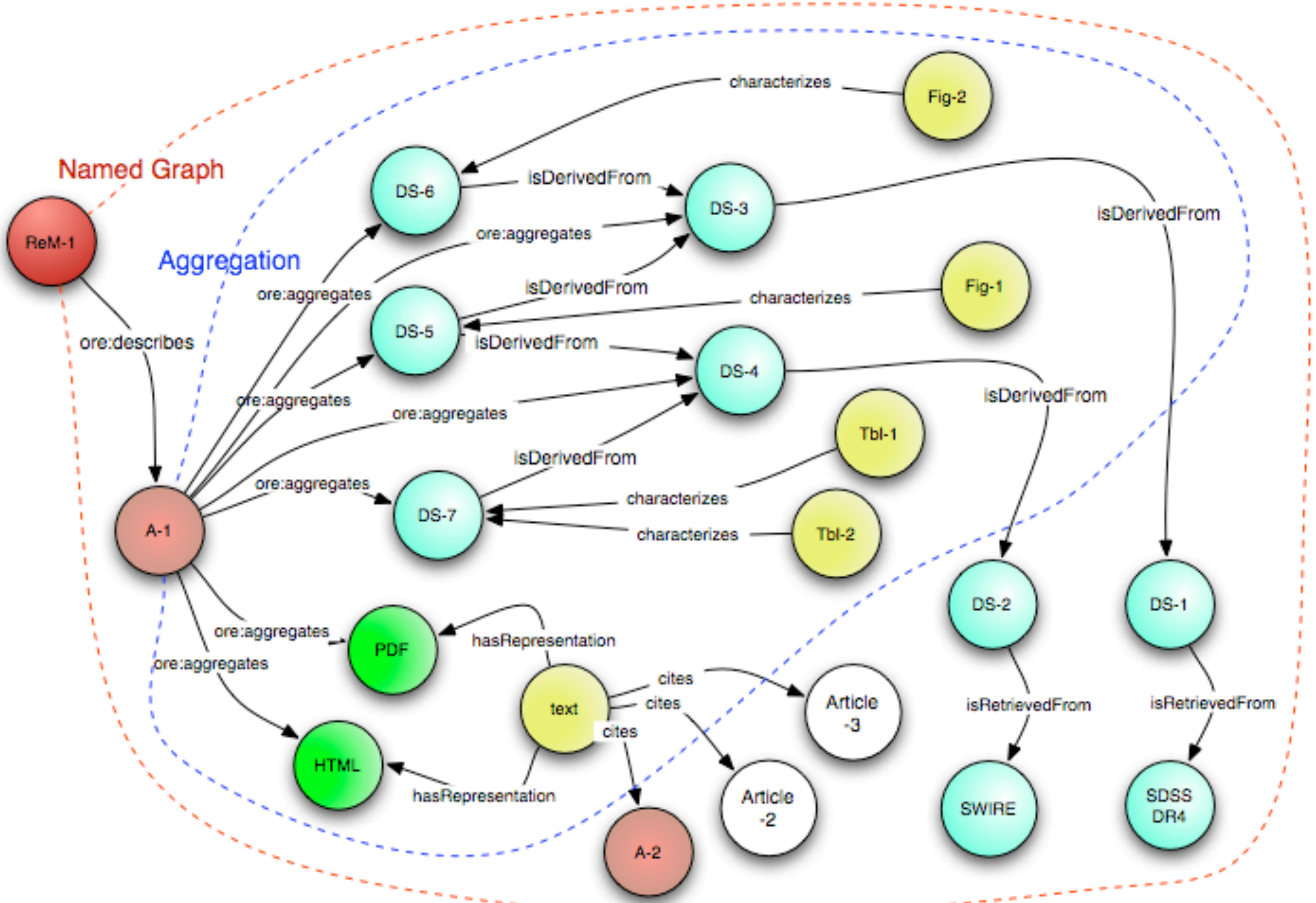
replication services

VOSpace

Data Access

- VO portals
- Journal portals
- Other after-market distributors
- Registry
- Logging

Open Archives Initiative – Object Reuse and Exchange



eScience Seminar – Max Planck Society – June 20, 2008

Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF-SPDA)

- Multi-disciplinary group with funding from US National Science Foundation and Andrew W. Mellon Foundation
- UK Joint Information Systems Committee nominated two representatives
- In-kind or staff support from Library of Congress, Council on Library and Information Resources, US National Archives and Records Administration, and NITRD
- Two year program of activities that began in January 2008 with kick-off meeting in Washington DC

Task Force Participants

Blue Ribbon Task Force:

- Paul Ayris, University College London
- Fran Berman, SDSC/UCSD
- Bob Chadduck, NARA Liaison
- Sayeed Choudhury, Johns Hopkins University
- 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

Sponsoring Agencies/Institutions:

