The Stanford Digital Repository (SDR), Fedora & Hydra
SDR is...

- a core Library service
- Stanford’s digital preservation system
- an increasingly robust digital asset management & access system
- in production since Dec 2006

http://library.stanford.edu/sdr
SDR Preservation Core

The Stanford Digital Repository (SDR) provides services to make scholarly resources available over the long term by helping ensure their integrity, authenticity, and reusability.

To fulfill its mission, the SDR must be secure, sustainable and trustworthy.
The Stanford Digital Repository

Contents as of Sept 2014

Management

Access

Preservation

Stanford Digital Repository

536k objects
193M files
616 collections
153 TB

Audit and Fixity
Replication
Storage
SDR Preservation Core

All Stats as of 9/2014
E.g., Google-Scanned Books

Download, process and preserve scanned 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
Technology

IBM Enterprise Tape Library

NetApp 6080 Disk
Hydra-based Applications at Stanford

ETD's – Electronic Theses & Dissertations

SALT – Self-Archiving Legacy Toolkit

EEMs – Everyday Electronic Materials

Argo – Repository Reporting and Management

Hypatia – Archives & Special Collections

SDR – Web UI (aka Hydrus)
Access Environments

PURL (Persistent URLs) Landing Pages

Image Galleries

Embeddable Web Widgets

International Image Interoperability Framework

http://iiif.io
PURLs (Persistent URLs)
Stanford Digital Access Services

Discovery
- SearchWorks
  - Bibliographic info
  - Embedded TOC and rendering
- Landing page
  - Table of Contents
  - Embedded rendering

Published metadata
- SOLR
- SOLR
- SOLR
- Index
- Public XML document cache
  - Privileged views

Index
- MARC
- DOR
- objects
  - DC/MODS, identity, rights & content MD
- policy

Published content
- Digital Stacks datastore

Delivery
- Rendering applications -- using contentMetadata
  - Book reader
    - a real page turner
  - Image viewer
    - photos, maps, mss
  - Media player
    - audio & video

Access services -- apply authorization per rightsMetadata, fetch & deliver content
- File Service
- Image Service
- Streaming Service
- Image Server
  - djakota
- Media Server
  - ???

links to
content metadata
- rights metadata
fetches from
direct access to digital stacks
full text
content

STANFORD UNIVERSITY LIBRARIES
## Standard Data Streams

<table>
<thead>
<tr>
<th>DOR datastream</th>
<th>Source</th>
<th>Description</th>
<th>Fedora Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>identityMetadata</td>
<td>Stanford</td>
<td>Core identifying information: title, creator, source and other identifiers, tags, etc.</td>
<td>Internal XML</td>
</tr>
<tr>
<td>descMetadata</td>
<td>Hydra</td>
<td>Currently MODS xml</td>
<td>Managed Content</td>
</tr>
<tr>
<td>technicalMetadata</td>
<td>Hydra</td>
<td>JHOVE generated metadata describing the content files that make up the digital object.</td>
<td>Managed Content</td>
</tr>
<tr>
<td>sourceMetadata</td>
<td>Hydra</td>
<td>Information identifying the source of the scanned object</td>
<td>Managed Content</td>
</tr>
<tr>
<td>rightsMetadata</td>
<td>Hydra</td>
<td>Holds human and machine readable expressions of Access and Use permissions/restrictions.</td>
<td>Managed Content</td>
</tr>
<tr>
<td>provenanceMetadata</td>
<td>Hydra</td>
<td>Contains a block of provenance information for each agent that handled the record.</td>
<td>Managed Content</td>
</tr>
<tr>
<td>contentMetadata</td>
<td>Hydra</td>
<td>Information about object structure, file sequence and type, and checksums.</td>
<td>Managed Content</td>
</tr>
<tr>
<td>RELS-EXT</td>
<td>Fedora</td>
<td>RDF triples describing selected object attributes and inter-object relationships.</td>
<td>Internal XML</td>
</tr>
</tbody>
</table>
## Standard Data Streams - continued

