Stanford University LIBRARIES & ACADEMIC INFORMATION RESOURCES

The Stanford Digital Repository

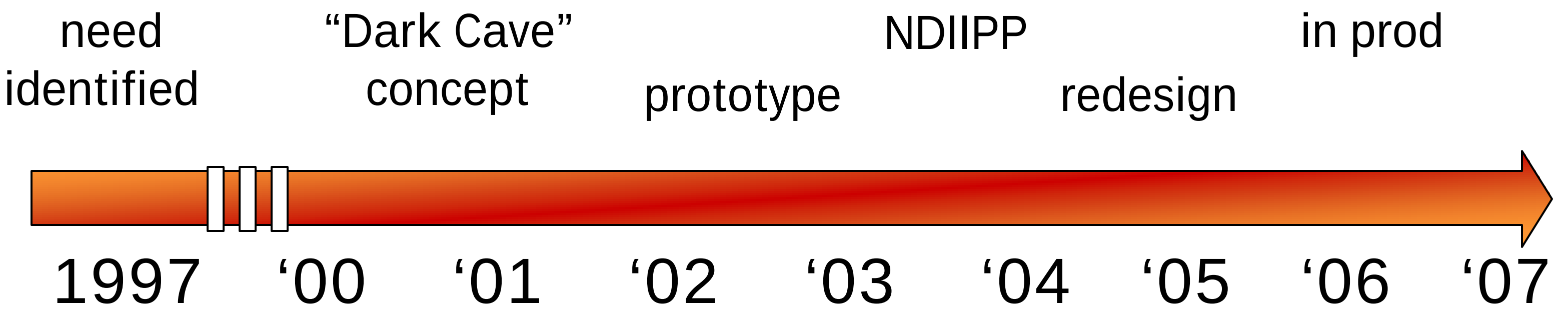
A Case Study in Building A Common Preservation Infrastructure

CNI Spring Task Force Meeting
Phoenix, AZ – April 17, 2007

Tom Cramer, tcramer@stanford.edu
Rachel Gollub, rgollub@stanford.edu

SDR is...

- a preservation system
- now in production with ~3 TB of content
- capable of ingesting ~200 GB/day
- a repository of a certain age



Three Major Areas of Preservation Needs

- Digital Library
  - SULAIR collections & resources
  - Digitization artifacts

Google Books	*('00s of TB)*
Parker Manuscripts	*(75 TB)*
MJF Media	*(50 TB)*
NGDA	*(10 TB)*
~30 other digi projects	*(15 TB)*
Purchased collections	*(25 TB)*

- Institutional Repository
  - Research data,
  - Publications, dissertations,
  - Learning objects, university assets
- External Depositors
  - Online preservation and access
  - Dark archive

Design Objectives & Assumptions

- Preservation-focused archive
- Replicated content
  - (multiple copies, geographically distributed)
- Secure
- Auditable
- Modular
- Tiered storage environment
  - (online, nearline, offline)
- Version rather than delete
- Content-agnostic

Core Repository Functionality

- Preserving access to digital information over time

  ...through generations of technology obsolescence and change.

