My Journey with Preservation and Curation

Sayeed Choudhury
Grand Challenges Summit

Archive Ingest and Handling Test

D-Lib Magazine
December 2005

Volume 11 Number 12

ISSN 1082-9873

The Archive Ingest and Handling Test

The Johns Hopkins University Report

Tim DiLauro, Mark Patton, David Reynolds, and G. Sayeed Choudhury
The Johns Hopkins University
{timmo, mpatton, davidr}sayeed@jhu.edu

Introduction

From very early in its existence, the Digital Knowledge Center (DKC) in the Sheridan Libraries at Johns Hopkins University (JHU) has focused on using automated and semi-automated processes to create workflows for the creation and ingestion of digital objects. What was missing was a place to put these objects, a standard way to put them there, and a way to preserve them. This has begun to change over the past two years.

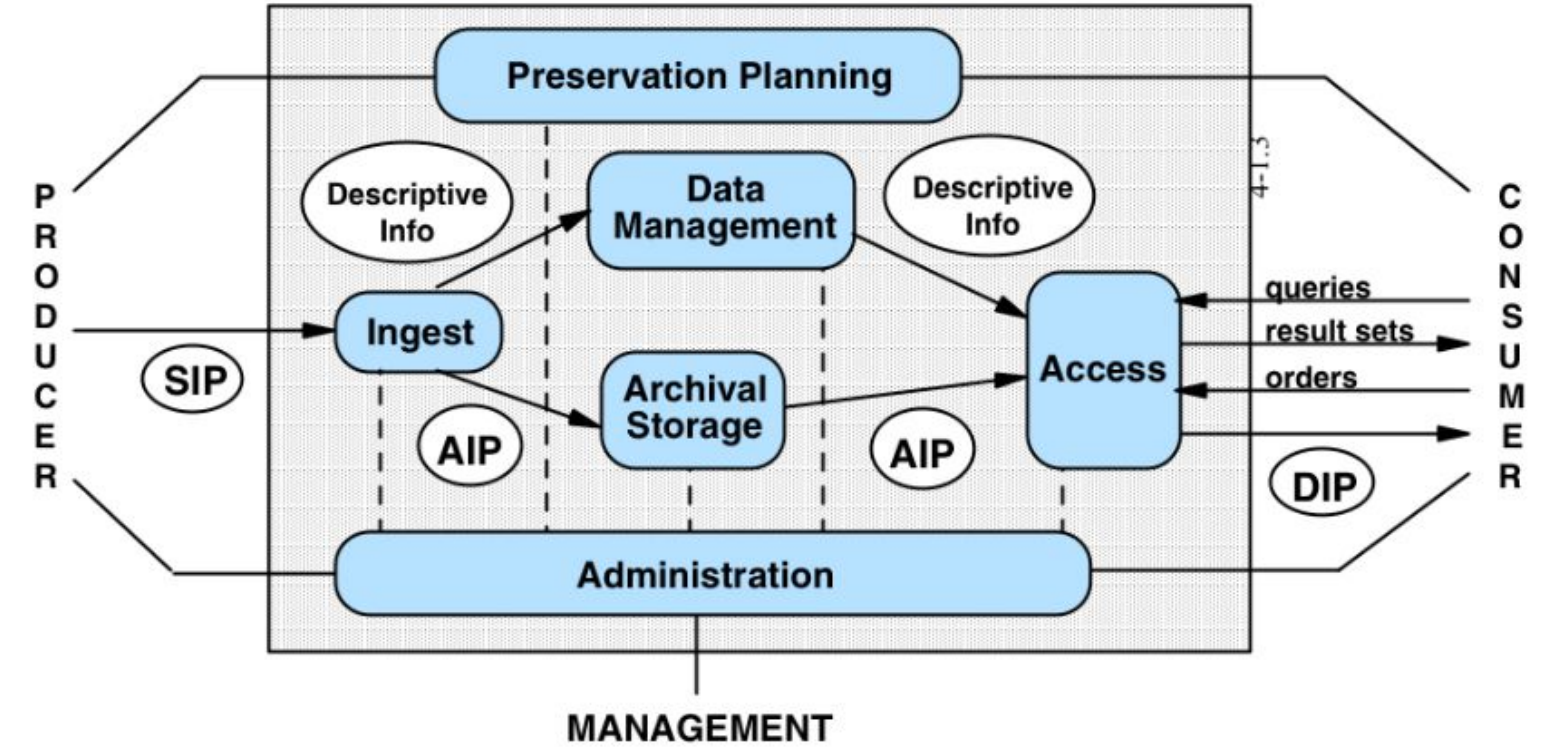
After participating in a series of workshops and meetings related to the Library of Congress' National Digital Information Infrastructure and Preservations Program (NDIIPP), we noted an emphasis on the aforementioned missing elements. When the Library of Congress (LC) announced the Archive Ingest and Handling Test (AIHT) as part of the NDIIPP, JHU saw participation as an opportunity to pursue several areas of interest.