<table>
<thead>
<tr>
<th>DOR Datastream</th>
<th>Source</th>
<th>Description</th>
<th>Fedora Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>versionMetadata</td>
<td>Stanford</td>
<td>A summary history of versions that have been accessioned/archived</td>
<td>Internal XML</td>
</tr>
<tr>
<td>workflows</td>
<td>Stanford</td>
<td>The workflows datastream is a view into all the workflows defined for an object, current and past. It is a concatenation of individual workflow XML in a <code>&lt;workflows&gt;</code> document container.</td>
<td>Redirect Referenced Content</td>
</tr>
<tr>
<td>DC</td>
<td>Fedora</td>
<td>Dublin Core XML, reflects author, title and identifier information here to take advantage of Fedora indexing for their administrative tools only.</td>
<td>Internal XML</td>
</tr>
<tr>
<td>embargoMetadata</td>
<td>Stanford</td>
<td>Operational metadata for managing embargo settings, extensions and release.</td>
<td>Internal XML</td>
</tr>
<tr>
<td>events</td>
<td>Stanford</td>
<td>A lightweight history of events relating to the management of the object, e.g., embargo extensions and release.</td>
<td>Internal XML</td>
</tr>
<tr>
<td>varies</td>
<td>Stanford</td>
<td>Applications may have their own datastream(s) to manage local settings and other pertinent information.</td>
<td>Internal XML</td>
</tr>
</tbody>
</table>
Basic Object Model
What Is Hydra?

• A robust repository fronted by feature-rich, tailored applications and workflows (“heads”)
  ➤ One body, many heads

• Collaboratively built “solution bundles” that can be adapted and modified to suit local needs.

• A community of developers and adopters extending and enhancing the core
  ➤ If you want to go fast, go alone. If you want to go far, go together.
A Tale of Two Worklites

Stephen Jay Gould

Museum of Comparative Zoology

Harvard University

Cambridge, MA 02138

Christopher Wren, the leading architect of London's reconstruction after the great fire of 1666, lies buried beneath the floor of his most famous building, St. Paul's cathedral. No elaborate sarcophagus adorns the site. Instead, we find only the famous epitaph written by his son and now inscribed into the floor: A monument requirs, says Wren — if you are searching for his monument, look around. A tall grandson perhaps, but I have never read a finer testament to the central importance — one may even say sacredness — of actual places, rather than relics, symbols, or other forms of vicarious resemblance.

An odd coincidence of professional life recently turned my thoughts to this most celebrated epitaph: after, for the second time, I received an office in a spot laden with history, a place still redolent with ghosts of past events both central to our common culture and especially meaningful for my own life and choices.

In 1973, I spent an academic term as a visiting researcher at Oxford University. I received a cranny of office space on the upper floor of the University Museum. As I set up my books, fossils, smalls, and microscope, I noticed a metal plaque affixed to the wall, informing me that this reconfigured space of shelves and cubicles had been, originally, the site of the most famous public confrontation in the early history of Darwinism. On this very spot, in 1860, just a few months after Darwin published the Origin of Species, T.H. Huxley had drawn his rhetorical sword, and soundly skewered the slick but superficial champions of creationism, Bishop "Napier Saint" Wilberforce.

As with most legends, the official version ranks as mere carthorse before a much more complicated and multifaceted truth. Wilberforce and Huxley did put on a splendid, and largely spontaneous, show — but no clear victor emerged from the scuffle and Joseph Hooker. Darwin's other champions, made a much more effective reply to the Bishop, however forgotten by history. See my essay on this debate, entitled "Knight takes Bishop!" and published in an earlier volume of this series: Biology for Bronnmonsia.

I can't claim that the lingering presence of these Victorian giants increased my resolve or improved my work, but I loved the sense of continuity conveyed to me by this happy circumstance. I even treated the controversy — for "circumstance" means "standing around" to me.
Point Solution Approach ...Welcome to Siloville

ETDs (Theses) → ETD
Books, Articles → IR
Images → Image DB
Audio-Visual → DAM
Research Data → ?
Maps & GIS → Geospatial Inf.
Documents → Records Mgmt.

