The Stanford Digital Repository (SDR)

PASIG3 – May, 2008
Three Major Areas of Preservation Needs

• Digital Library
  – SULAIR collections & resources
  – Digitization artifacts

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

• External Depositors
  – Online preservation and access
  – Dark archive

<table>
<thead>
<tr>
<th>External Depositors</th>
<th>Storage Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google Books</td>
<td>(‘00s of TB)</td>
</tr>
<tr>
<td>Parker Manuscripts</td>
<td>(75 TB)</td>
</tr>
<tr>
<td>MJF Media</td>
<td>(50 TB)</td>
</tr>
<tr>
<td>NGDA</td>
<td>(10 TB)</td>
</tr>
<tr>
<td>~30 other digi projects</td>
<td>(15 TB)</td>
</tr>
<tr>
<td>Purchased collections</td>
<td>(25 TB)</td>
</tr>
</tbody>
</table>
The Stanford Digital Repository (SDR) provides services to help ensure that digital information is available over the long term with fixity, intelligibility, and authenticity.

To fulfill its mission, the SDR must be secure, sustainable and trusted.
Design Objectives & Assumptions

• Preservation-focused archive
• Replicated content
  – multiple copies, geographically distributed
• Auditable
• Simple
• Modular
• Tiered storage environment
  – online, nearline, offline
• Version rather than delete
• Content-agnostic
SDR Serves As Common Preservation Infrastructure
SDR Serves As Common Preservation Infrastructure

while specialty archives and applications provide focused digital content collection, access and value-added services

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

- **SUL Digital Bookshelves**
  - (Google Books, internally digitized, vendors' e-books)

- **Digital Library Applications**
  - (images, mss, media, Special Collections showcases)

- **Institutional Repository**
  - (faculty- and student submitted papers, data, websites, etc.)

---

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

• Preserving access to digital information over time  
  ...through generations of technology obsolescence and change.

• Maintaining integrity of that information over time  
  ...through generations of migration and reformatting.

Repository Services Functionality

• All (or almost all) user-facing services
• Enhanced access & delivery through applications
• Data mining, dry research, new indexing, e-science, etc.
• Federation
E.g., The Parker Project

• 525 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
SDR Component Diagram
SDR Storage Platforms

Honeycomb

Tape
On the Near Term Horizon for SDR

- Parallelize workflow to support multiple, simultaneous content streams
- Deploy & integrate with repository services for deposit & access
- Build out more robust administration modules
- Build out more robust access modules
- Upgrade storage subsystems
  - Storage abstraction layer
  - New tape infrastructure (robot & library)
- Initiate next wave of audits
Questions for PASIG Members

• What designs can we copy, what code can we adopt for repository services such as deposit & access?

• (When) will we need to evolve our metadata requirements as our queues grow and content types flourish?

• What administrative functions do we need, but haven’t yet built out (or envisioned)?

• What type of auditing procedures do we need to put in place?
Bootcamp Lessons: Getting Started

• Be clear about your objectives
  – Preservation or access?
  – Map out the enterprise-level architecture

• Start a program, not a project
  – Sustainability is paramount
  – Staff, funding & systems
  – Create a full team: client services, development, operations, sys admins

• Keep it modular & keep it simple
  – Everything will evolve
  – “Building tomorrow’s legacy code today”