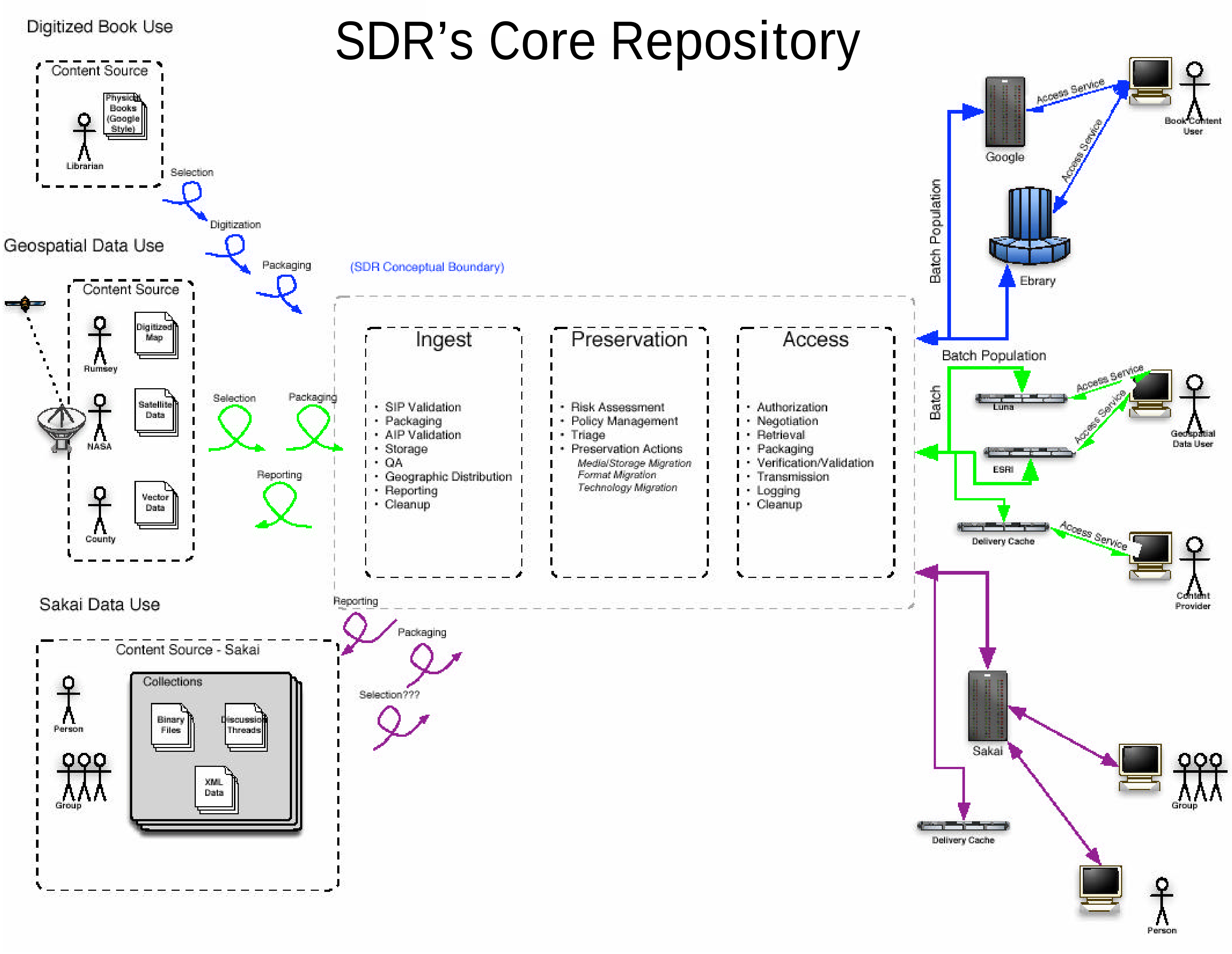
- Maintaining integrity of that information over time

  ...through generations of migration and reformatting.

Repository Services Functionality

- All (or almost all) user-facing services
- Enhanced access & delivery through applications
- Dry research, new indexing, e-science, etc.
- Federation

SDR's Core Repository



Ingest

- SIP Validation
- Packaging
- AIP Validation
- Storage
- QA
- Geographic Distribution
- Reporting
- Cleanup

Preservation

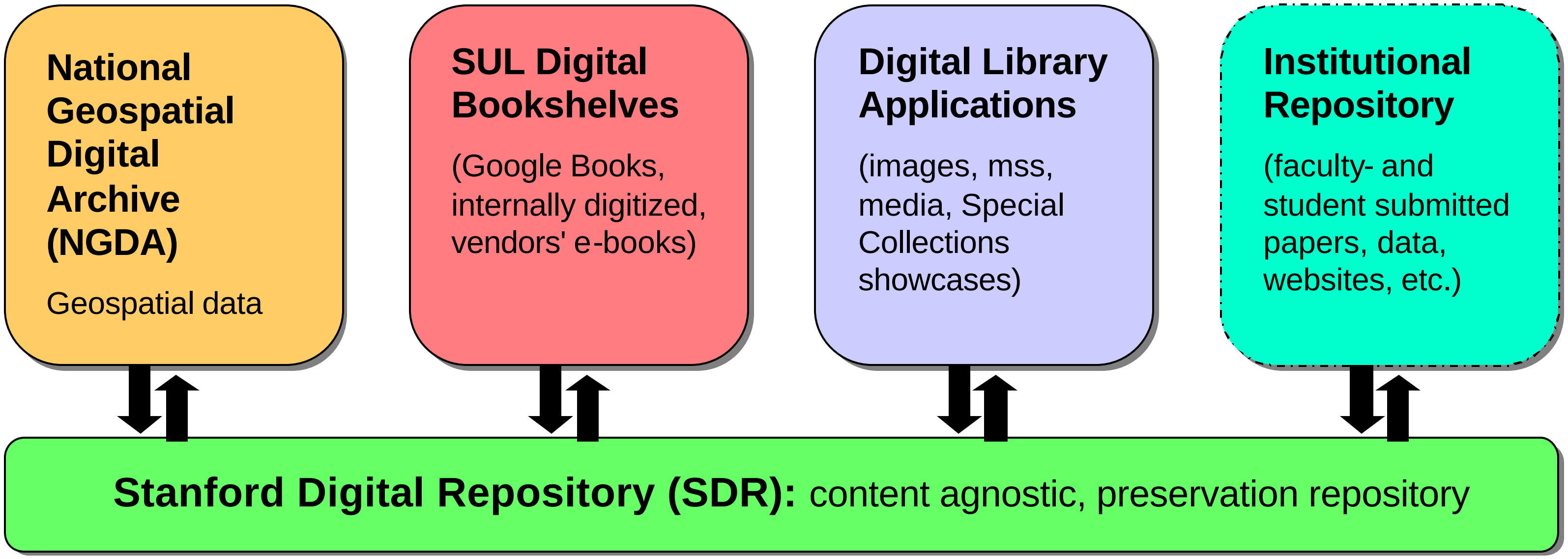
- Risk Assessment
- Policy Management
- Triage
- Preservation Actions
  - *Media/Storage Migration*
  - *Format Migration*
  - *Technology Migration*

