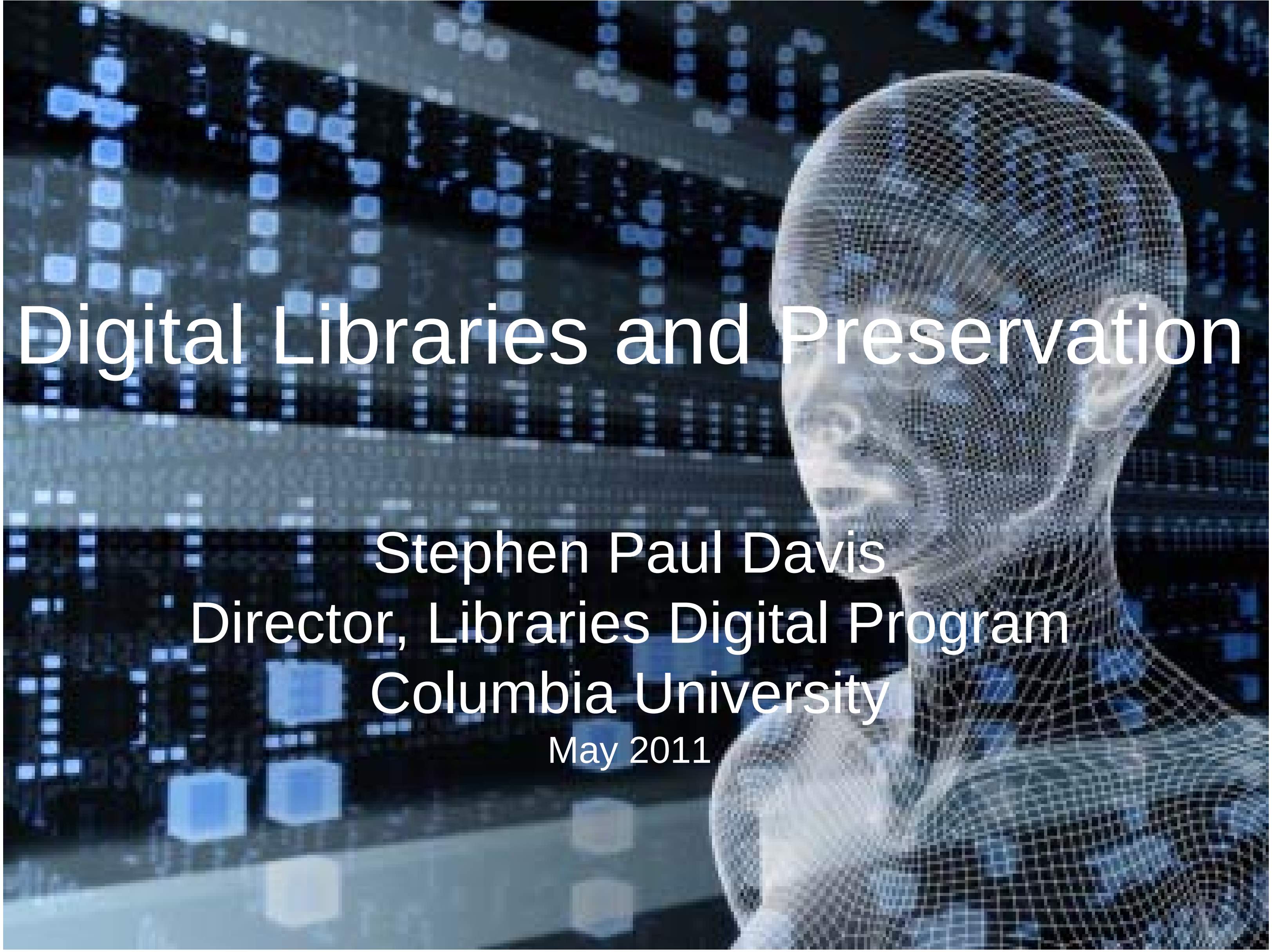


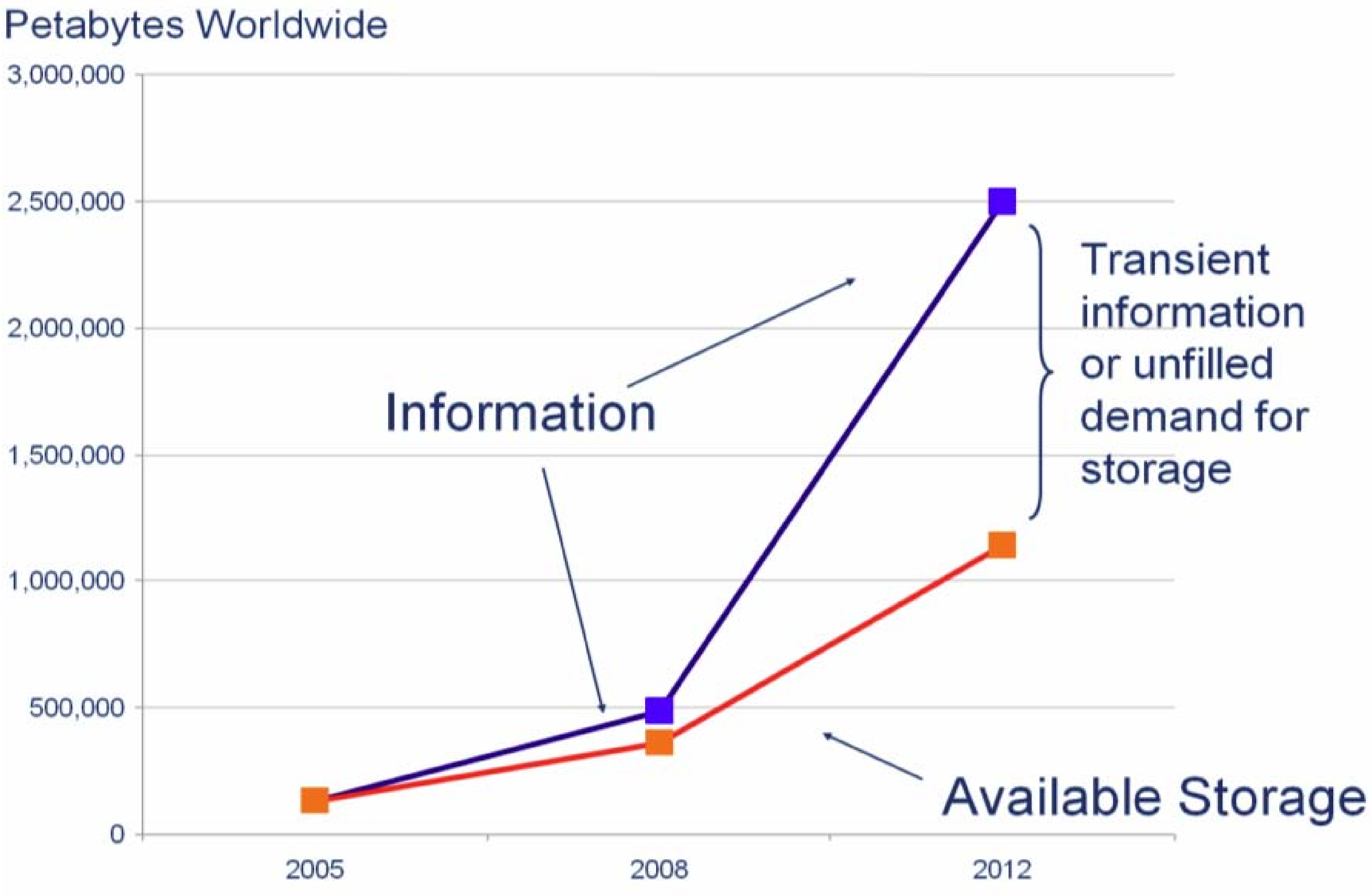
Digital Libraries and Preservation

Stephen Paul Davis
Director, Libraries Digital Program
Columbia University
May 2011

So What’s the Problem?

Digital Content Disappears

