Repository Development Center (RDC)

Office of Strategic Initiatives

LIBRARY OF CONGRESS

Babak Hamidzadeh, Director

What we do

- Build & deploy to production, processes and software systems that enable management of digital collections in their lifecycle.
- Digital Collections lifecycle includes:
  - Production
  - Selection
  - Transfer
  - Preservation
  - Access

Our Vision

- Build tools for librarians and archivists to operate (not for technologists to operate).
  - User interfaces become important
- Design & build to scale & to reduce cost
  - Less forensics & manual processing over time
- Human in many of the links in the loop (semi-automated?)
  - Workflows become important

Our Vision

- One monolithic system is unlikely to work for all content types, formats & uses
  - Interoperability, interfaces & standards become important
- Requirements come in small, varying packages, over time.
  - Iterative development & deployment become important
- Expose content (at item level)!
  - Websites, portals & access applications become less important

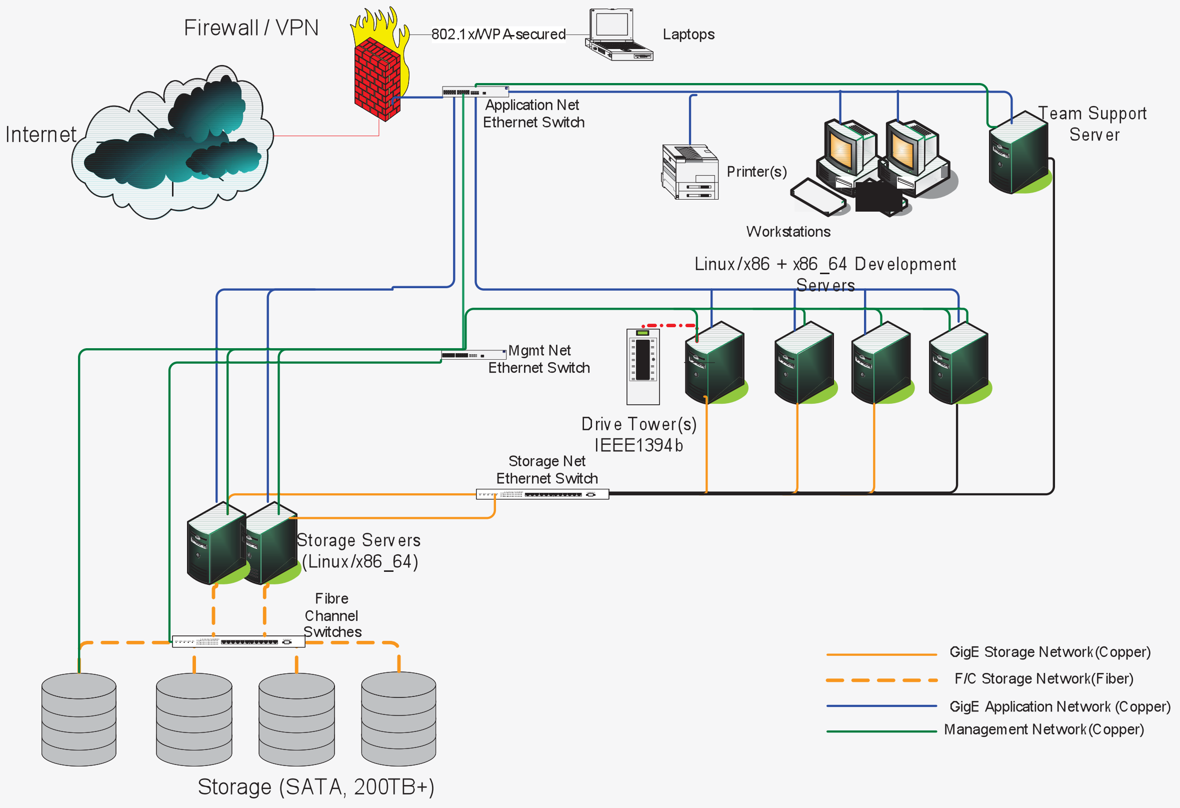
Team

- Technical project management
- Software development
- Software quality assurance
- System operations and maintenance
- System deployment