Access

- Authorization
- Negotiation
- Retrieval
- Packaging
- Verification/Validation
- Transmission
- Logging
- Cleanup

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)

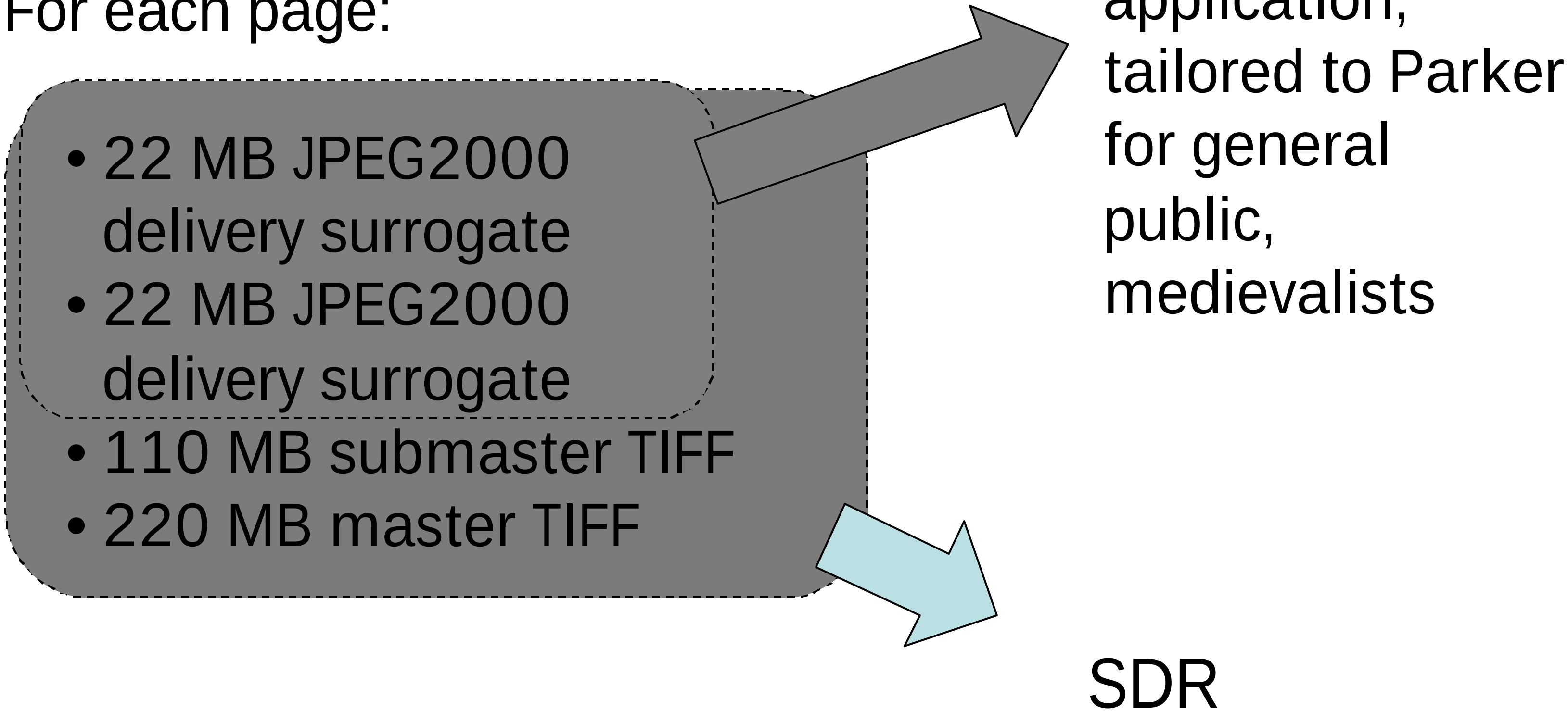
Institutional Repository

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

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

E.g., Preservation of the Parker Manuscripts

- 530 Manuscripts
- 200,000 pages
- For each page:

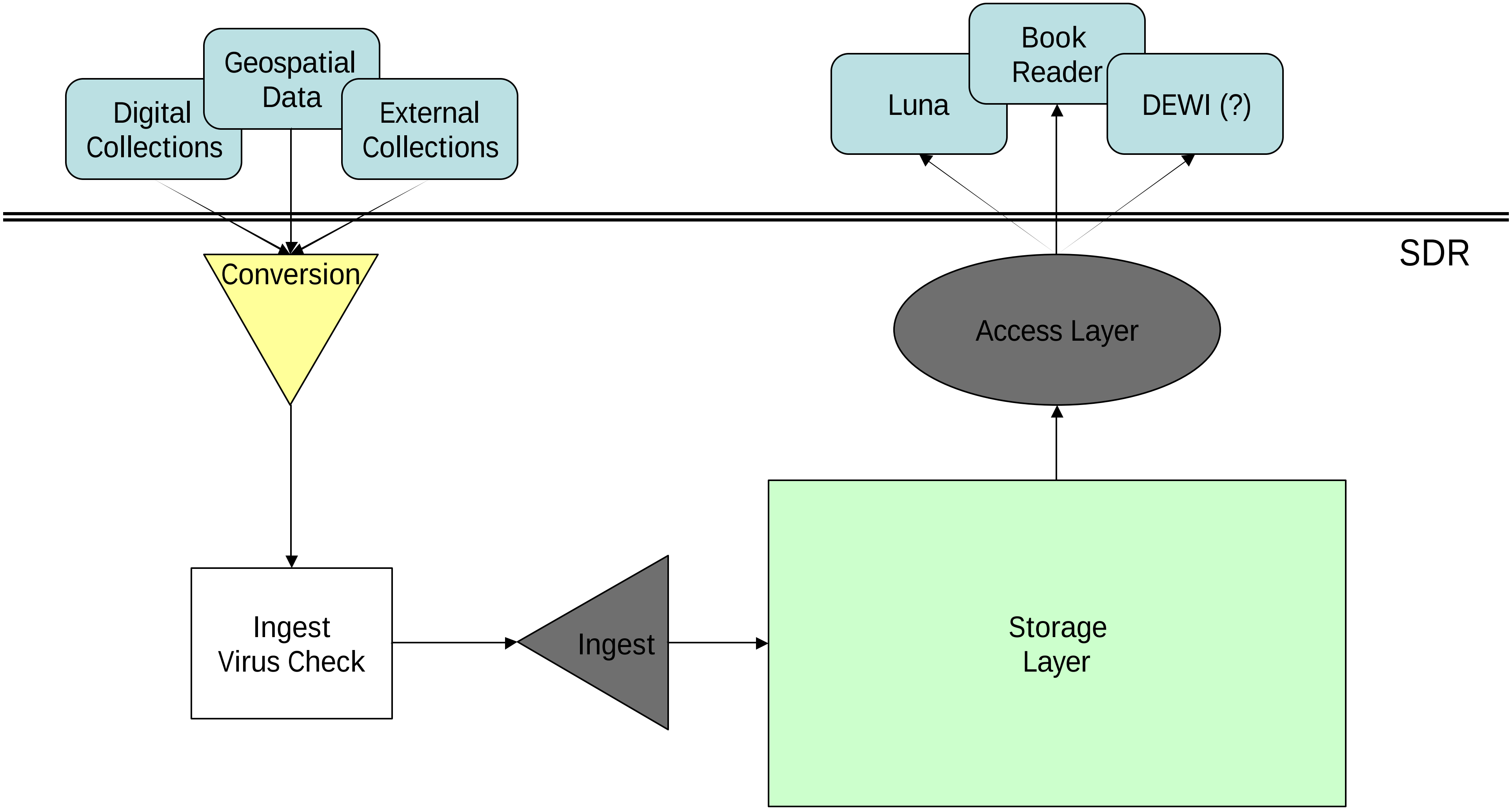


- 22 MB JPEG2000 delivery surrogate
- 22 MB JPEG2000 delivery surrogate
- 110 MB submaster TIFF
- 220 MB master TIFF