“An estimated 44 percent of Web sites that existed in 1998 vanished without a trace within just one year. The average life span of a Web site is only 44 to 75 days.” Jim Barksdale, Francine Berman, Washington Post, 5/16/2007.



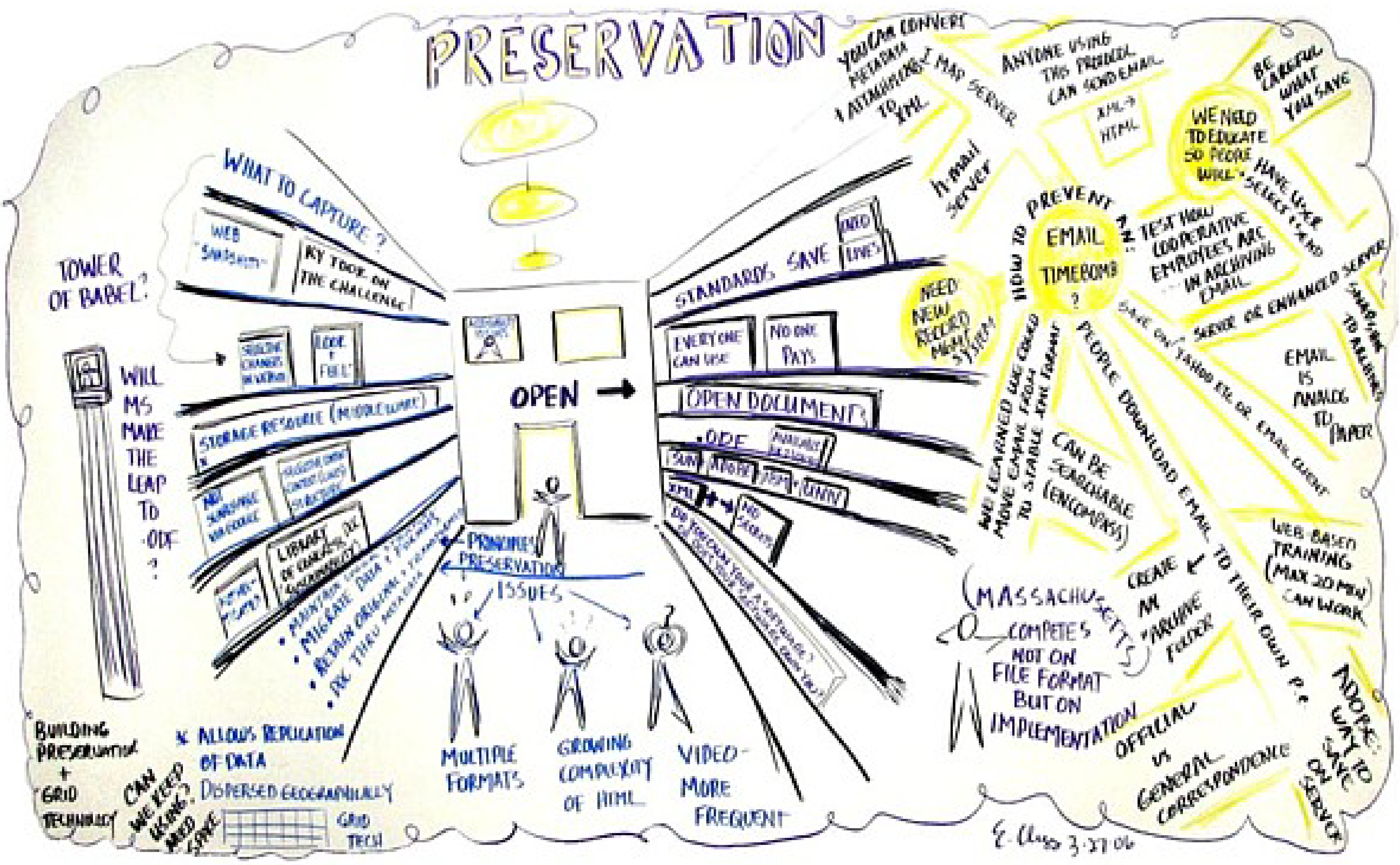
From: [Sustainable Economics for a Digital Planet (2010)]