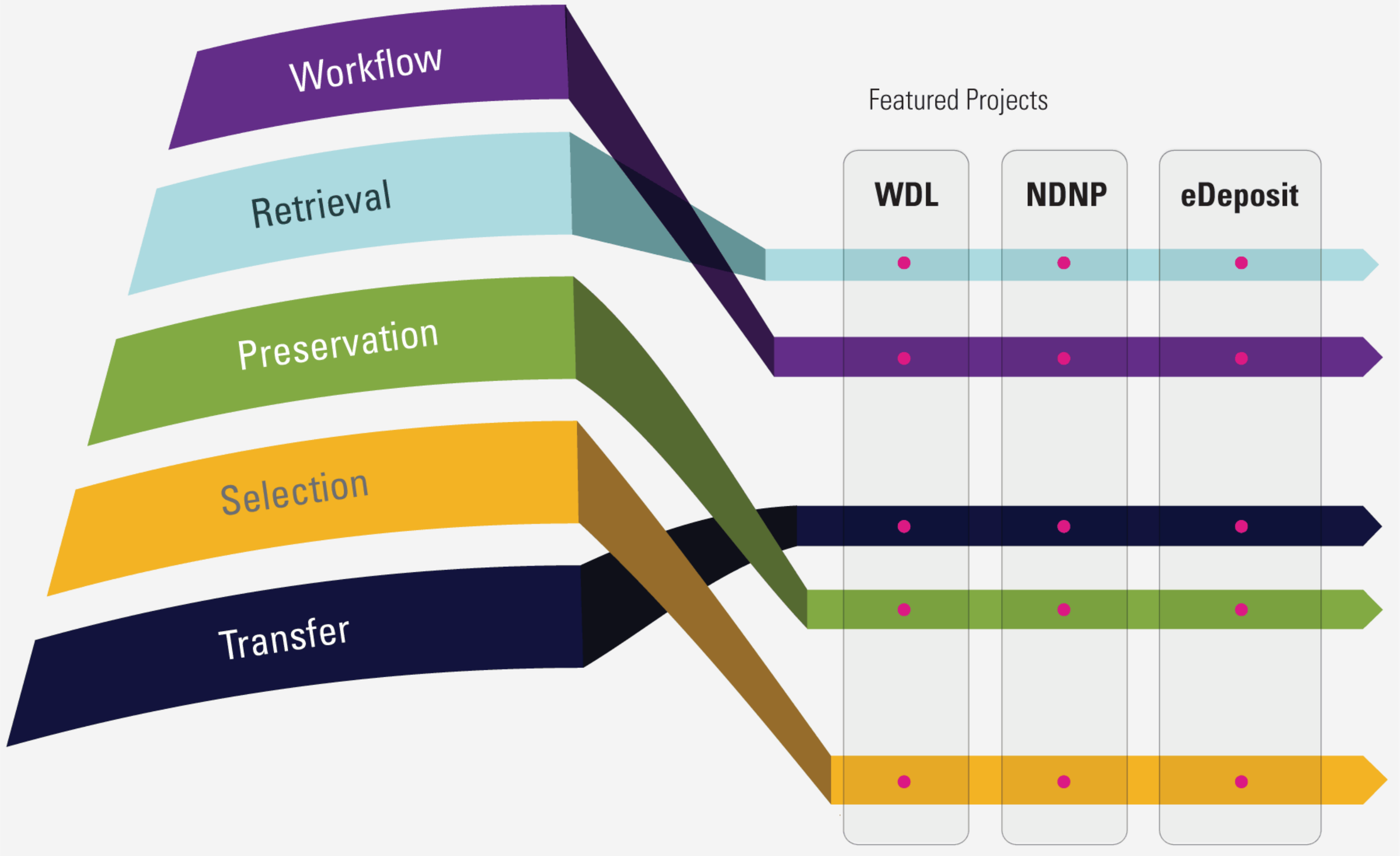
Process

- Project Charter
- Requirements Document
- Technical Development
- Deployment Plan