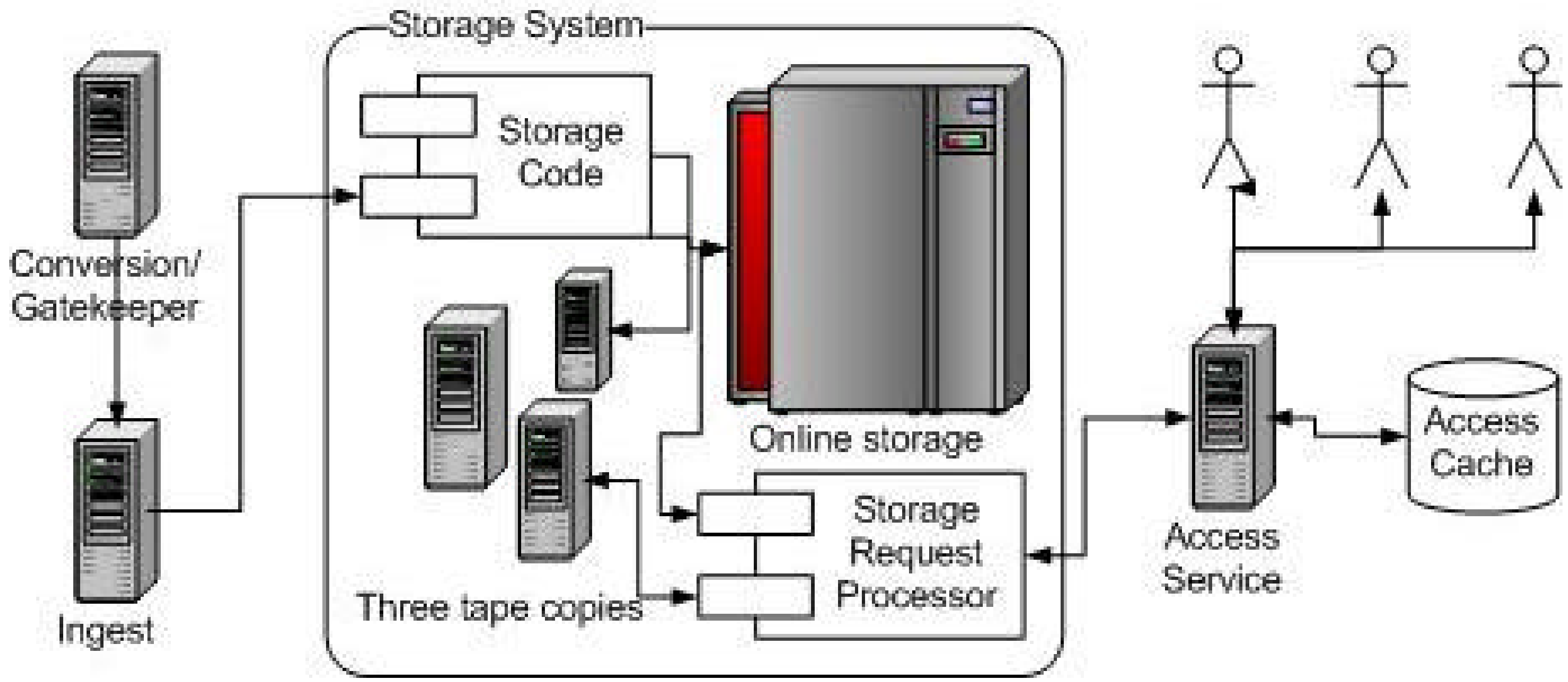
Rich web application, tailored to Parker for general public, medievalists