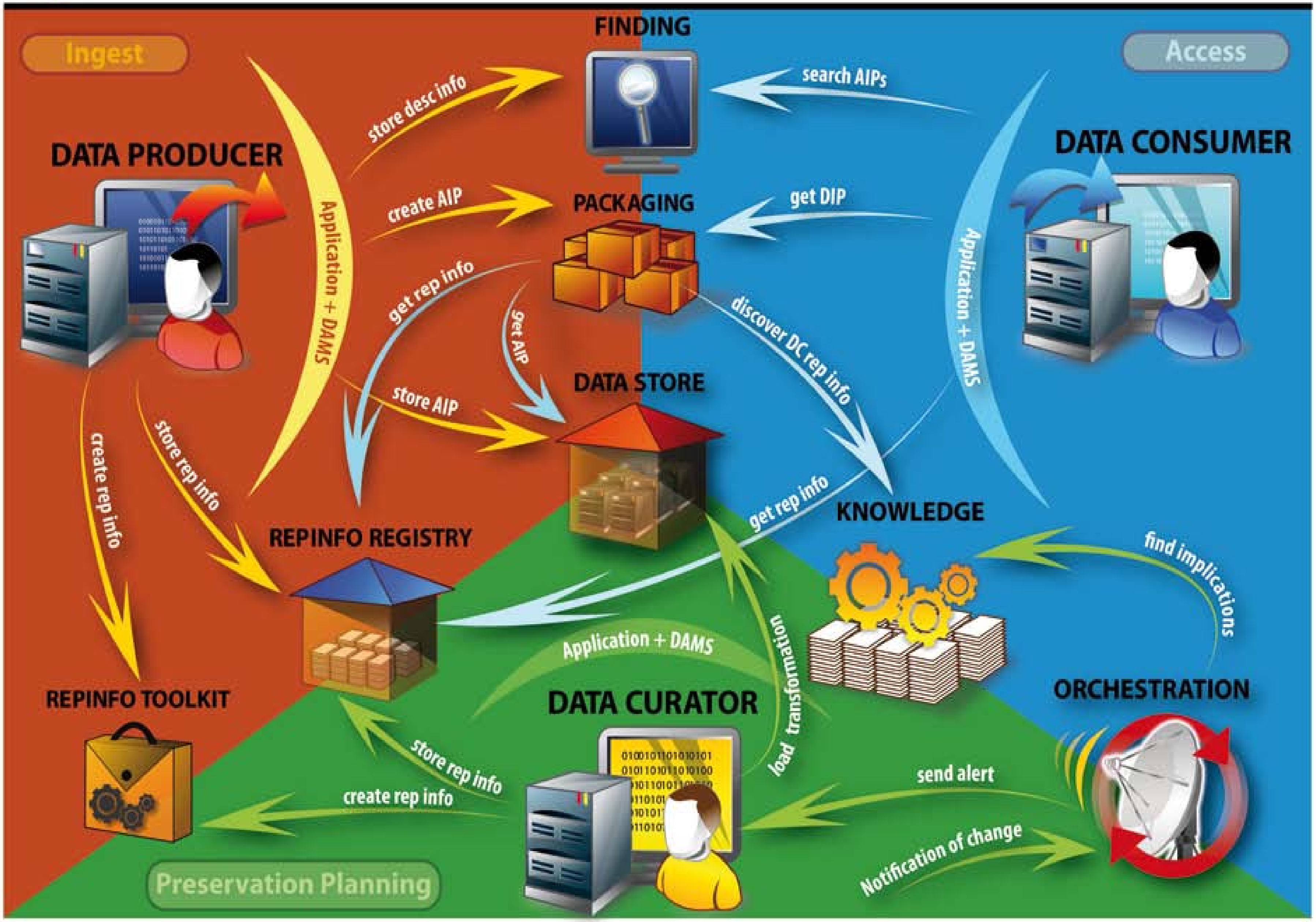
Should everything really be saved?

- Nina Matheson (no)
- Frank Moretti (no)
- Stephen Davis (stupid question)

Ensuring Access

- Digital information is vital
- Digital information is fragile
- Access in the future means action today
- Digital preservation is complicated


PRESERVATION
WHAT TO CAPTURE ?
WEB "SNAPSHOT"
KY TOOK ON THE CHALLENGE
TOWER OF BABEL?
LOOK + FEEL
STORAGE RESOURCE (MIDDLEWARE)
LIBRARY OF CONGRESS, DC
WILL MS MAKE THE LEAP TO .ODF ?
• MAINTAIN
• MIGRATE DATA + FORMATS
• RETAIN ORIGINAL + TRANSFORM
• DC THRU METADATA
BUILDING PRESERVATION + GRID TECHNOLOGY
CAN WE KEEP USING? NEED SPACE
* ALLOWS REPLICATION OF DATA
DISPERSED GEOGRAPHICALLY
GRID TECH
OPEN →
PRINCIPLES
PRESERVATION
ISSUES
MULTIPLE FORMATS
GROWING COMPLEXITY OF HTML
VIDEO - MORE FREQUENT
STANDARDS SAVE
NEED LINES
EVERYONE CAN USE
NO ONE PAYS
OPEN DOCUMENTS
.ODF
SUN
ADOBE
IBM
UNIV
XML
NO SECRETS
YOU CAN CONVERT METADATA + ATTACHMENTS TO XML
MAP SERVER
ANYONE USING THIS PROTOCOL CAN SEND EMAIL
XML → HTML
IN MAIL SERVER
BE CAREFUL WHAT YOU SAY
WE NEED TO EDUCATE SO PEOPLE WILL
HAVE USER SELECT + ADD
NEED NEW RECORD MGMT SYSTEM
HOW TO PREVENT AN EMAIL TIMEBOMB ?
TEST HOW COOPERATIVE EMPLOYEES ARE IN ARCHIVING EMAIL
SERVER OR ENHANCED SERVER
SAVE ON YAHOO ETC OR EMAIL CLIENT
EMAIL IS ANALOG TO PAPER
WE LEARNED WE COULD MOVE EMAIL FROM TO STABLE XML FORMAT
CAN BE SEARCHABLE (ENCOMPASS)
PEOPLE DOWNLOAD EMAIL TO THEIR OWN P.C.
WEB BASED TRAINING (MAX 20 MIN CAN WORK)
MASSACHUSETTS
COMPETES NOT ON FILE FORMAT BUT ON IMPLEMENTATION
CREATE AN "ARCHIVE" FOLDER
OFFICIAL vs GENERAL CORRESPONDENCE
ADOBE: WAY TO SAVE ON SERVER
E. Class 3.27.06


Ingest
Access
Preservation Planning
FINDING
PACKAGING
DATA STORE
KNOWLEDGE
ORCHESTRATION
DATA PRODUCER
DATA CONSUMER
DATA CURATOR
REPINFO REGISTRY
REPINFO TOOLKIT
store desc info
create AIP
get rep info
get AIP
store AIP
Application + DAMS
store rep info
create rep info
search AIPs
get DIP
discover DC rep info
Application + DAMS
get rep info
find implications
Application + DAMS
load transformation
store rep info
create rep info
send alert
Notification of change

What Needs to be Preserved?

- Digitally-published commercial content
- Digitally-published non-commercial content
- Born-digital research data
- Born-digital organizational data
- Born-digital personal data
- Digitized analog collections

How do we preserve it?