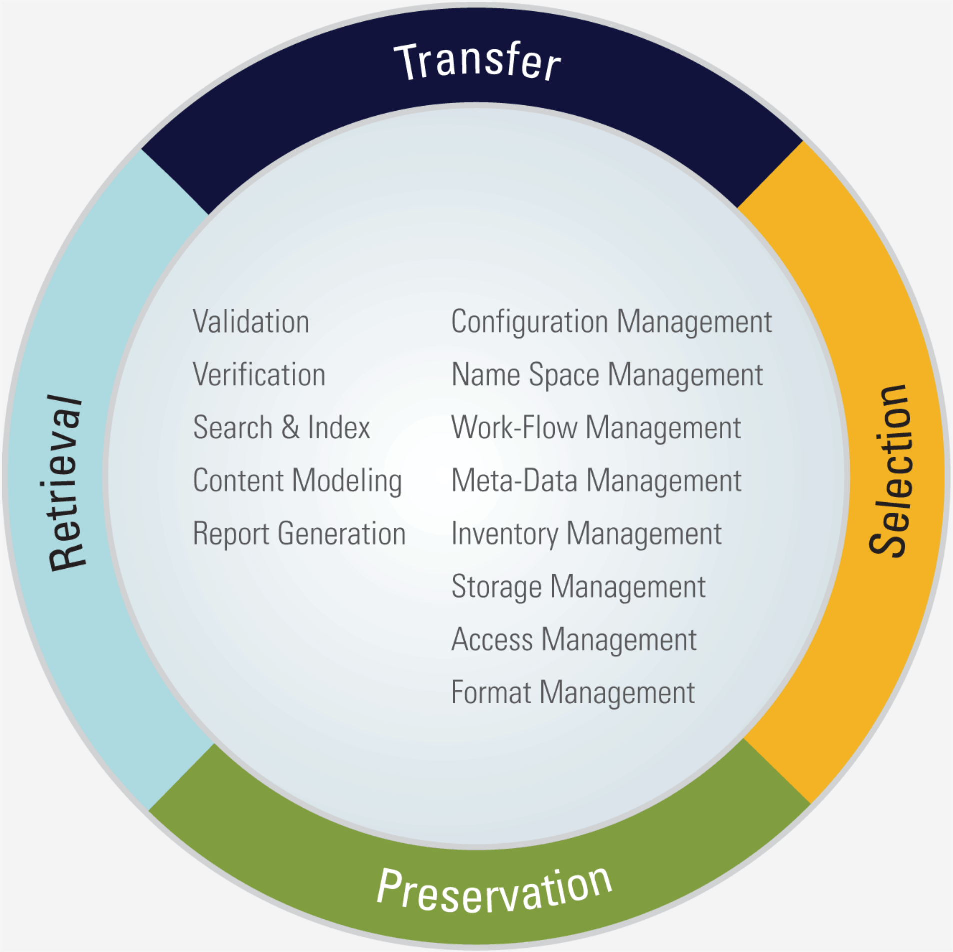
Development Environment



Repository Services



Repository Components



Transfer

Selection

Preservation

Retrieval

Validation

Verification

Search & Index

Content Modeling

Report Generation

Configuration Management

Name Space Management

Work-Flow Management

Meta-Data Management

Inventory Management

Storage Management

Access Management

Format Management

Repository Attributes

- Unique, consistent & persistent identifiers
- Consistent file system structures across collections
- Initially, tools using simple file & directory operations
- Inventory of all digital objects, their associated files & their integrity information
- Audits based on the inventory system
- In-severable, two-way link between items & their meta-data

Repository Attributes

- Ability to recognize & validate formats
- Semantic content models for preservation & access
- Ability to salvage files/objects independently of repository or other software
- Versioning for content, meta-data and identifiers
- Automated ingest in production, by operators
- Access vs Preservation: Separate mechanisms, formats

