Defining and Designing a Cyberinfrastructure for the Library of the Future

Tom Cramer, Lynn McRae, Rachel Gollub
Digital Library Systems & Services
Stanford University
Fall DLF, Philadelphia, 2007
The Digital Library: Content, Services, Infrastructure

- **Ebooks**
- **Google Books**
- **Digitized mss, texts, images, media**
- **Born-digital materials (data sets, theses, articles, new media, etc.)**

**Content**

- **Digital preservation infrastructure**
- **Content management & content middleware services**
- **Security, DRM**
- **Server, storage and data center facilities**

**Infrastructure**

- **Discovery**
- **Delivery**
- **Use**
  - **Analysis**
  - **Annotation**
  - **Citation**
  - **Collaboration**
  - **Mining**

**Services**
Another View of Underlying Infrastructure

Digital Content is the raw material for 21st century research & teaching.

Information Services give scholars the prerequisite tools and environments to work with digital materials.

Underlying Infrastructure supports content acquisition and user-facing information services.
Environmental Scan at SULAIR

1. Digital preservation
   • SDR (the Stanford Digital Repository)
   • NGDA

2. Google Book Search scanning

3. Internal digitization
   • workflows, processing & delivery

4. Discovery

5. Content delivery
   • mss, image & book application environments

6. Ongoing process reengineering & optimization
The Challenge

Format Registries
  Authn
  Authz
  DRM

Annotation Tools
  Data Mining
  Text mining

Analytic Environments

Image servers
  Geospatial apps
  Media players

Discovery Apps

Visualizations
  Semantic text indices

ERM

SFX

ILS

Digitization Workflow
  Google BookSearch

SDR
  collections.stanford.edu
  SULAIR LOCKSS

Sakai

EEMS

Aquifer ADL Etc.

Federated Content Stores

ebrary
  Luna

stanford.edu

GeoServer

Media players

Geospatial apps

Text mining

Data Mining

Annotation Tools

Analytic Environments

Image servers

Discovery Apps

Visualizations

Semantic text indices

ERM

SFX

ILS

Digitization Workflow

Google BookSearch

SDR

collections.stanford.edu

SULAIR LOCKSS

Sakai

EEMS

stanford.edu
Strategy 1: Hardware Approach

- ILS
- Authn
- Authz
- DRM
- Format Registries
- Discovery Apps
- Delivery Apps
- Geospatial apps
- Media players
- Analytic Environ
- Annotation Tools
- Data Mining
- Text mining
- Image servers
- SFX
- ERM
- ILS
- ADL
- SDR
- collections.stanford.edu
- Sakai
- ebrary
- Luna
- SULAIR LOCKSS
- EEMS
- Aquifer
- Federated Content Stores
- Etc.
- Digitization Workflow
- Google BookSearch
- Semantic indices
- Full text indices
- Strategy 1: Hardware Approach
Strategy 2: Repository Centric

SDR

- Format Registries
- Authn
- Authz
- DRM
- Annotation Tools
- Data Mining
- Text mining
- Analytic Envrons
- Image servers
- Geospatial apps
- Media players
- Delivery Apps
- Visualizations
- Semantic indices
- Discovery Apps
- SFX
- ILS
- ERM
- SULAIR
- LOCKSS
- EEMS
- Digitization Workflow
- Google BookSearch
- ebrary
- Luna
- Sakai
- SDR
- collections.stanford.edu
- SULAIR LOCKSS
- Etc.
- Aquifer
- ADL
- Federated Content Stores

Apps
- Analytics
- Delivery
- Discovery
- Image servers
- Geospatial apps
- Media players
- Visualizations
- Semantic indices
- Analytics
- Delivery
- Discovery
- Image servers
- Geospatial apps
- Media players
- Visualizations
- Semantic indices

SDR

SDR
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
Strategy 3: MacGyver It

- Discovery Apps
- Delivery Apps
- Analytic Environments
- SFX
- ERM
- Unicorn
- Sakai
- SULAIR
- LOCKSS
- EEMS
- Google BookSearch
- Digitization Workflow
- Ebrary
- Luna
- SDR
- Collections.stanford.edu

- Format Registries
- Authn
- Authz
- DRM
- Aquifer
- ADL
- Etc.
- Federated Content Stores

- Annotation Tools
  - Data Mining
  - Text mining
- Image servers
  - Geospatial apps
  - Media players
- Visualizations
  - Semantic
  - text indices

- Analytic Environments
- Delivery Apps
- Discovery Apps
- Image servers
- Geospatial apps
- Media players
- Visualizations
- Semantic text indices

- Strategy 3: MacGyver It
- Delivery Apps
- Image servers
- Geospatial apps
- Media players
- Analytic Environments
- Discovery Apps
- Visualization
- Semantic text indices