- Digital preservation is not a one-time action
- Digital preservation requires ongoing digital “lifecycle management”
- Digital preservation is more complicated and expensive than anything libraries have ever attempted
- We believe we need to build “Trusted Digital Repositories”

So what is a “trusted digital repository”?

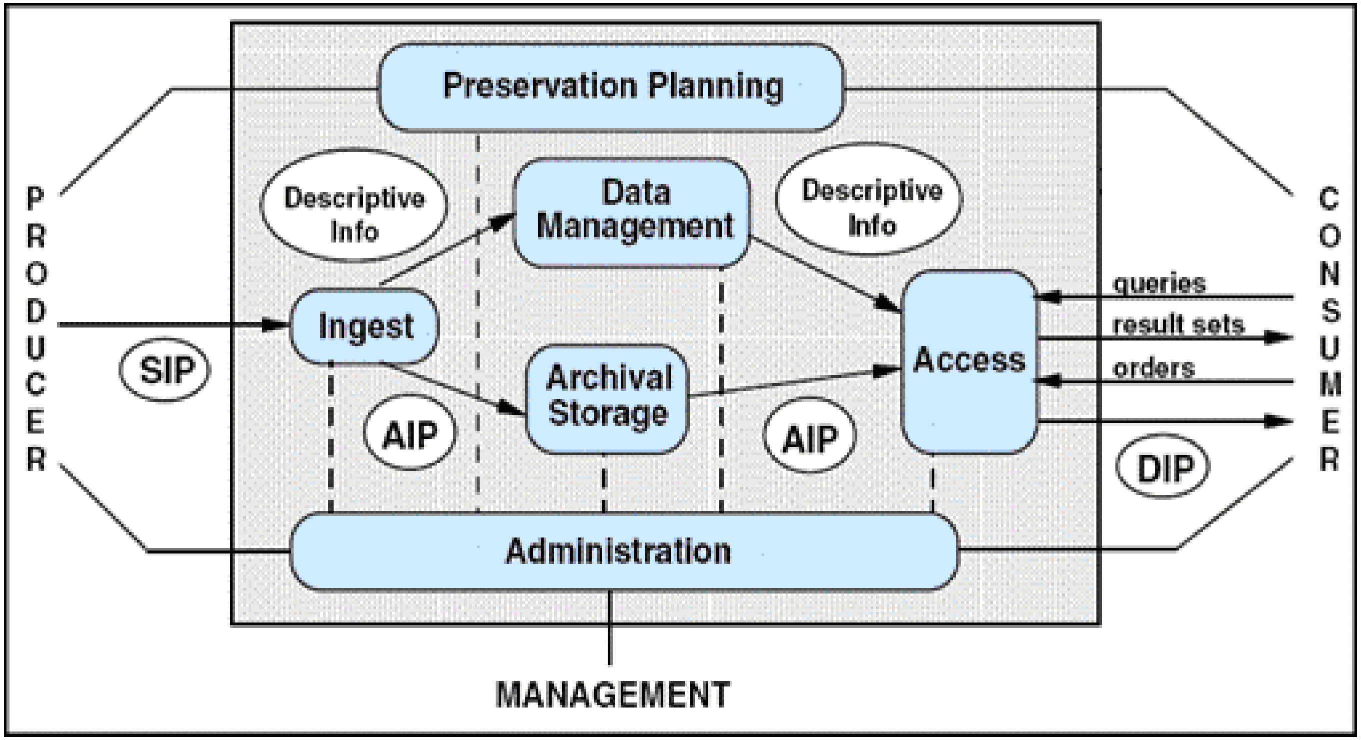
“Trusted Digital Repositories”

- *Trustworthy Repositories Audit & Certification: Criteria and Checklist (2007)*
- *Drambora: “Digital Repository Audit Method Based on Risk Assessment”* (Digital Curation Center / Digital Preservation Europe) – toolkit for self-assessment (2007)
- Other, domain-specific approaches

TRAC Compliance #1

OAIS compliance

System must comply with OAIS Standard



TRAC Compliance #2

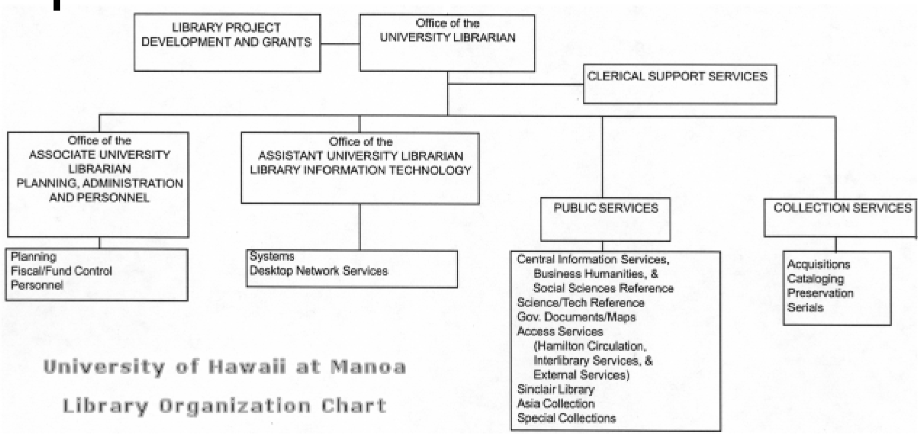
Administrative responsibility

A commitment to track and comply with current and emerging standards embraced by the preservation community.

TRAC Compliance #3

Organizational viability

Capacity to receive, store, preserve, and provide access to digital materials under its care, encompassing legal, fiscal, and ethical considerations and requirements.



LIBRARY PROJECT DEVELOPMENT AND GRANTS

Office of the UNIVERSITY LIBRARIAN

CLERICAL SUPPORT SERVICES

Office of the ASSOCIATE UNIVERSITY LIBRARIAN PLANNING, ADMINISTRATION AND PERSONNEL

Planning
Fiscal/Fund Control
Personnel

Office of the ASSISTANT UNIVERSITY LIBRARIAN LIBRARY INFORMATION TECHNOLOGY

Systems
Desktop Network Services

PUBLIC SERVICES

Central Information Services,
Business Humanities, &
Social Sciences Reference
Science/Tech Reference
Gov. Documents/Maps
Access Services
(Hamilton Circulation,
Interlibrary Services, &
External Services)
Sinclair Library
Asia Collection
Special Collections

COLLECTION SERVICES

Acquisitions
Cataloging
Preservation
Serials

University of Hawaii at Manoa

Library Organization Chart

TRAC Compliance #4

Financial Sustainability

Accounting and budget policies and procedures that are part of a business plan to define and protect requisite resources for the digital preservation program.