Primary among these was the evaluation of content repositories as platforms for digital preservation. We have been concerned for some time that many in the digital library community conflate the storage of digital objects in a repository with the preservation of those objects. Participating in this test would give us an opportunity to experiment with repositories and digital preservation. JHU became especially interested in validating a level of activity that could be described as a necessary, minimal level of digital preservation.

JHU was already experimenting with Fedora and had tested ingestion of content into DSpace. The opportunity to get more hands-on experience with the facilities of these two open source efforts was an additional motivator. At a higher level, though, we were even more interested in the possibility of implementing a layer of abstraction (an application programming interface or API) over existing repository applications. Such an abstraction would allow other applications to interact with at least some facilities of a repository without knowing with which repository application (e.g., DSpace, Fedora) it is interacting. The AIHT gave us an opportunity to test the feasibility of constructing such a layer and to determine the ease with which such a layer could be applied in practice.

With funding from the Andrew W. Mellon Foundation, we are conducting a technology analysis of repositories and services to continue and build on this work. More information is available on our project wiki.[1]

OAIS Functional Entities



Data Management Layer or “Stack” Model

Levels of Services and Curation for High Functioning Data

G. Sayeed Choudhury[1], Carole L. Palmer[2], Karen S. Baker[2], Timothy DiLauro[1]

The Sheridan Library ~ Johns Hopkins University Libraries

[1] Sheridan Libraries, Johns Hopkins University

[2] Center for Informatics Research in Science & Scholarship

Graduate School of Library & Information Science, University of Illinois, Urbana-Champaign

Graduate School of Library and Information Science, The iSchool at Illinois

CIRSS Center for Informatics Research in Science & Scholarship

Introduction

The growing volume and variety of data brings new demands and opportunities. This conceptual model represents levels of data repository services and the cumulative nature of curation.

The Data Management Stack model integrates contributions from two groups within the Data Conservancy Initiative (http://dataconservancy.org):

- The Technical team and Data Management Services team at Johns Hopkins University, focused on designing and implementing systems (Choudhury & Hanisch, 2009; Mayernik et al, 2012)
- The Data Practices team at the University of Illinois, focused on social studies of data curation (Palmer et al., 2011; Weber et al, 2012).

The Model

The model represents four levels of activity and capacity shown in the center panel. It builds on definitions offered by Lord and Macdonald (2004). Today, the use of these terms, together with the notion of data stewardship (NAP, 2009), is fluid and inconsistent. Caution is advised in applying these concepts (BRTF, 2010).

Progress with Shared Vocabulary

The Stack Model has proven useful for communicating with researchers who often use

Data Management Layers

Layers	Characteristics	Implication for PI	Implication relative to NSF
Curation	• Adding value throughout life-cycle	• Feature Extraction • New query capabilities • Cross-disciplinary	• Competitive advantage • New opportunities
Preservation	• Ensuring that data can be fully used and interpreted	• Ability to use own data in the future (e.g. 5 yrs) • Data sharing	• Satisfies NSF needs across directorates
Archiving	• Data protection including fixity, identifiers	• Provides identifiers for sharing, references, etc.	• Could satisfy most NSF requirements
Storage	• Bits on disk, tape, cloud, etc.	• Responsible for: • Restore	• Could be enough for now but not

The Stack

Increasing layers of support and functionality; each level depends on the level below. (Choudhury, 2009).

- **Storage**: lowest service; basic physical storage with backup and restore services.
- **Archive**: following BRTF, "activities that enable long-term retention of digital materials"; DC focus on data protection through replication, fixity, and identifiers.
- **Preservation**: providing enough representation information, context, metadata, fixity, etc. to support use and interpretation by agents other than the original data producer.
- **Curation**: processes that add value to foster discovery and reuse.