PROGRAMS

- Digital Content Transfer
- National Digital Newspaper Program (NDNP)
- World Digital Library (WDL)
- eDeposit

Digital Content TRANSFER

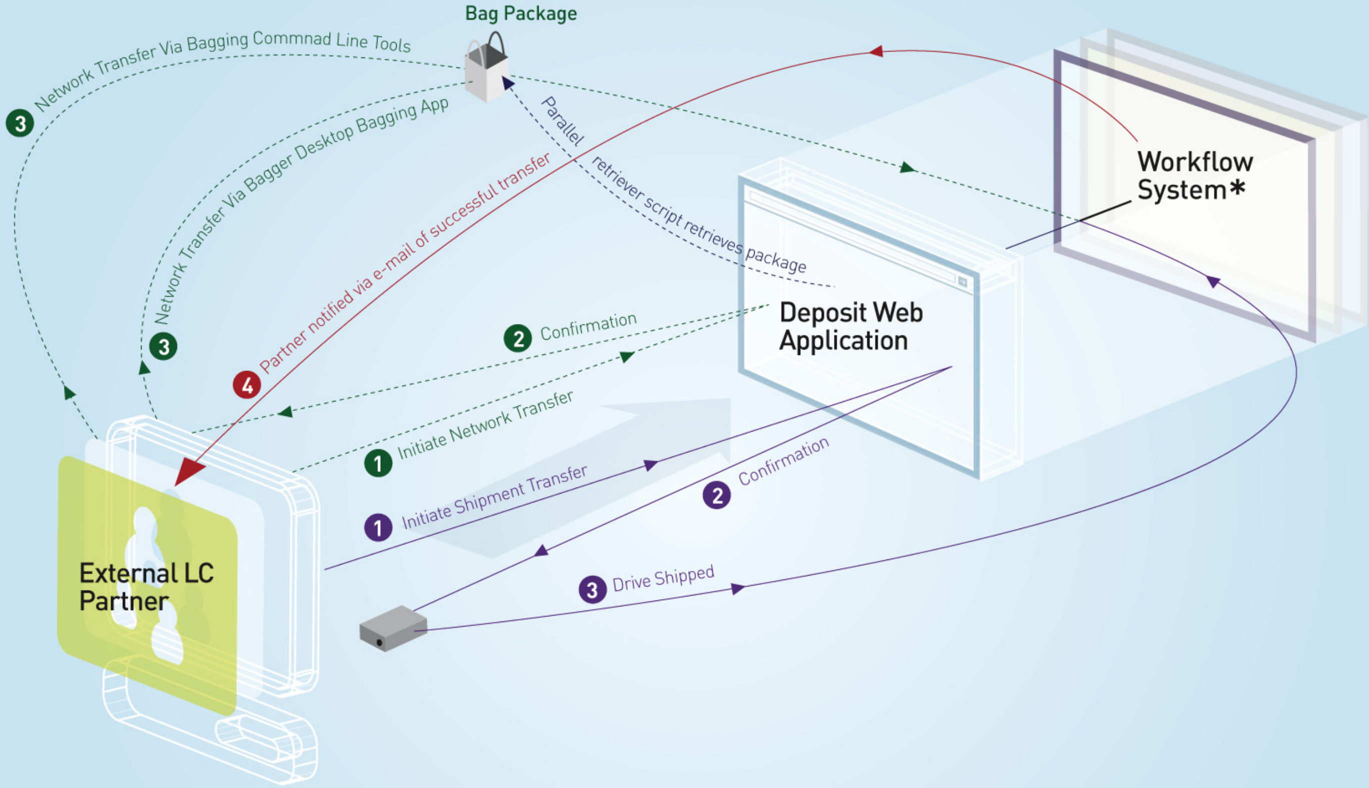
Overview

- Basic repository service to allow movement of large-scale digital content between entities (e.g. persons, organizations).
- Content type agnostic.
- Ensures content integrity.
- Maintains an inventory of content received.
- Does not require high technical capability from the sender.
- Accommodates organizations' workflows.