TRAC Compliance #5

Technological Suitability

Capacity to develop and maintain requisite hardware, software, expertise, and techniques to support and enable the digital preservation program, including adherence to relevant standards and industry best practice.

TRAC Compliance #6

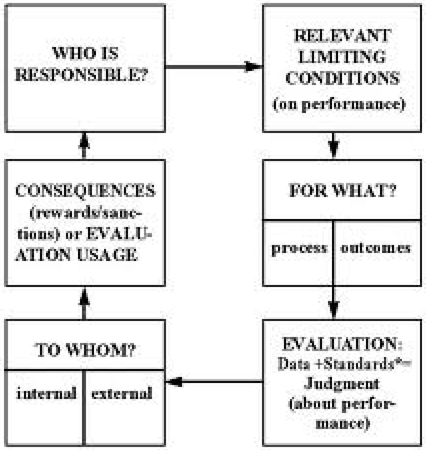
System Security

A commitment to maintaining a constant and appropriate level of environmental and online protection; surveillance; and risk detection, response, and mitigation to safeguard the integrity of digital assets.

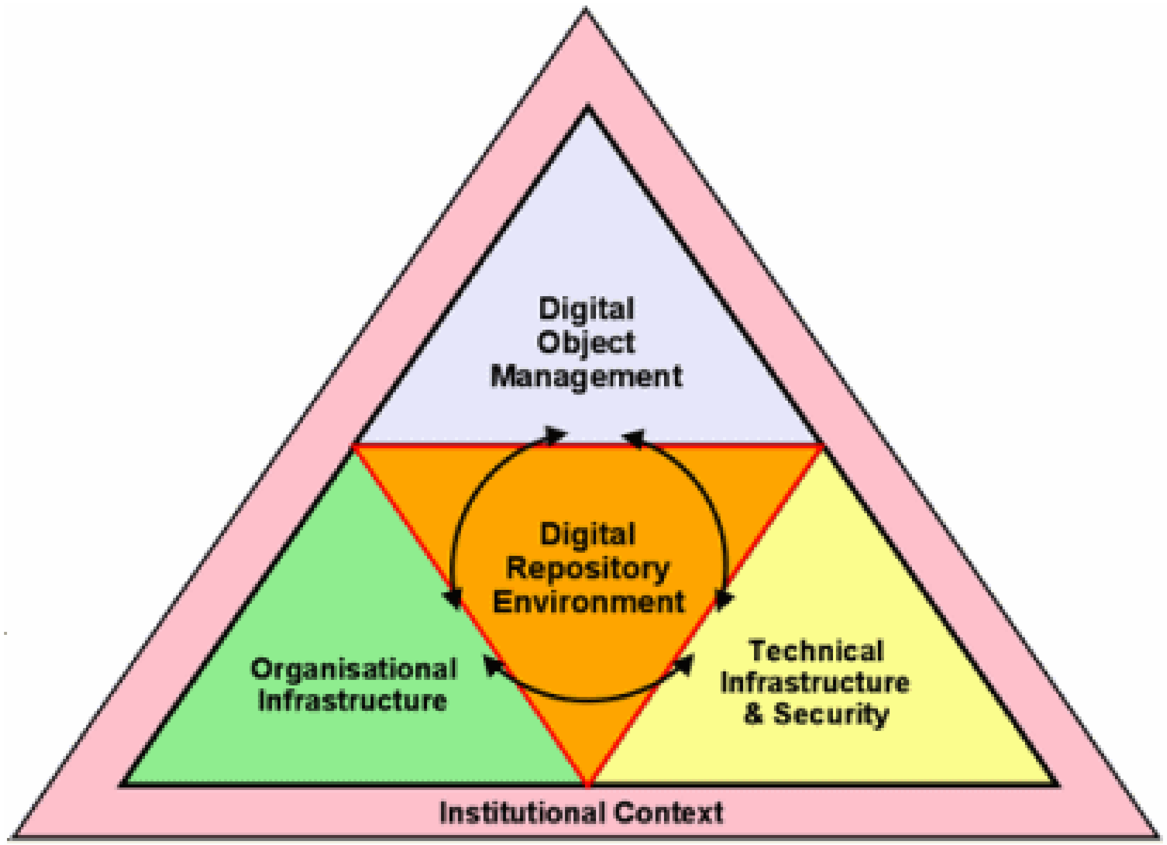
TRAC Compliance #7

Procedural Accountability

A means for documenting, sharing, and applying the set of policy statements and associated procedures and prevailing practice.


WHO IS RESPONSIBLE?
RELEVANT LIMITING CONDITIONS (on performance)
FOR WHAT?
process
outcomes
EVALUATION: Data +Standards⇒ Judgment (about performance)
TO WHOM?
internal
external
CONSEQUENCES (rewards/sanctions) or EVALUATION USAGE

Key Vectors of a Trusted Digital Repository

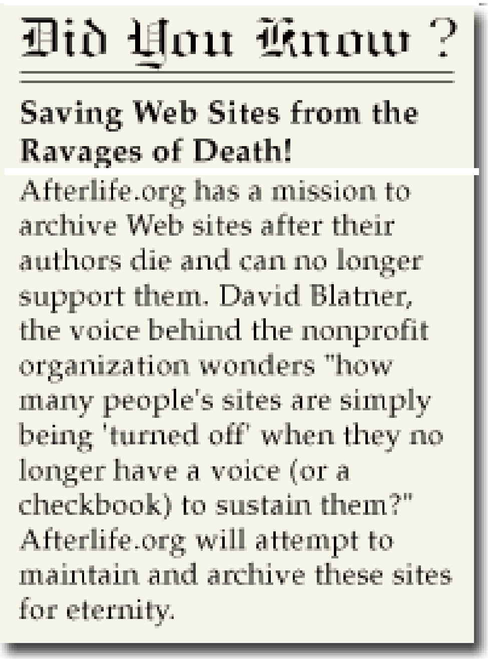

Digital Object Management
Digital Repository Environment
Organisational Infrastructure
Technical Infrastructure & Security
Institutional Context

Digital Preservation Strategies

- Bitstream copying
- Refreshing
- Technology preservation
- Digital archaeology & forensics
- Analog backups
- Content migration, emulation
- Replication
- Reliance on Standards

Where are “trusted digital repositories”?

- •LOCKSS (2006)
- •ICPSR (2006)
- •Portico (2010)
- •Hathi Trust (2010)
- •Columbia (2012?)