The curation level identifies a range of services, enabling use for purposes not necessarily envisioned by the data producers.

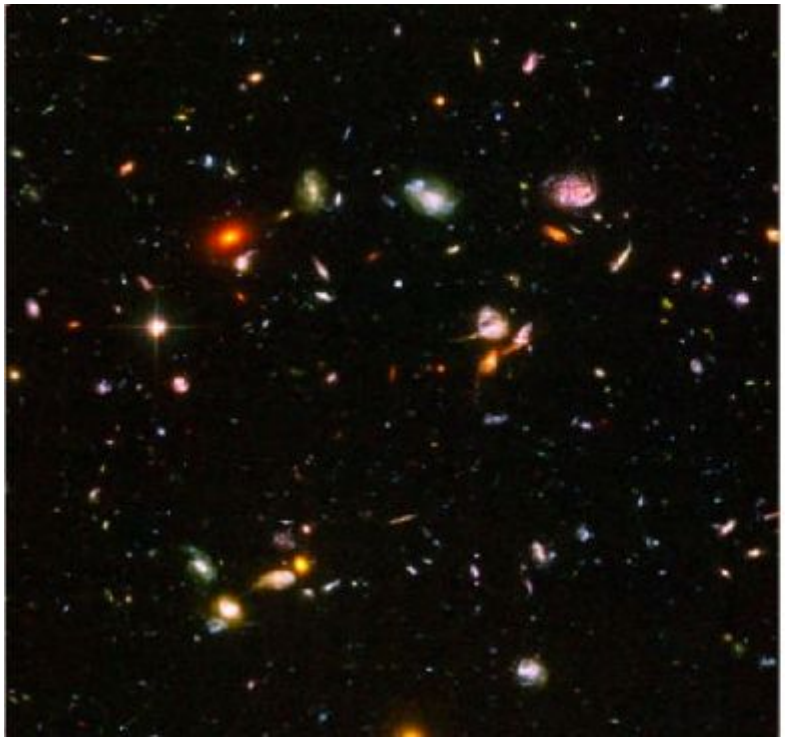
References

BRTF (2010). Blue Ribbon Task Force Report on Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information by the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

Choudhury, S. and R. Hanisch (2009). The Data Conservancy: Building a Sustainable System for Interdisciplinary Scientific Data Curation and Preservation.

Lord, P., A. MacDonald, et al. (2004). From data deluge to data curation. Proceedings of the UK e-Science All Hands Meeting, Nottingham.