TRANSFER TOOLS

- BagIt: A content packaging specification for file transfers.
- Bagger: Graphical desktop application to create/update/validate Bags.
- LoCDrop: Web application to register transfers.
- Workflow System: Reconfigurable tool to capture & enforce various content transfer scenarios.
- Inventory System: Tool to inventory Bags, files, their locations, file integrity information, & lifecycle events (e.g. moving, copying, creation of derivatives).
- Parallel Retriever: Tool to exploit available network bandwidth for Bag transfer.
- VerifyIt: Application to verify file integrity during transfer.
- BagIt Library (BIL): Used for application & command line tool development.

Process & Control Flow


Bag Package
3 Network Transfer Via Bagging Commnad Line Tools
3 Network Transfer Via Bagger Desktop Bagging App
4 Partner notified via e-mail of successful transfer
Parallel retriever script retrieves package
Workflow System*
Deposit Web Application
2 Confirmation
1 Initiate Network Transfer
1 Initiate Shipment Transfer
2 Confirmation
3 Drive Shipped
External LC Partner

LEGEND

- - - - - Network Transfer

——— Shipping Transfer

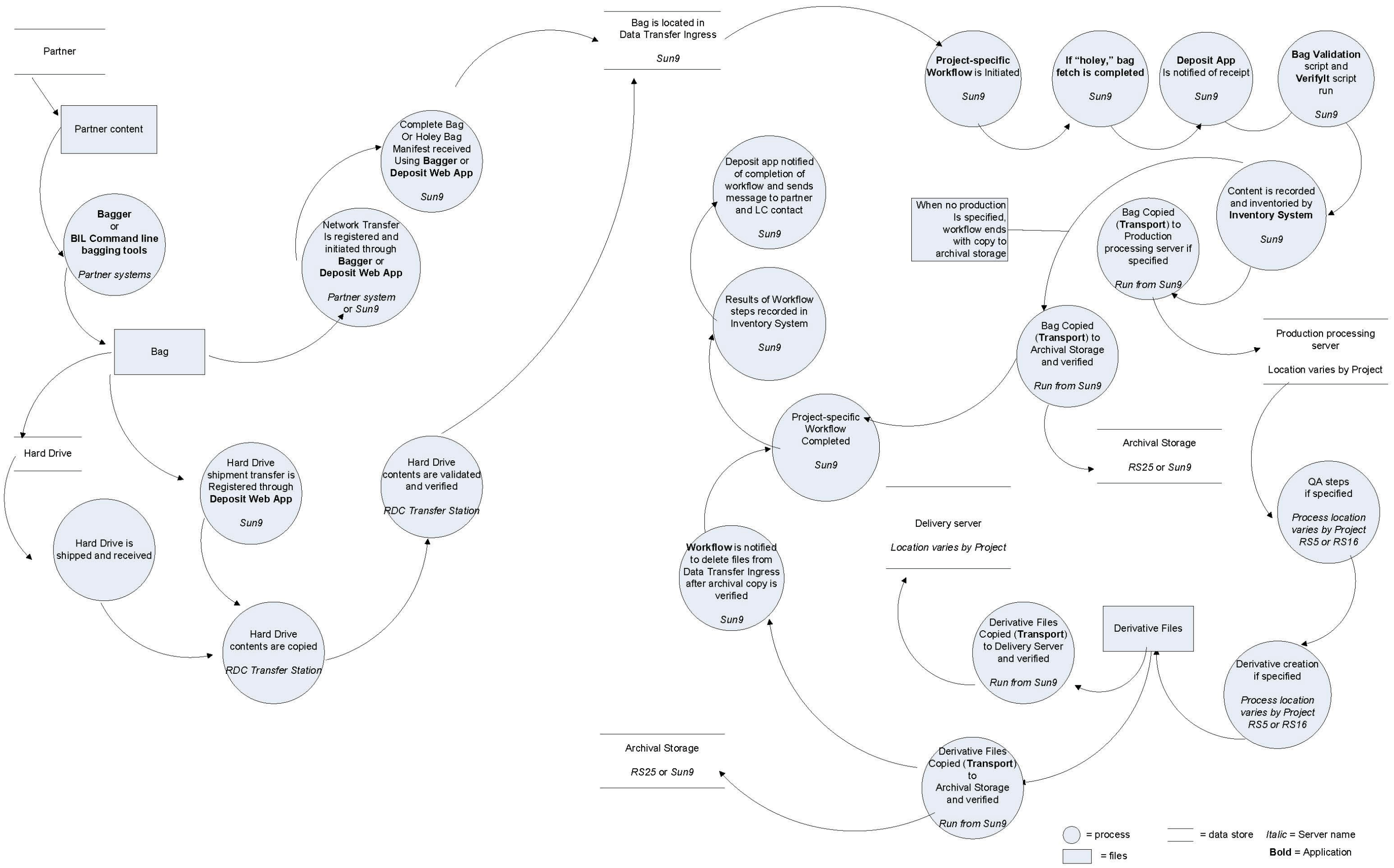
* Workflow System

Appropriate project-based **Workflow UI** (NDNP, NDIIPP, Internet Archive Capture,eDeposit, etc) launched.

Tasks vary by project, but includes **Bag validation** using Bag Validator script, file fixity checking using **Verifyit** script, format validation using JHOVE or DVV, transport of files to a production server, and transport of files to Sun29 for archival storage

DATA FLOW & WORK FLOWS

Transfer Data Flow Diagram: External Partner to LC



Partner

Partner content

Bagger or **BIL Command line bagging tools**
Partner systems

Bag

Hard Drive

Hard Drive is shipped and received

Hard Drive shipment transfer is Registered through **Deposit Web App**
Sun9

Hard Drive contents are copied
RDC Transfer Station

Hard Drive contents are validated and verified
RDC Transfer Station