Did You Know ?

Saving Web Sites from the Ravages of Death!

Afterlife.org has a mission to archive Web sites after their authors die and can no longer support them. David Blatner, the voice behind the nonprofit organization wonders "how many people's sites are simply being 'turned off' when they no longer have a voice (or a checkbook) to sustain them?" Afterlife.org will attempt to maintain and archive these sites for eternity.

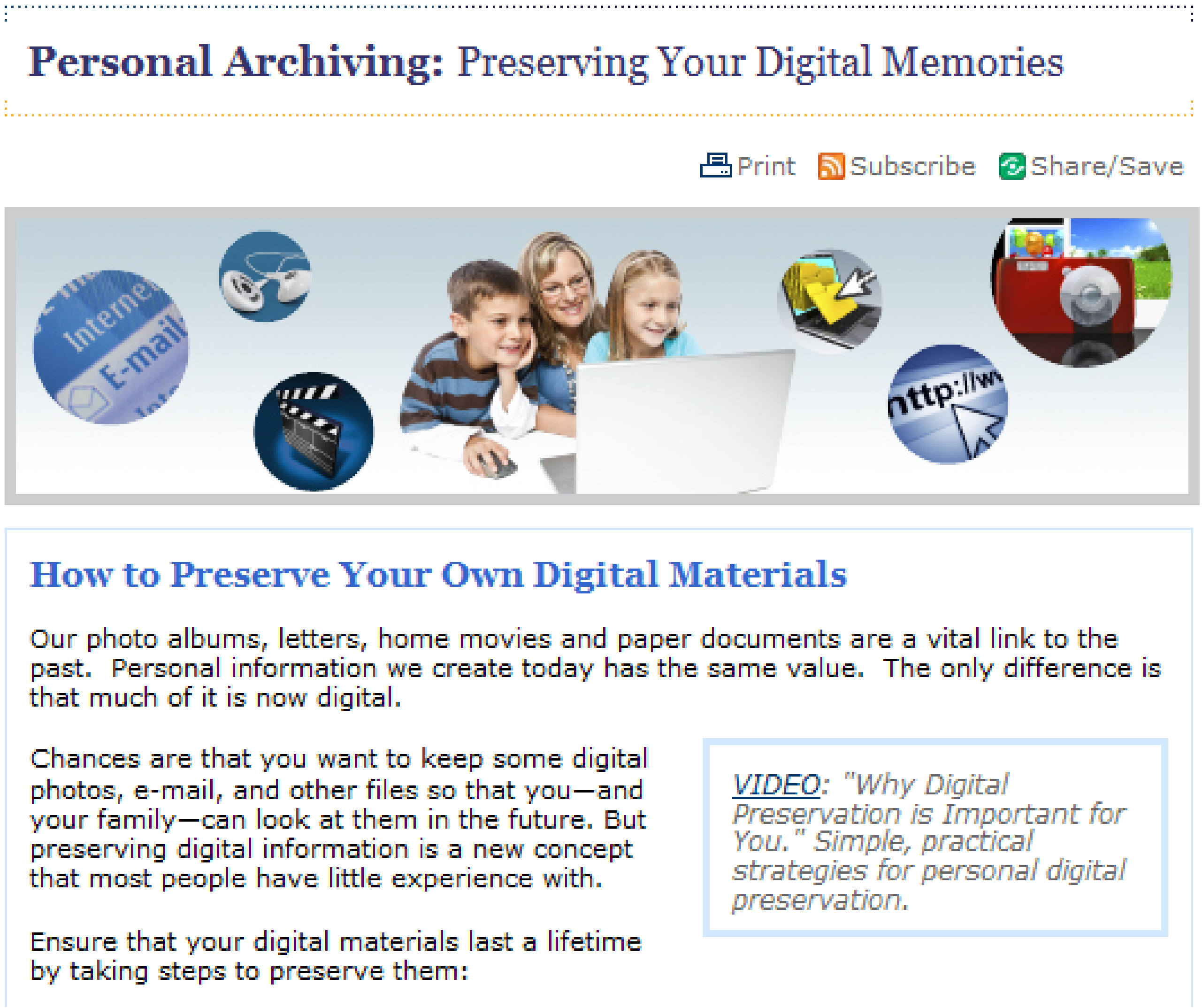
Afterlife.org ?

Digital Genome Project (Planets)

Digital Genome Project

- Five major at risk formats - JPEGs, JAVA source code, .Mov files, websites using HTML, and PDF documents
- Versions of these files stored in archival standard formats – JPEG2000, PDFA, TIFF and MPEG4 – to prolong lifespan for as long as possible
- 2500 additional pieces of data – mapping the genetic code necessary to describe how to access these file formats in future
- Translations of the required code into multiple languages to improve chances of being able to interpret in the future
- Copies of all information stored on a complete range of storage media – from CD, DVD, USB, Blu-Ray, Floppy Disc, and Solid State Hard Drives to audio tape, microfilm and even paper print outs

“Personal Archiving”



Personal Archiving: Preserving Your Digital Memories

How to Preserve Your Own Digital Materials

Our photo albums, letters, home movies and paper documents are a vital link to the past. Personal information we create today has the same value. The only difference is that much of it is now digital.

Chances are that you want to keep some digital photos, e-mail, and other files so that you—and your family—can look at them in the future. But preserving digital information is a new concept that most people have little experience with.

Ensure that your digital materials last a lifetime by taking steps to preserve them:

VIDEO: "Why Digital Preservation is Important for You." Simple, practical strategies for personal digital preservation.

Columbia University Libraries

Preservation

Strategies and Technologies

Preservation Focus at Columbia

1. Local Digitization Projects
2. Institutional Repository / Data Sets
3. Born Digital Archival Content
4. Archived Web Sites

Local Digitization Projects

Preservation of *unique digitized content* created from print, manuscript and multimedia collections

E.g.,

- papyri, medieval manuscripts, image and object collections, rare books and journals, archival collections, useful reference and curricular material, oral histories

Institutional Repository

Preservation of *University-generated content* of all kinds (working papers, conference proceedings, theses, preprints, ***research data sets***)

Academic Commons (Columbia' Institutional Repository)

NSF Data Management Plan Support

Born Digital Archival Content

Preservation of *born-digital* personal and organizational archival collections (e.g., of authors, political figures, publishing houses, philanthropic organizations)

E.g.,

- Human Rights Watch Records
- Bomb Magazine Records
- Carnegie Corporation of New York Records

"International Non-Profit Organization"

- 22 office in 18 countries
- Entire organization being dismantled in 2014
- Columbia to develop repository-based services to acquire, ingest, process, preserve and make accessible their born-digital archival content
- Build infrastructure for additional such archival projects going forward

Archived Web Sites

Preservation of *significant and at-risk Web sites* of potential value to scholars and researchers of the future

E.g.,

Columbia Human Rights Web Archive

(see e.g. savetibet.org)

… archived via "Achive-It," a service of the *Internet Archive*.

