The Stanford Digital Repository

Tom Cramer
Chief Technology Strategist
Stanford University Libraries
May 2011
Agenda

- Purpose of PASIG
- Stanford Use Cases
- Stanford Ecosystem
- SDR 2.0 - Preservation Core
- Observations, lessons, next steps
Why PASIG?

- A practical forum for practitioners in the preservation & archiving space
- Exchange of requirements, designs, specifications, architectures & practices
- Cross-pollination of technology industry experts and preservationists
SDR is...

- in production since Dec 2006
- a second generation preservation system
- one component in a larger ecosystem of digital library infrastructure
SDR today

- 97+ TB of unique content
- 300+ TB of managed data
- 200,000+ objects
- 62,000,000 files
- 7 content types: books, images, audio, video, manuscripts, GIS data, software
- 5 storage migrations
Three Major Areas of Preservation Needs

• Digital Library
  - Legacy collections
  - Digitized collections
  - Licensed, locally loaded content
  - Born digital collections

• Institutional Repository
  - Research data,
  - Publications, dissertations,
  - Learning objects, university assets

• External Depositors
  - Publishers
  - Discipline-specific repositories
  - Reciprocal deposits with peer institutions

Summary of Digital Collections Size:

- Google Books: (’00s of TB)
- Manuscripts: (75 TB)
- Media: (50 TB)
- Geospatial Data: (10 TB)
- ~30 other digi projects: (15 TB)
- Purchased collections: (25 TB)
E.g., Google-Scanned Books

Download, process and preserve 8 million volumes in SDR for...

• local indexing,
• text mining,
• selective delivery, and
• long-term access.
E.g., Monterey Jazz Festival

• Festival founded in 1958: longest running jazz festival in the world.

• Rich collection of recordings from inception, spanning over 50 years, in varying states of condition & decay.

• Archives held at Stanford’s Archive of Recorded Sound

• ~800 audio recordings, 1.6 TB audio files in SDR

• ~250 video recordings, 22 TB video files in SDR

Access:
- complete database of digital recordings online at collections.stanford.edu/mjf
- Access via in-site visit to ARS
- New commercial releases on MJF Records
E.g., National Geospatial Digital Archive

- Some 27,000 “at risk” geospatial objects
- TIFFs, GeoTIFFs, Shapefiles, Digital Elevation Models, Digital Orthophoto Quadrangle files
E.g., Preserving Virtual Worlds

Stanford University Libraries
Second Life Open House,
31 July 2009
E.g., Forensically Extracted Born Digital Files

- Digital Forensics lab extracting original computer files from legacy media
- Actively building pipeline from extraction to preservation store
- Support for both immediate and deferred archival processing & description
E.g., Electronic Theses and Dissertations

Submit your dissertation to the Stanford Digital Repository
Dissertation ID: 0000000060

Read-only administrative view
This is a read-only view of the student's ETD submission

Submission approved
Your dissertation has been accepted by the University and is currently being processed by Stanford University Libraries.

» Download your submitted file (including the copyright and signature pages).

When processing is complete and it is released for online access, the dissertation will be accessible from http://purl.stanford.edu/rb768bf2323. External access (i.e., from outside Stanford) to the link may be limited pending any release delay applied at the time of submission.

1 Verify your citation details
This information establishes how your work will be cited, and how you will be credited.

Confirm: Yes, the information in this section is correct and complete.
If the information that appears here is incorrect, please consult with your department in order to correct the data in Axess.

Progress
- Citation details verified
- Abstract provided
- Dissertation uploaded
- Copyrighted material checked
- License terms applied

- Submitted
04-Dec-2009 02:52 PM

- Verified by Final Reader
04-Dec-2009 03:50 PM

- Approved by Registrar
04-Dec-2009 03:54 PM

Congratulations!
NSF Policy Position on Data Archiving

“NSF's policy position on data is straightforward: all science and engineering data generated with NSF funding must be made broadly accessible and usable, while being suitably protected and preserved.”

---

NSF and NIH Grants to Stanford
Preservation Is One Leg of a Stool

- Preservation without Access is pointless
  - Further, all signs points indicate that it is not economically viable

- Access without Preservation is myopic

- Robust Management services are prerequisite for accessioning, archiving and providing access to content
  - The “pre-ingest” phenomenon

Can one system handle it all? or
Stanford’s Digital Library Ecosystem

SDR Preservation Core

Digital Object Registry (DOR)

ETDs

SALT

EEMs

Digitization Workflow

Google Books

Symphony ILS

SearchWorks

Socrates

Text

Images

Media

Files

Data

“Digital Stacks” Delivery Systems

SDR Preservation Core
Three Spheres: Management, Preservation and Access

Digitization, Deposit & Management
- EEMs
- ETDs
- SALT
- Digitization Workflow
- Google Books
- Symphony ILS

Discovery & Delivery
- Digital Object Registry (DOR)
- SearchWorks
- Socrates
- "Digital Stacks" Delivery Systems
- Text
- Images
- Media
- Files
- Data

Preservation
- SDR Preservation Core
SDR in Stanford’s DL Ecosystem

Specialty applications provide context-specific, user-facing deposit, and access services tailored to content types and disciplines

- **Library Management Applications**
  - EEMS (acquiring born digital content), digitization workflow, etc.

- **Institutional Repository**
  - ETDs, open access articles, faculty “papers”, research data, web sites, etc.