Network Transfer Is registered and initiated through **Bagger** or **Deposit Web App**
Partner system or *Sun9*

Complete Bag Or Holey Bag Manifest received Using **Bagger** or **Deposit Web App**
Sun9

Bag is located in Data Transfer Ingress
Sun9

Project-specific Workflow is Initiated
Sun9

If "holey," bag fetch is completed
Sun9

Deposit App Is notified of receipt
Sun9

Bag Validation script and **VerifyIt** script run
Sun9

Content is recorded and inventoried by **Inventory System**
Sun9

Bag Copied (**Transport**) to Production processing server if specified
Run from Sun9

When no production Is specified, workflow ends with copy to archival storage

Bag Copied (**Transport**) to Archival Storage and verified
Run from Sun9

Production processing server
Location varies by Project

Archival Storage
RS25 or *Sun9*

QA steps if specified
Process location varies by Project RS5 or RS16

Derivative creation if specified
Process location varies by Project RS5 or RS16

Derivative Files

Derivative Files Copied (**Transport**) to Delivery Server and verified
Run from Sun9

Delivery server
Location varies by Project

Derivative Files Copied (**Transport**) to Archival Storage and verified
Run from Sun9

Archival Storage
RS25 or *Sun9*

Project-specific Workflow Completed
Sun9

Results of Workflow steps recorded in Inventory System
Sun9

Deposit app notified of completion of workflow and sends message to partner and LC contact
Sun9

Workflow is notified to delete files from Data Transfer Ingress after archival copy is verified
Sun9

○ = process
▭ = files
═ = data store
Italic = Server name
Bold = Application

STATUS

- BagIt in use in several institutions (e.g. Portico, CDL, IA).
- LC's first Open Source software release via SourceForge.
- 30 Tb received from NDIIPP partners
- 20 Tb received in web crawls from the Internet Archive
- Dozens of hard drives received with licensed, partner & vendor-supplied content
- Content was in all types and formats.
- From 10 GB to over 2 Tb in a single transfer over the network.