SDR

SDR Workflow

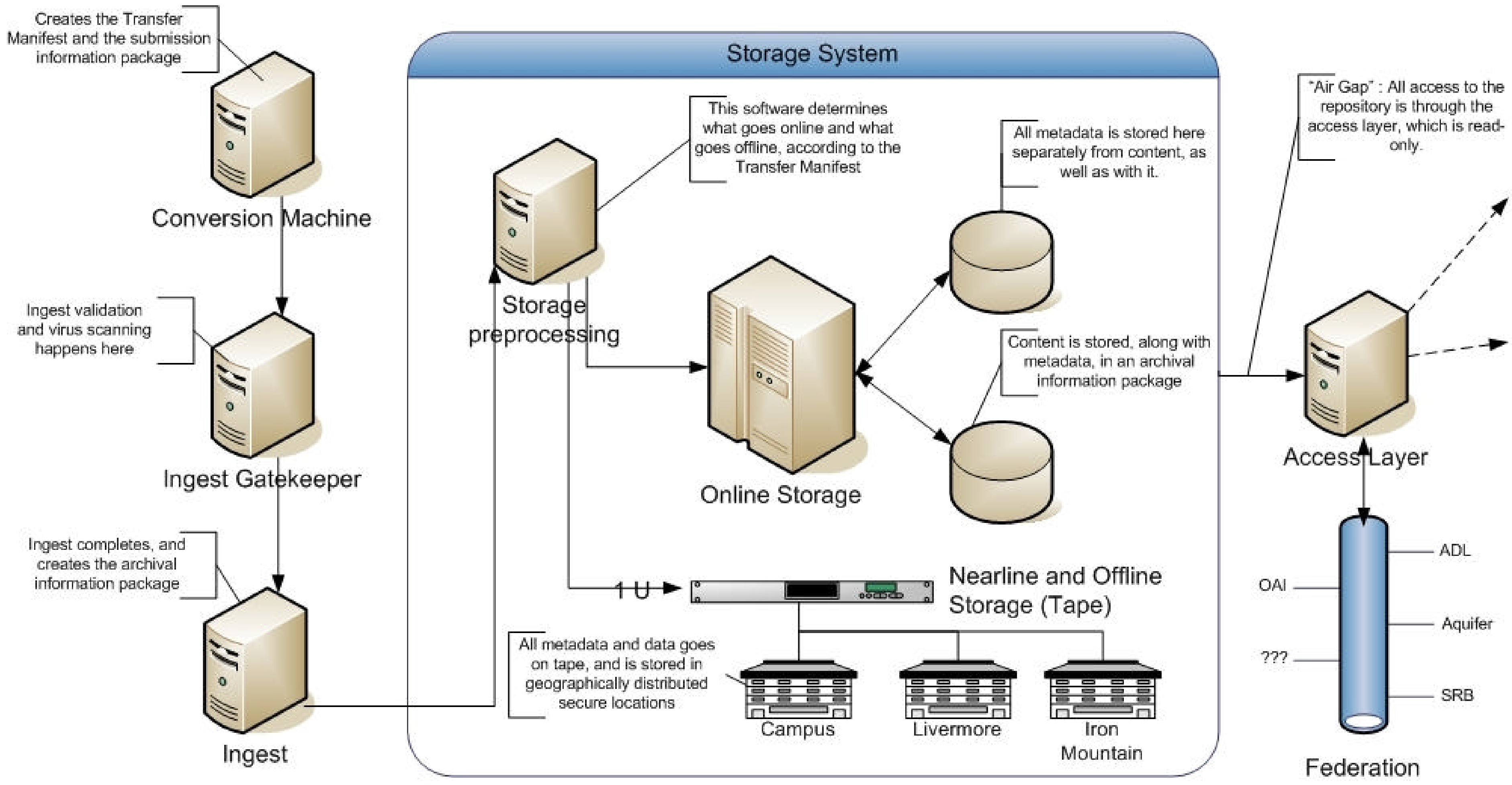


SDR High-Level Architecture



SDR Architecture

March 2006



Creates the Transfer Manifest and the submission information package

Ingest validation and virus scanning happens here

Ingest completes, and creates the archival information package

This software determines what goes online and what goes offline, according to the Transfer Manifest

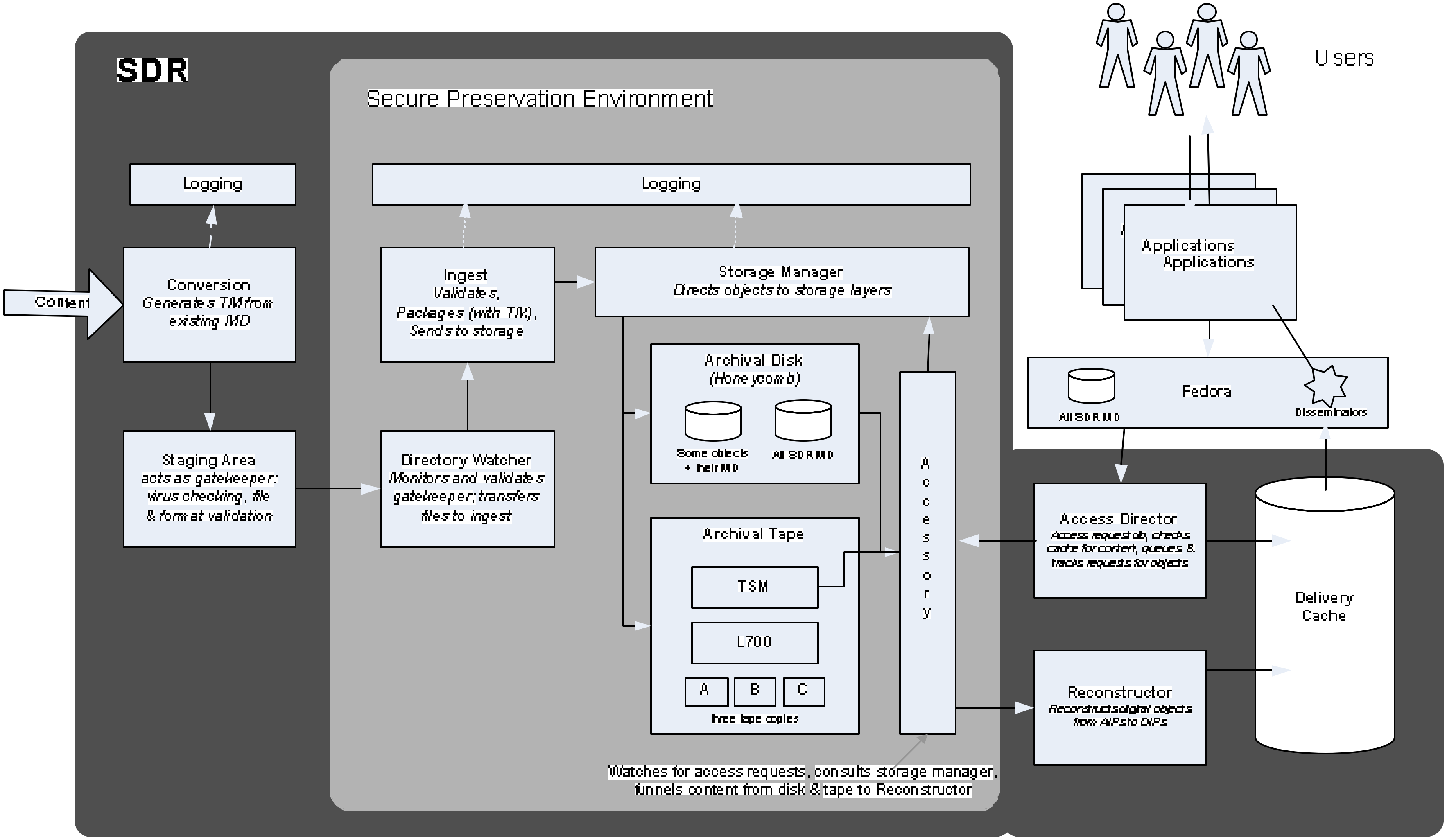
All metadata is stored here separately from content, as well as with it.