- Strategy 3: MacGyver It
- Delivery Apps
- Image servers
- Geospatial apps
- Media players
- Analytic Environments
- Discovery Apps
- Visualization
- Semantic text indices

- Strategy 3: MacGyver It
- Delivery Apps
- Image servers
- Geospatial apps
- Media players
- Analytic Environments
- Discovery Apps
- Visualization
- Semantic text indices

- Strategy 3: MacGyver It
- Delivery Apps
- Image servers
- Geospatial apps
- Media players
- Analytic Environments
- Discovery Apps
- Visualization
- Semantic text indices
Strategy 4: Library Middleware

Content & Service Middleware – conceptual boundary

Reporting

Intelligence

Monitoring

Access Broker

Collections Registry

Content Registry

Format Registries

Authn

Authz

DRM

Aquifer

ADL

Etc.

Federated Content Stores

Annotation Tools

Data Mining

Text mining

Analytic Environ

Image servers

Geospatial apps

Media players

Delivery Apps

Visualizations

Semantic

Text indices

Discovery Apps

SDR

collections.stanford.edu

SULAIR

LOCKSS

ebrary

Luna

Sakai

SFX

ERM

ILS

Digitization Workflow

Google BookSearch

EEMS
Parallels to Identity Management

In enterprise computing, the middleware that reconciles representations of entities (people, organizations, groups, etc.) across disparate systems, distributes this information across the infrastructure, and serves as the basis both for authentication and authorization (privilege assignment & management).

Data change from source system or Registry interface

Data change in source system

Data reconciliation, management & policy application in the registry = identity management

Data delivery triggered via events, published via LDAP, XML, Java, WS, etc. to integrated systems
The process

1. Recognize need
2. Think through narratives
3. Identify parts
   • Applications
   • Services
   • Infrastructure
4. Assemble in an architecture
5. A diversion into ESB’s
6. Validate concepts & find a name (still searching)
Narratives

1. Support for scholarly workflows
2. Personalized academic work environment
3. The DIY Library/Research Environment
4. Integrated, comprehensive content discovery
5. Creating, managing, publishing dynamic digital collections
6. Extend library infrastructure & workflows to already-digital content
7. Flexible digitization workflow
8. Library management, operations and reporting
1. Scholarly workflows

*Support collaboration, research and publication through a full information and creative lifecycle.*

- Research: raw data, to distilled, to published, to reference data.
- Scholar/author: notes, to drafts, to preprint, to published article, to archival version
- Faculty: preparing a course, to conducting a course, to materials added by students, to archiving the course
- Generically: transition from private to group to public
- Post publication lifecycle, e.g., continued access to post-project web sites
Capabilities

• Personal identity -- status, role, relationship to resources
• Collaborators -- roles, groups, rights
• Repository services -- data capture, backup, sharing, archiving
• Preservation -- archived data that’s verifiable, citable, reusable
• Capture annotations, corrections, additions, versioning
• Transformation into publication formats
2. Personalized Academic Work Environment

*Highly personalized services and resources based on persistent and intimate knowledge of the scholar's identity, roles, background, explicit choices, and implicit preferences.*

- Endow all services with an awareness of personal identity, preferences and contributions
- Connect authors to their publications and citations
- Support consistent user tools & experience across UIs
- Preferences
  - Favorite resources (creating resource lists), ranked search results
  - Saved sources (citations), searches, ratings, tags and annotations
  - Language, delivery formats (accessibility)
  - Saved visibility settings (private, group-restricted, public)
3. The DIY Library/Research Environment

Enable scholars to use the tools of their choice to incorporate library data, metadata, and complementary services into their workspace

• Open access/APIs to content, metadata, library services
• Content extraction/delivery to local environment
• Support for syndicating & aggregating content
• Alerts
• Widgets interoperating via internally linked APIs a la iGoogle
• Metrics/usage profiles
• Backup and long term preservation
4. Comprehensive content discovery

*Ability to discover across all content stores and deliver requested content regardless of location (local, licensed, federated, external).*

- Integrated / federated search capabilities
  - across internal digitized collections (with unified metadata)
  - to external sources (via library supported access)
- Export and import of metadata (e.g., OAI data providers, harvesting)
- Common & interoperable data models
- Metadata flattening, translation
- Interoperable content delivery, API's
- Aware of community tagging
5. Creating, managing, publishing dynamic digital collections

The ability for librarians & scholars to create & curate their own digital collections by selectively collecting other digital collections, objects or their components, and remixing them into new collections for personal use, group use or publication.

- Fedora-like support for content disaggregation/reaggregation
- Open access/APIs to content, metadata, library services
- Ability to annotate objects/collections with new commentary
- Authn and authz services, plus application of DRM
6. Extend library infrastructure and workflows to already-digital content

Extensible framework to support simple (monolithic) datatypes and structured information through a full processing lifecycle -- receipt/capture, preservation, annotation, cataloging, indexing, delivery -- for information that is already in digital form.