Columbia's Digital Preservation Infrastructure

Columbia is building a repository system and robust application development platform for:

- Digital asset management
- Digital asset 'curation'
- Controlled access to digital assets and collections
- Long-term digital preservation

Essential infrastructure for digital preservation.

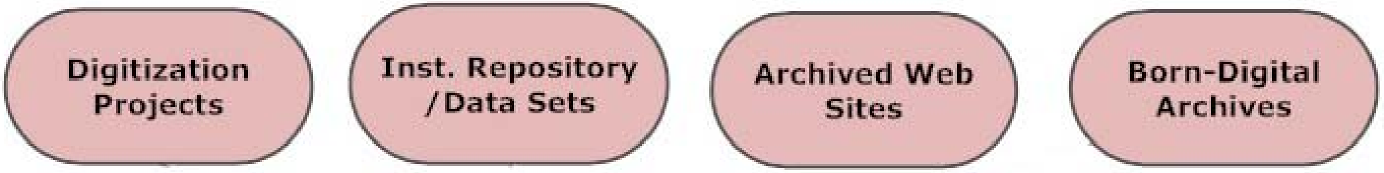
Fedora Commons Repository Software

- Robust open-source development community
- Supported by Duraspace consortium & several funding agencies
- Broad adoption within higher education (see User Registry)
- *Columbia is a "gold" member of the Duraspace and one of our programmers is a Fedora "committer*

CUL/IS Fedora Architecture

- Fedora Software Platform
- Digital Preservation Storage System
- Application and authentication middleware
- Applications to support Long Term Preservation Archive

Columbia University Libraries / Information Services
Digital Archiving Overview



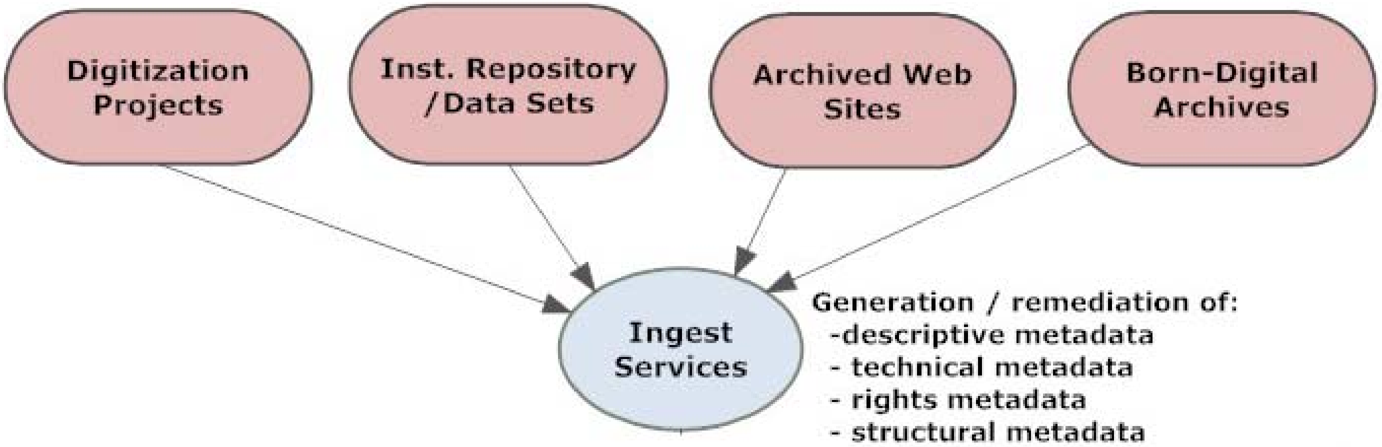
Digitization Projects

Inst. Repository /Data Sets

Archived Web Sites

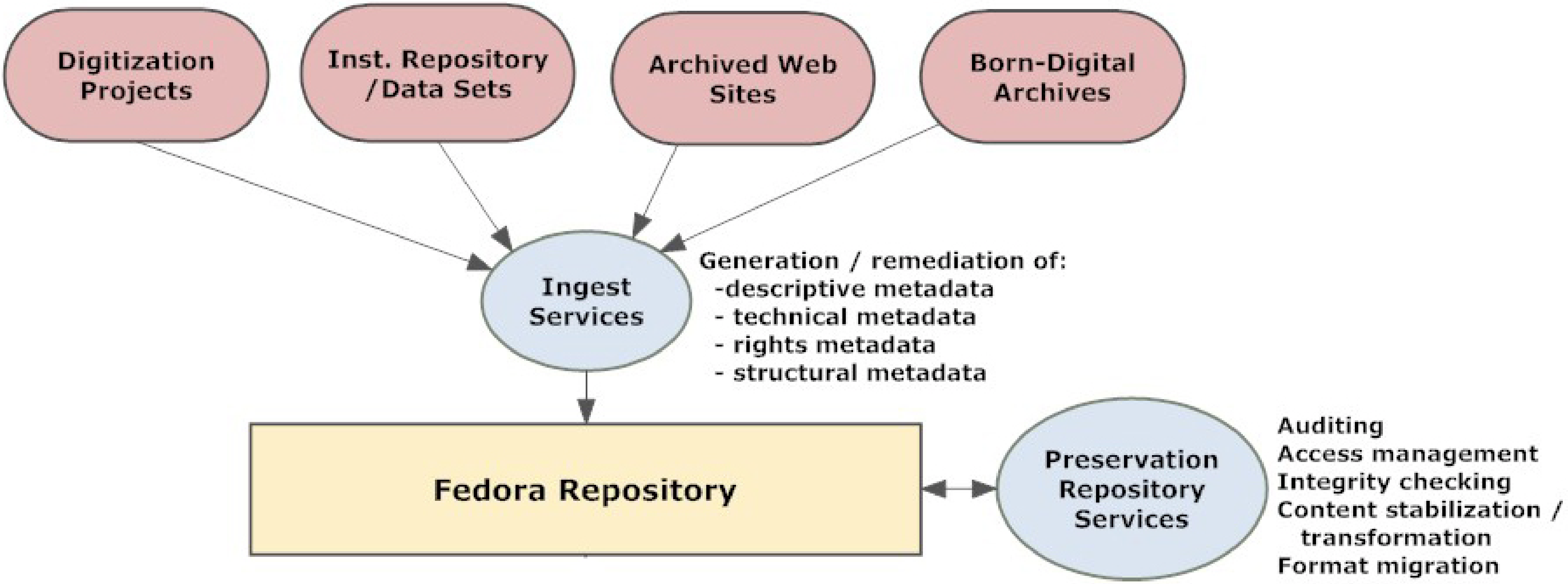
Born-Digital Archives

Columbia University Libraries / Information Services
Digital Archiving Overview



Generation / remediation of:
-descriptive metadata
- technical metadata
- rights metadata
- structural metadata