- National Science Foundation
- Mellon Foundation
- Library of Congress
- National Archives and Records Administration
- CLIR
- NITRD
- JISC
- Member institutions

Specific Responsibilities

- Fran Berman / co-Chair
- Amy Friedlander / First Report Editor
- Ann Kerr / Panel Rapporteur
- Brian Lavoie / co-Chair
- Susan Rathbun / Task Force Support
- Abby Smith / Second Report Editor
- Jan Zverina / Communications Lead
- Lucy Nowell / NSF Program Officer
- Don Waters / Mellon Program Officer
- Laura Campbell, March Anderson / LC representative

Charge to the Task Force

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*

Key Areas for Recommendations

- *What are appropriate **roles and responsibilities for institutions** in the international, federal, academic, commercial, and non-profit/foundation sectors?*
  - What special capabilities do each of these sectors bring to the table and what are their inherent limitations?

- *What are the appropriate **roles and responsibilities for data authors, data users, and community groups**?*
  - What incentives exist or are needed to encourage and enable data authors to deposit data and metadata for preservation and reuse?

- *What **sustainability models** exist, can be adapted, and/or should be investigated to support long-term, sustainable digital information preservation?*
  - What cost/benefit analysis models exist or should be developed to evaluate these institutions and their operational models?
  - What incentives exist or are needed to encourage and enable institutions to preserve digital information over the long term?

- *How can we **characterize long-term digital preservation as an economic activity**, and what models can we bring to bear to understand its characteristics and policy implications?*
  - What are the alternative strategies for organizing digital preservation capacity (e.g. centralized services, distributed local capacity, etc.) and what are the pros and cons of each?

Deliverables

- **First Year Report (positive, “what is”):**
  - Describe past and current models (case studies, etc.);
  - Identify points of convergence/divergence; “lessons learned”;
  - What we know so far, and what our key knowledge gaps are.
- **Second Year Report (normative, “what should be”):**
  - General cost framework: key cost categories of digital preservation
  - Set of economic models/“scenarios”: alternate ways of organizing digital preservation activities, within the context of the cost framework
  - Describe each model: features, pros, cons, trade-offs, etc.
  - 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.”

TF Outreach:

- Community web resource and bibliography
- Articles designed to enlighten and broaden the community of stakeholders

DRAFT Initial Report

- Please keep in mind that the report is still in DRAFT format and the Task Force is still reviewing the document

- Definition of Economic Sustainability:

The set of business, social, technological, and policy mechanisms that 1) encourage the gathering of important information assets into digital preservation systems, and 2) support the indefinite persistence of the digital preservation systems, thus securing access to and use of the information assets into the long-term future.

Economic Sustainability (continued)

Economically sustainable digital preservation requires:

- Recognition of the benefits of preservation on the part of key decision-makers, as part of a process of selecting digital materials for long-term retention;
- Appropriate incentives to induce decision-makers to act in the public interest;
- Mechanisms to secure an ongoing allocation of resources, both within and across organizations, to digital preservation activities;
- Efficient use of limited preservation resources;
- Appropriate organization and governance of digital preservation activities.

Key Themes

- Definition of economic sustainability is meant to be sufficiently detailed enough to scope the work, yet also general enough to be applicable in most contexts
- The report focuses on sustainable economic models for digital materials for which there is a clear public interest
- Several concepts that should be considered over entire life-cycle of content including stakeholders, (both start-up and ongoing) costs, value, incentives, and organizational frameworks

Literature Review

- Focus on economic dimensions of digital preservation (not a complete review of all digital preservation literature)
- While there are other studies, reports, etc., seven major studies are examined in greater depth:
  1. Roquade Project
  2. Harvard Depository and OCLC, Inc. Digital Archive
  3. Sweden’s Riksarkivet/National Archives
  4. Digital Preservation Testbed, Nationaal Archief of the Netherlands
  5. Digital Motion Pictures (Science Technology Council of the Academy of Motion Picture Arts and Sciences)
  6. LIFE Project
  7. Beagrie-Chruszcz-Lavoie (BCL) Cost Model

Report (DRAFT!) Observations

- Difficult to compare and contrast different studies given diverse definitions of costs, units of measurement, scope or accounting methods
- Most recent studies such as LIFE or BCL cost model attempt to account for entire life-cycle of costs and also intertemporal considerations (e.g., inflation/deflation, interest rates)
- Important to acknowledge costs associated with human dimensions (e.g., management) and infrastructure elements of storage (e.g., power and cooling)
- "Format matters; scale matters"

Acknowledgements

- Robert Hanisch for Virtual Observatory slides
- US Institute of Museum and Library Services and Microsoft for funding related to Virtual Observatory data curation prototype development
- Fran Berman and Brian Lavoie for BRTF-SPDA slides
- Members of BRTF-SPDA for contributions to draft initial report
- Max Planck Society for invitation