- Workflow & rules, integrated with other workflows
- Track, manage, reconcile content
- Support for complex, structured objects -- web sites, wikis, blogs
- Web Forms (simplified/uniform UI layer, e.g., XForms)
- eCommerce/ePayment (departmental and personal)
- Digital Rights Management
7. Flexible digitization workflow

Support full range of individual patron requests to project-based collections to mass digitization initiatives, with a mix of formats & media types, metadata structures, and internal & external agents.

- Workflow & rules
- eCommerce/ePayment
- Track, manage, reconcile content
- Quality Control, e.g., format verification, tools for editing
- Location tracking, data moving
- Editing of archived material
- Scoping of front matter
- Scanning licenses
8. Library management, operations, reporting

Provide full administrative capabilities over all aspects of the digital library collections and processing flow to support management, analysis and reporting requirements needed for digital resources. Integrate these capabilities with information from the ILS to support a comprehensive view of Library assets,

- Report on what's held, what’s preserved
- Track context (relationship, use) of materials/objects/collections
- Status of processing queues
- Vendor management
- License management & renewal
- Coordinate information about print and related e-content
- Flexible, distributed reporting
Five themes derived from narratives

1. Identity
2. Preservation
3. Personalization
4. Access
5. Management

Bound together in a common service infrastructure
Theme 1 -- Identity

• Identity Management
  – Unique work/resource
  – Editions, versions, manifestations, formats
  – Whole to parts; parts to whole
  – Rules, policies

• Extensible, multi-faceted
  – Base bibliographic record
  – Aggregated/sourced metadata
  – Formal and informal (community) tagging, commentary, annotations, etc.
Theme 2 -- Preservation

• **Ingest**
  – Extensive metadata
  – Risk analysis
  – Validation

• **Storage**
  – Archived masters and metadata
  – Safety, redundancy and permanence
  – Multi-generational timespan
  – Format migration
Theme 3 -- Personalization

- **User centric**
  - Personal identity, connected to roles, status, authorship
  - Personalized experience (preferences)
  - Consistent across user experience

- **Source and Resource centric**
  - “Personal library”, favorites, ranking
  - Identity enrichment -- annotations, tags, corrections, commentary, reviews, etc.
Theme 4 -- Access

• Discovery
  – Comprehensive across sources
  – Multiple search heuristics
  – Automated, programmatic, alerts, agents
  – Multi-faceted, ontologies, tags

• Presentation
  – Context sensitive, e.g., ranked results
  – Visualizations -- contexts, timelines, relationships
  – Interoperable with personal tools
Services, Infrastructure

• Services
  – Format and protocol conversions
  – Indexing and search
  – UUID assignment / Persistent URL
  – Link Resolution

• Infrastructure
  – Service Registry
  – Policy and Rules engine
  – Messaging
  – Workflow
  – Logging, statistics, metrics, diagnostics
Five themes derived from narratives

1. Identity
2. Preservation
3. Personalization
4. Access
5. Management

Bound together in a common service infrastructure
The Identity Equation
The Preservation Ultimatum
The Access Imperative
The Personalization Supremacy
The Management Sanction
Translating function into architecture
Content management

• Data/metadata stores:
  – Content Registry -- Digital Object Registry (DOR)
  – Personalization (PPPPPPPPPP ...)
  – Fedora
  – Collections -- DCR
  – Rights registry
  – Service registry
  – Schema registry

• Data/content sources
  – Preservation Repository -- SDR
  – collections.stanford.edu
  – NGDA
  – ILS
Services -- Preservation & Management

- Format conversion
- Purl service
- Link reassigner
- Loading records into DOR
- Data source CRUD
- Content mover
- Link checker
- Checksum validation

- Encrypt/decrypt
- JHOVE2/DROID
- Repository ingest
- Reporting
- Monitoring
- Logging
- alert
- scheduling
Services -- Access/Identity/Personalization

- Schema converter
- Link resolver
- Digital signature
- Alert/update
- Data mining
- Preferences api
- Licensing
- Rights
- OAI dp
- SRU/SRW
- authn
- authz
- ecommerce
Infrastructure services

- Service registry
- logging
- workflow
- monitoring
- diagnostics
- rules engine
- messaging
- Ecommerce

- Connections to campus infrastructure
  - Directory/LDAP lookups
  - Calendars
  - Lists
  - File services
  - Groups
  - Connections to other infrastructure (e.g., Higher Ed)
Services are associated with core data stores
What’s In A Name?

‘Digital Library Management System’
‘Content Management System’
‘Content Middleware’
‘Library Cyberinfrastructure’
‘Libra-infrastructure’
‘Lyberinfrastructure’
‘Lyberstructure’
‘Lybermanagement’
‘Lyberware’

- Information Architecture
- Library
- Infrastructure
- Management
- Cyber/digital
- Middleware