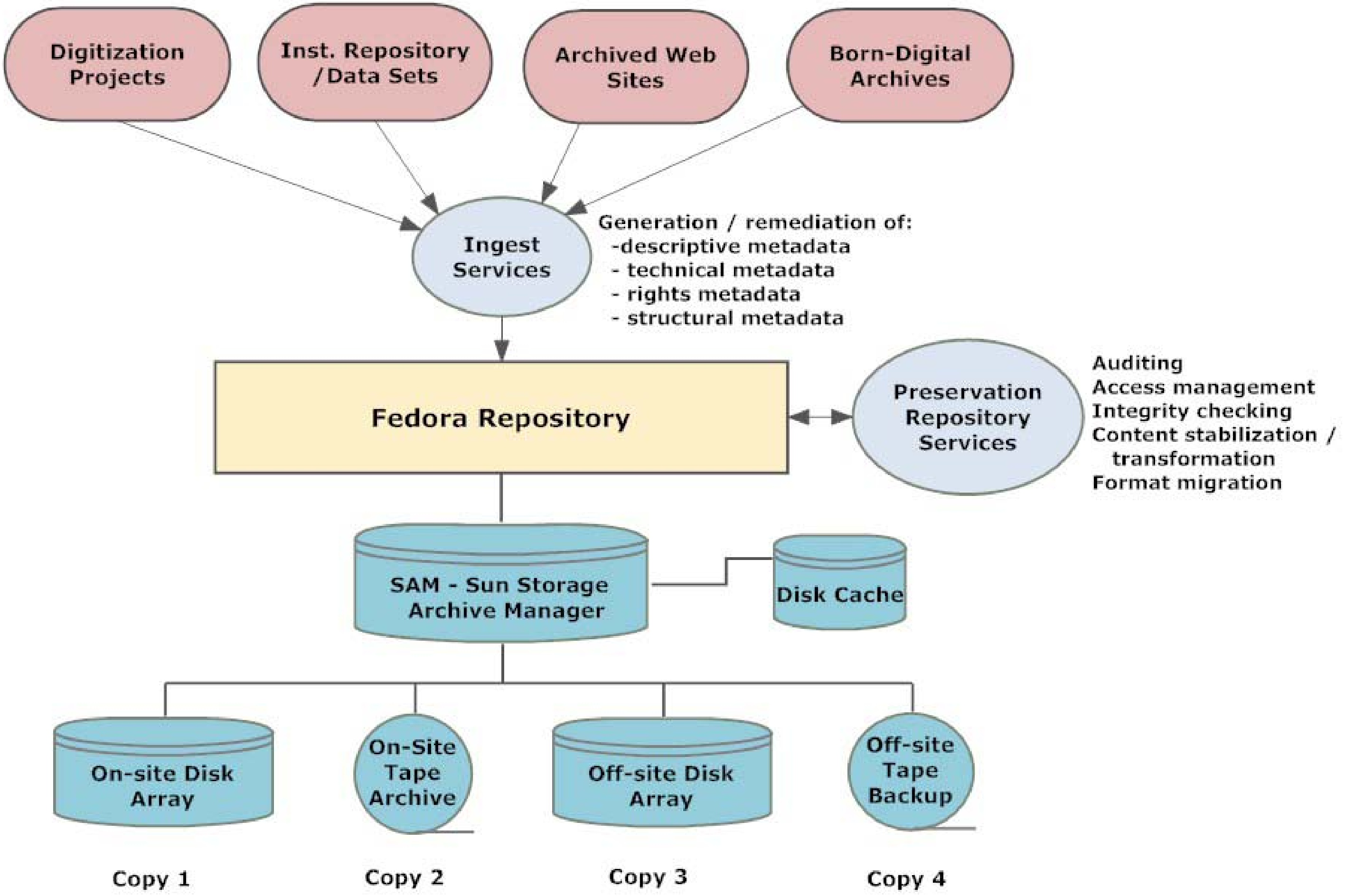
Columbia University Libraries / Information Services
Digital Archiving Overview



Generation / remediation of:
-descriptive metadata
- technical metadata
- rights metadata
- structural metadata

Auditing
Access management
Integrity checking
Content stabilization / transformation
Format migration

Columbia University Libraries / Information Services
Digital Archiving Overview



Some Digital Preservation “Macro Issues”

Who can / should preserve digital content?

- Commercial publishers
- Non-profit archiving organizations
- National libraries
- Museums
- Research universities

Who will pay to preserve?

- Businesses?
- Governments?
- Foundations?
- Research institutions?

Digital Preservation Challenges

- **Uncertainty** about selection criteria for assessing long-term value, especially with large-scale data sets, small “hand-crafted” digital collections, and the emerging genres of collective authorship on the Web;
- **Misalignment** of incentives between those who are in a position to preserve and those who benefit from preservation and access;
- **Lack of clear responsibility** for digital preservation, coupled with a prevailing assumption that it is someone else's problem;
- **Little coordination** of preservation activities across diffused stakeholder communities;
- Difficulty in separating **preservation costs** from other costs, that is, in distinguishing between the processes of making things available now and making things available in the future; and
- Difficulty in valuing or monetizing the **costs and benefits** of digital preservation, which are necessary to secure funding and investment.

The Future

- Mixed, distributed preservation environment
- Efforts to coordinate within subcommunities
- A few large trusted digital archives in the U.S. and elsewhere
- Many smaller digital archives that provide initial stabilization and packaging for later deposit in larger archives
- Much lost knowledge

What Would You Preserve ?

A. Harlem Hospital Murals

B. Encyclopedia Iranica

C. Mapping Gothic France

D. Top 100 CUL Oral Histories

The End Is Only The Beginning.

Questions:

daviss@columbia.edu

CU Libraries Digital Program

http://www.columbia.edu/cu/libraries/inside/units/ldpd/