Astronomy Image from Hubble Space Telescope



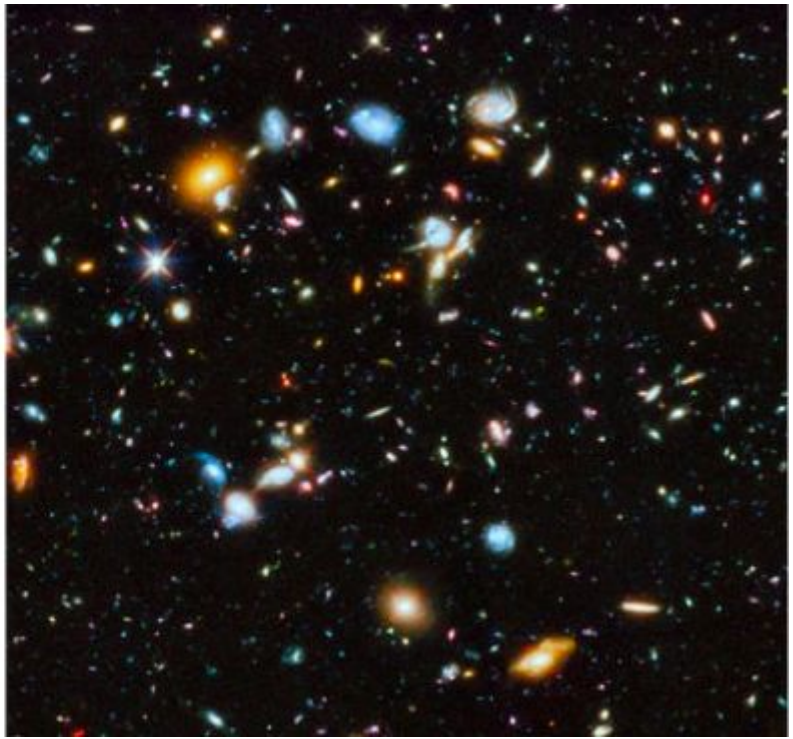
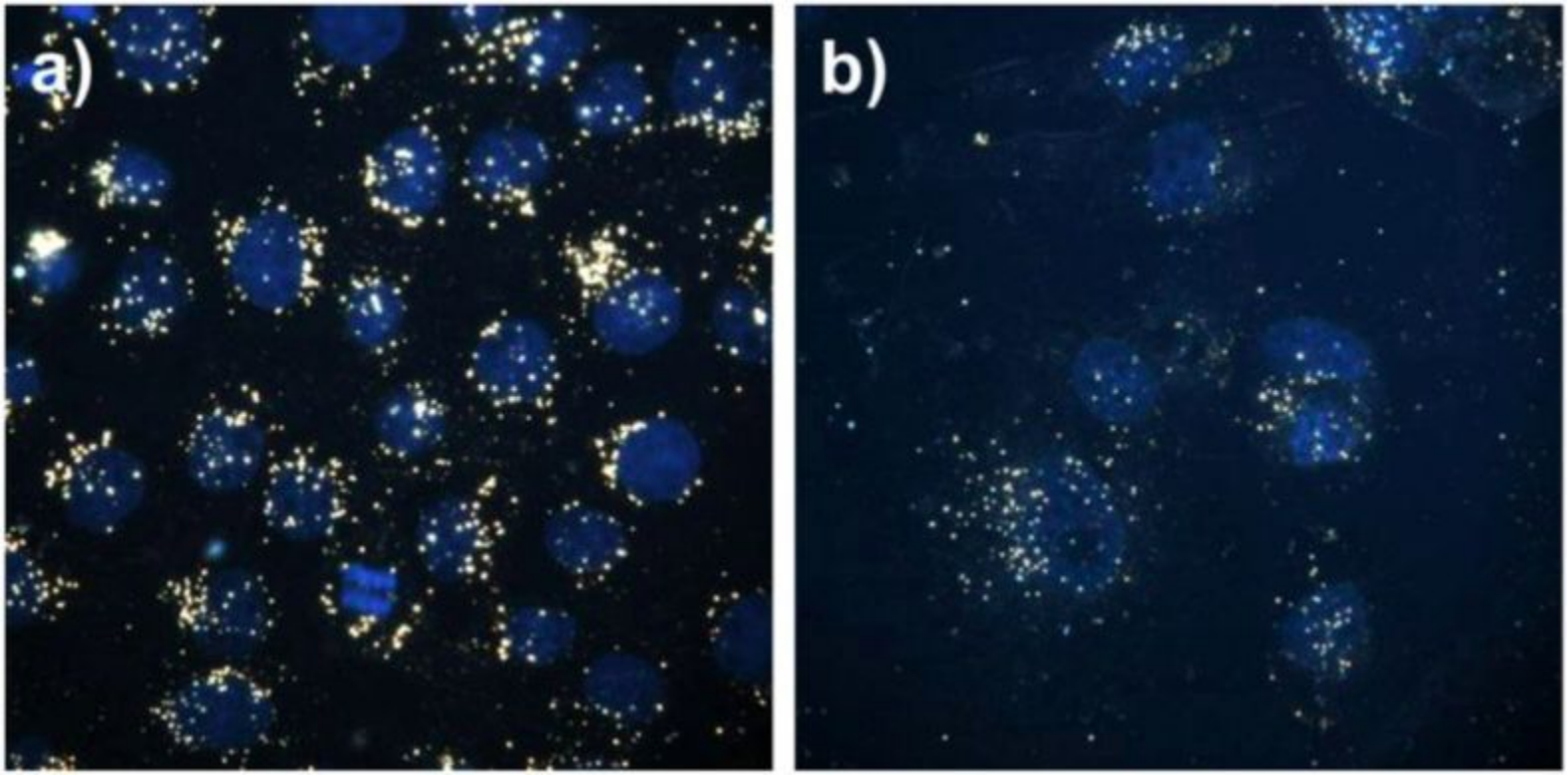
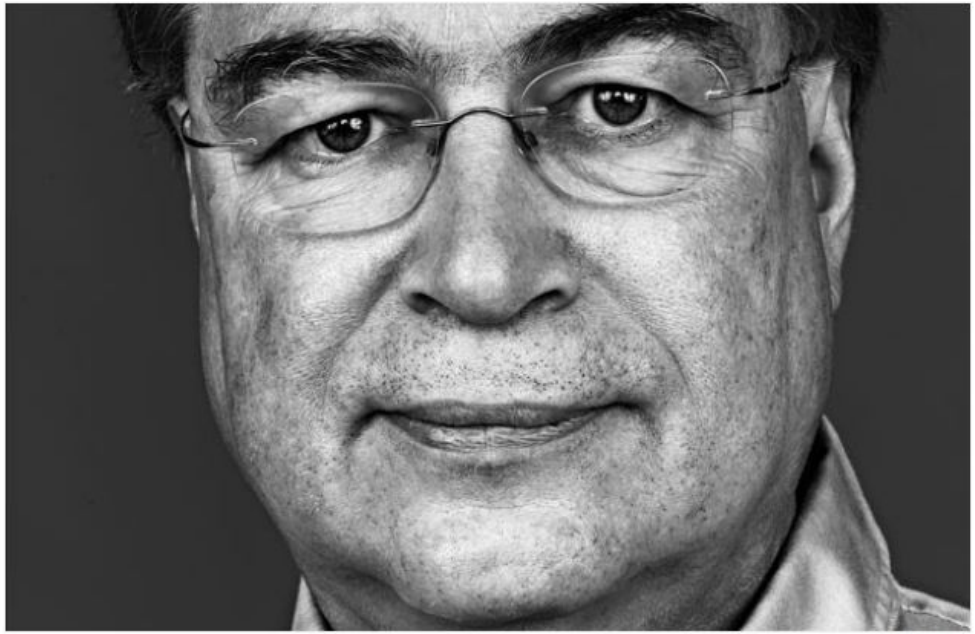