Management  Access  Preservation(?)
Effective?  Sustainable?
Repository-Powered Approach

ETDs (Theses) | Books, Articles | Images | Audio-Visual | Research Data | Maps & GIS | Documents

Digital Repository

Scalable, Robust, Shared Management and Preservation Services
One Body, Many Heads...

- ETDs (Theses)
- Books, Articles
- Images
- Audio-Visual
- Research
- Maps & GIS
- Documents

Scalable, Robust, Shared Management and Preservation Services
Hydra Heads of Note

Avalon & HydraDAM for Media

UCSD DAMS

BPL Digital Commonwealth

Sufia

Northwestern Digital Image Lib.
Hydra Partners and Known Users

OR = Open Repositories Conference
Hydra Technical Framework
CRUD in Repositories

Create/Submit/Edit (CUD)  
Search/View (R)  
Repository/ Persistent Storage
CRUD in Repositories

Create/Submit/Edit (CUD)

Search/View (R)

Repository/Persistent Storage
Major Hydra Components

- hydra-head
  - Rails Plugin
    - (CUD)
- Blacklight
  - (Read Only)
- Fedora
- Solrizer
- Solr
The Hydra Stack

Browser/Client Side

HTML, CSS & JavaScript or REST Client Applications

Hydra REST API

Custom Controllers

Custom Views

Blacklight Controllers

Blacklight Views

Hydra Access Controls

Ruby on Rails (Server)

Custom Models

OM

ActiveFedora

Rubydora

Solrizer

Blacklight Search & Facet Logic

Fedora REST API

Solr REST API

Java & SQL (System Services)

Fedora

Solr

User Model

User & Group Solution (SQL, LDAP, Shibboleth, etc.)
Hydra’s Core Gems

Rails engine with “everything you need to build a user-facing app”

DSL (domain specific language) for easy mgmt of XML metadata

Applies ActiveModel pattern to working with Fedora objects

Client API for consuming Fedora’s REST API via Ruby
Three Active Hydra-Fedora 4 Pilots

• Penn State University / ScholarSphere

• UCSD DAMS

• Stanford / Annotation Management
Project Cerberus at Stanford

- Bib records (Bibframe)
- Local authority file (person, org, place, subject, titles)

• Biblio & Entity Graph

• Digital Objects (SDR)

• Annotations

Open Annotation-based assertions on other entities...
- Virtual collections
- Tags
- Commentary
- Translation
- Etc.

- Books
- Images
- Manuscripts
- Media
- Software
- Etc.
Project Cerberus at Stanford

SDR -> BibFrame
CAP -> Person
MARC (Bib) -> MARC (Auth)
MARC XML -> BibFrame

BibFrame

HighWire

Inventory all data sets, analysis of entities in data sets, analysis of conversion processes, analysis of target ontologies, establish strategy for relating data across entity banks & across institutions, begin conversions, begin data loading, begin data relating, wash rinse repeat, develop services/app that sit on top for value add (??) -- what are these use cases?

Pile o' Bibframe
- OR-
Post Bibframe BibRDF with robust converters/deconverters

Fedora 4? Possibly, for scale testing. No need for UI. Probably just use a triple store here or a file system.

Fedora 4 seems like a perfect fit. Trigops link things together. Elasticsearch stores stories for body of anno. Nice API that is LDP compliant. Separate from DOR and ILS logically and almost certainly physically. Supports L4L UC1 & DMS Tech.

Use out of the shelf tools like annotator or hypothesis as UIs.

DMS Tech use case is indexing, search and display of anno. (Also LD4L?) Prop BL on top of 3rd party

Comment Tag Review

Annotat'n

Graph Search

Applicat'n

Anno Store

Authority

MARC

pre-process

post-process

converter

marc2xml

pre-process

converter

marc2xml

Install Fedora 4, supports management of annotations for DMS Tech and LD4L (anno + body). Integration of annos into one library/DLSS system (BL, Mirador, DMS Index, SW). Integration of front end to write annos into Anno Store. Possible Resources: Rob, ND, WM, Drew (Mirador), Cabeer, Ben.