Content is stored, along with metadata, in an archival information package

All metadata and data goes on tape, and is stored in geographically distributed secure locations

"Air Gap" : All access to the repository is through the access layer, which is read-only.

SDR Component Diagram


SDR
Secure Preservation Environment
Users
Logging
Logging
Content
Conversion
Generates TM from existing MD
Ingest
Validates, Packages (with TM), Sends to storage
Storage Manager
Directs objects to storage layers
Applications
Applications
Staging Area
acts as gatekeeper: virus checking, file & format validation
Directory Watcher
Monitors and validates gatekeeper; transfers files to ingest
Archival Disk
(Honeycomb)
Some objects + their MD
All SDR MD
Archival Tape
TSM
L700
A
B
C
three tape copies
Accessory
Fedora
All SDR MD
Disseminators
Access Director
Access request ob, checks cache for content, queues & tracks requests for objects
Delivery Cache
Reconstructor
Reconstructs digital objects from AIPs to DIPs
Watches for access requests, consults storage manager, funnels content from disk & tape to Reconstructor

SDR Physical Topology

March 2006

Module(s)	**Hardware**
Conversion, Gatekeeper	Sun Fire X4100 Server 4 TB Nexsan SATA Disk
Ingest, Storage code, Storage Request Processor	Sun Fire X4100 Server 4 TB Nexsan SATA Disk
Online storage	32 TB Sun Honeycomb Storage System
Tape Copies	Sun StorEdge L700 Tape Library, with LTO2 drives IBM Tivoli Storage Manager Iron Mountain data protection plan
Access Service, Access Cache	Sun Fire X4100 Server 8 TB of Nexsan SATA Disk

Metadata Strategy

- A “Transfer Manifest” is generated during conversion (pre-ingestion): includes descriptive, administrative, and structural metadata for the object
- METS wrapper; descriptive MD in MODS
- Parsed and error-checked automatically by Ingest
- Minimum required set is very small
- Ideal: a finite and manageable number of schemas. E.g., simple book, manuscript, image…

A Sample Transfer Manifest

```
<mets ID="library_stanford_edu_e33914b2-fa74-11da-83e8-db2a90744a3c"
OBJID="library_stanford_edu_e33914b2-fa74-11da-83e8-db2a90744a3b"
LABEL="Generic Bit Preservation Agreement for SULAIR DPG collections"
TYPE="SUL_SDR__transferManifest" ...>
    <metsHdr CREATEDATE="2006-06-13T13:41:17" RECORDSTATUS="TM">
       <agent ROLE="CREATOR" TYPE="OTHER" OTHERTYPE="SOFTWARE">
          <name>SDR_CSN_CONVERTER_V1.1</name>
       </agent>
       <altRecordID TYPE="SUL_CSN_objectID">90990250</altRecordID>
    </metsHdr>
    <dmdSec ID="DMD_CSNObjID">
       <mdWrap MDTYPE="MODS" LABEL="CSN_BookSimple, CSN_SUL_ProjectName">
          <xmlData>
             <mods:mods>
                <mods:titleInfo>
                   <mods:title>
                      SDR Preservation Agreement: Bit Preservation,
Generic, v1.0
                   </mods:title>
...
```

Using References Among Transfer Manifests for...