Image of Cancer Cells



The Measured Man

Larry Smarr, an astrophysicist turned computer scientist, has a new project: charting his every bodily function in minute detail. What he's discovering may be the future of health care.



Grant Delin

MARK BOWDEN | JULY/AUGUST 2012 ISSUE | TECHNOLOGY

Definition of Preservation -- 2010

“Data preservation involves providing enough representation information, context, metadata, fixity, etc. such that someone other than the original data producer can use and interpret the data -- without having to contact the original data producer.”

-- Sayeed Choudhury

Don’t Panic

“We need to become comfortable with the idea that not everything will not done.”

-- Kate Zwaard

Who Preserves?

Internet Archive Blogs

A blog from the team at archive.org

Blog | Announcements | Internet Archive Store | archive.org | About | Events | Developers | Donate

← 27 Public Libraries and the Internet Archive Launch "Community Webs" for Local History Web Archiving

10 Ways To Explore The Internet Archive For Free →

Andrew W. Mellon Foundation Awards Grant to the Internet Archive for Long Tail Journal Preservation

Posted on March 5, 2018 by jefferson

The Andrew W. Mellon Foundation has awarded a research and development grant to the Internet Archive to address the critical need to preserve the "long tail" of open access scholarly communications. The project, *Ensuring the Persistent Access of Long Tail Open Access Journal Literature*, builds on prototype work identifying at-risk content held in web archives by using data provided by identifier services and registries. Furthermore, the project expands on work acquiring missing open access articles via customized web harvesting, improving discovery and access to this materials from within extant web archives, and developing machine learning approaches, training sets, and cost models for advancing and scaling this project's work.

Search

Recent Posts

- Let's Build a Great Digital Library Together...Starting with a Wishlist
- TV News Record: Glorious ContextuBot making progress
- Archive video now supports WebVTT for captions
- 10 Ways To Explore The Internet Archive For Free
- Andrew W. Mellon Foundation Awards Grant to the Internet Archive for Long Tail Journal Preservation

Recent Comments

- noklion on 10 Ways To Explore The Internet Archive For Free
- ایکا on 10 Ways To Explore The Internet Archive For Free

Definition of Preservation -- 2014

“For all my thoughts about data, engineering, or scholarship, ultimately preservation is about keeping people’s stories alive for future generations.”

-- Sayeed Choudhury

Definition of Curation -- 2018

“For all my thoughts about data, engineering, or scholarship, ultimately curation is about keeping people’s stories alive for future generations.”

-- Sayeed Choudhury

RMap -- Linked Data Protocol and Service