- **National Geospatial Digital Archive (NGDA)**
  - Geospatial data

- **SULAIR Digital Stacks**
  - Delivery for text, images, mss, media, data, & curated collections

---

**Stanford Digital Repository (SDR):** content agnostic, preservation repository

and SDR provides “back-office” preservation services: replication, auditing, migration, and retrieval in a secure, sustainable, scalable stewardship environment
E.g., Parker Manuscripts

- 559 Anglo-Saxon manuscripts, 200,000 pages

- For each page:
  - 22 MB JPEG2000 delivery surrogate
  - 22 MB JPEG2000 delivery surrogate
  - 110 MB submaster TIFF
  - 220 MB master TIFF

Parker.stanford.edu:
Rich web application, tailored for general public, medievalists

SDR – Preservation Core
Separation of Concerns

• Scoped repository: differentiation between preservation (provided by SDR) and content management (provided by DOR) access (provided by the Digital Stacks apps)

• Implications:
  - Reduces pressure on SDR to be all things to all depositors, for all content
  - Reinforces need to provide managed & secure storage at scale
  - Reinforces requirement to focus on content integrity services
  - Emphasizes need to integrate SDR to management & access services through stable API’s
SDR 2.0: New Technical Architecture
SDR 2.0: New Technical Architecture

- Adopt Fedora as a metadata management system
  - Clean mapping of new data model to Fedora content models
  - Reuse same design pattern, core technology as in DOR
- Support for parallelized & asynchronous operations
  - Multiple ingest streams to increase throughput
  - Decompose one process (e.g., “ingest”) into discrete, loosely coupled operations (“checksum”, “package”, “transfer”)
- Adopt a RESTful architecture & common workflow service
### Workflow: googleScannedBookWF

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Register new objects in DOR</td>
</tr>
<tr>
<td>2</td>
<td>Request GRIN items to be converted for download</td>
</tr>
<tr>
<td>3</td>
<td>Make descriptive metadata (MODS &amp; DC) from Symphony/MARC</td>
</tr>
<tr>
<td>4</td>
<td>Download content from Google</td>
</tr>
<tr>
<td>5</td>
<td>Process content, create tech/rights/provenance/content metadata</td>
</tr>
<tr>
<td>6</td>
<td>Shelve contents in the Digital Stacks</td>
</tr>
<tr>
<td>7</td>
<td>Prepare bagit transfer package, push to SDR-Stage</td>
</tr>
<tr>
<td>8</td>
<td>Call from SDR on completed ingest</td>
</tr>
<tr>
<td>9</td>
<td>Cleanup workspace; make room for more!</td>
</tr>
<tr>
<td>10</td>
<td>Call from SDR on completed archiving</td>
</tr>
</tbody>
</table>
Management: Hydra-based Applications

Under Development...

- SDR’s Front End – Institutional Repository for Stanford
- Hypatia – Archival Arrangement, Description & Access
- SDR Preservation Core Administrative Application

ETD’s – Electronic Theses & Dissertations

SALT – Self-Archiving Legacy Toolkit

EEMs – Everyday Electronic Materials
Hydra for Repository Management

- Reusable application framework that sits on top of Fedora – providing CRUD (create, read, update, delete) apps
- Open source application with dozens of contributors
- Basis for Stanford’s user-facing (institutional) repository services (electronic theses, open access articles, data deposit, etc.)
- Interoperates with workflow, robot and web service framework for process orchestration
- Also the basis for an SDR administrative user interface (under development)
SDR 2.0: Revised Data Model

SDR 1.x’s METS-based SIP, AIP and DIP, had many issues:

- Each Transfer Manifest was content & collection specific → Doesn’t scale
- Transfer manifests require too much interpretation and analysis to change, augment
- Too complex: Stanford METS structure breaks apart related data across the object
- Wraps (somewhat dynamic) metadata with (mostly static) data files in same envelope
- Recursive nature of transfer manifest makes versioning self-referential, complex
- No one speaks METS natively: depositors, SDR & clients all forced to perform translation at handshakes
Content Structures and Flavors of Metadata

- Flexible data model can take any type of data, packaged in “bags”
  - A “bag” is a directory with standardized top-level structure and syntax

- Minimizes analysis & processing required on ingest

- Preserves options for future processing & transformations based on future needs

Each object has seven discrete metadata files:
- *Identity* metadata
- *Descriptive* metadata
- *Content* metadata (aka structural metadata)
- *Technical* metadata
- *Rights* metadata
- *Source* metadata
- *Provenance* metadata
SDR Deposits: Content Transfer via Bagit

druid
/bagit-info.txt
:
  Stanford-Content-Metadata: data/metadata/contentMetadata
  Stanford-Identity-Metadata: data/metadata/identityMetadata
  Stanford-Provenance-Metadata: data/metadata/provenanceMetadata
/data
/metadata
  /contentMetadata
  /descMetadata
  /identityMetadata
  /provenanceMetadata
  /rightsMetadata
  /sourceMetadata
  /technicalMetadata
/content
  /file1
  /file2
  :
Lessons Learned Over 5 Years

• Custom code, maintained by evolving & smaller team, was inefficient & unsustainable
  – Adopted Fedora for metadata management, Hydra for application framework
  – Shared technology & design patterns with rest of digital library ecosystem
  – API’s for management, ingest, retrieval, reporting

• Bottlenecks
  – Need to be quicker to add new content types & collections: simplify the data model, support “Zip & SIP”
  – Need to increase the throughput to the storage layer led to parallelization of processes

• Need to refine & hone the SDR service model
  – Need to deemphasize curation approach, shift away from “just in case” to a “just in time” mentality
Next Steps

• Scale to 1 PB
• Migrate legacy content and SDR 1 content to new environment
• Relate & manage preserved objects through RDF
• Develop robust management, administration and workflow UI’s via Hydra
• More granular encapsulation of work into “robots” and web services
• Defining meaningful provenance and rights metadata for coherent audit trails and access management
Growth in Disk and Computing at SULAIR