- Preservation Agreements
  - signed, approved agreement
  - specifies long term plan, details of preservation
  - access rights by depositor maintained group
  - ingested as a digital object, referenced by the TM
- Collection Level Metadata
  - collection schema
  - ingested as a digital object, referenced by other digital objects' TMs
- Versioning
  - file by file
  - most current data referenced in most current TM
  - each TM refers to the TM immediately previous, so no version is lost
- File format information
  - format registry (GDFR)
  - links from TMs to ingested format information from the registry

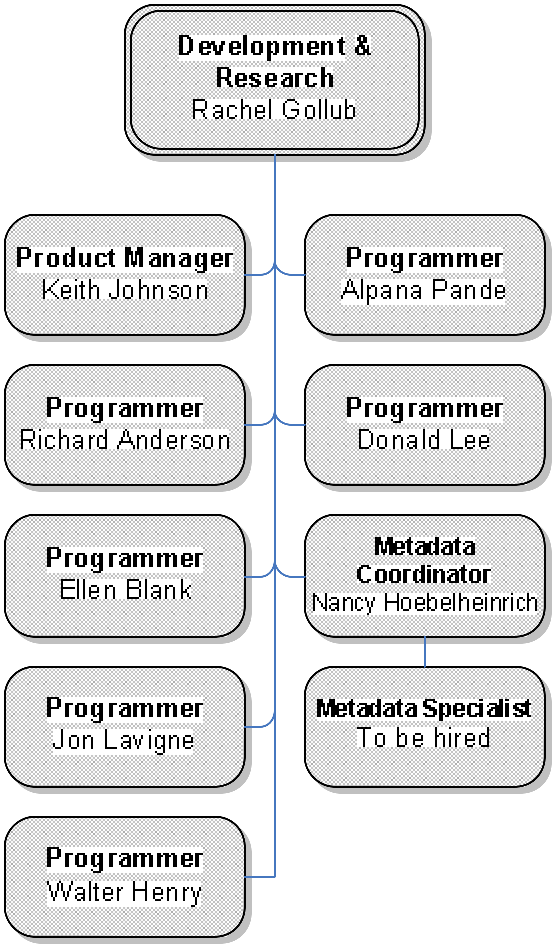
What Do We Wonder About?

- MD extensibility, flexibility
- Duplication of content b/t access and preservation systems. Big storage overhead.
- Need for DOR (digital object registry) – metadata tracking, reconciliation outside of SDR
- Long term overhead of tape infrastructure
- Sustaining focus, progress once we:

  1.) go into full production

  2.) phase out NGDA-funded development

Organization(al Learnings)



Development & Research
Rachel Gollub

Product Manager
Keith Johnson

Programmer
Alpana Pande

Programmer
Richard Anderson

Programmer
Donald Lee

Programmer
Ellen Blank

Metadata Coordinator
Nancy Hoebelheinrich

Programmer
Jon Lavigne

Metadata Specialist
To be hired

Programmer
Walter Henry

Mix of roles needed:

- Dedicated development manager, architect, project manager (Rachel)
- one ‘product manager’, digital preservationist (Keith)
- embedded, expert metadata design, support (Nancy)
- Five developers (plus one borrowed) swarming development through reassignment into one group
- discrete operational/production support group
- NGDA grant mgmt by someone outside the group (Julie)
- In house system administration & storage administration support
- organizational emphasis/top priority over 2+ years

What Has Worked?

- Dedicated team (albeit with other duties). Can’t develop a system of this scope and scale on the margins
- Core repository v. repository services distinction. Helped segment the problem into bite-sized chunks, and move forward. Helped tamp down expectations that SDR would be all things to all people.
- Modularity mindset. Developing tomorrow’s legacy code today. Commitment to making progress, even if we don’t have the perfect answer.
- Concrete use cases - NGDA, Google, Parker, etc. helped frame what we need. SDR development not predicated on hypothetical IR use cases.

What Has Worked? (continued...)

- Sys & storage admins & dedicated infrastructure.
- Surprised even ourselves by sheer volume of machines & storage
- Decision for tiered storage–tape gives us significant capacity, relieves (some of the) pressure on selection, management.
- Leveraging NGDA activity as a catalyst–to start production development–and tying SDR's progress to NGDA milestones