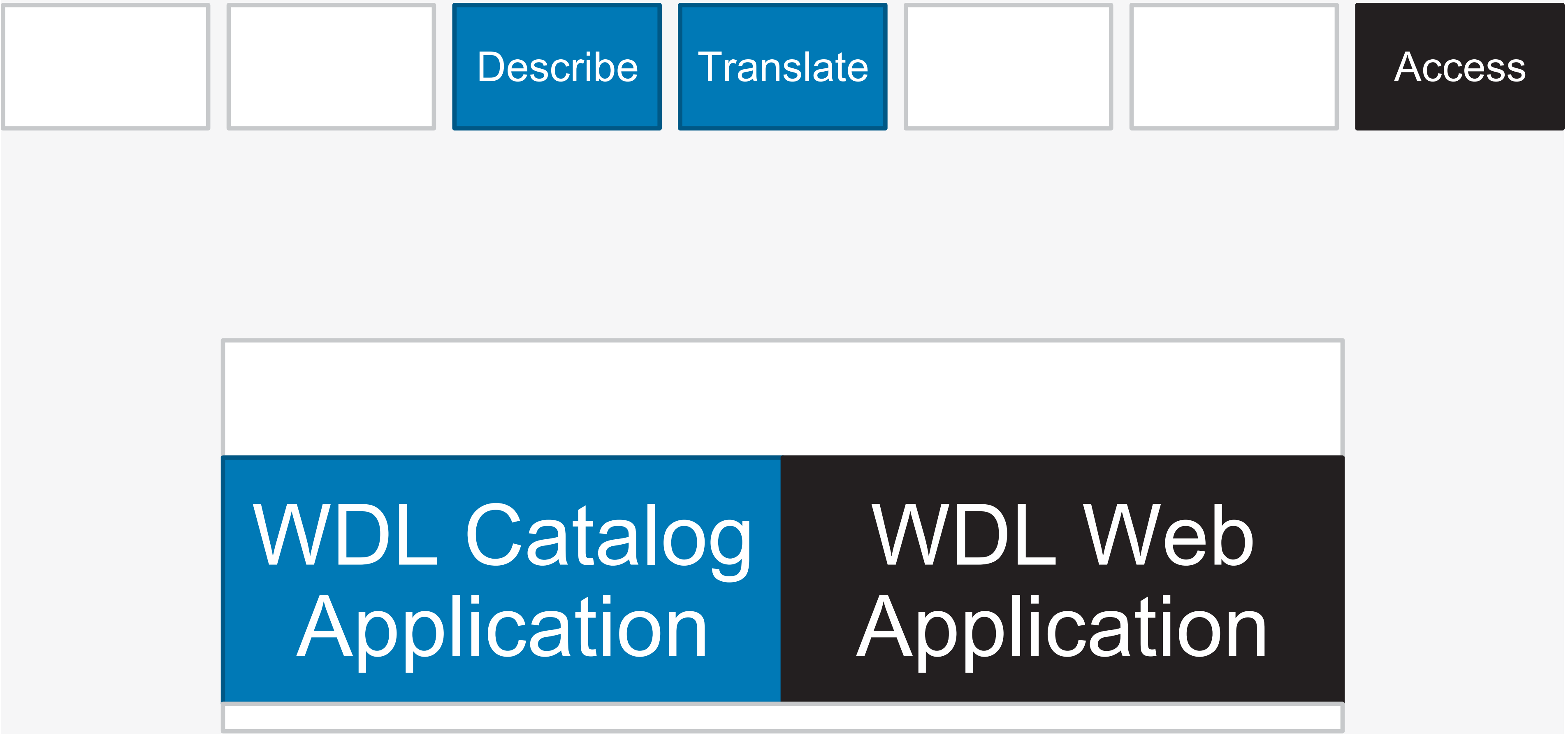
WORLD DIGITAL LIBRARY (WDL)

wdl.org

Overview

- Pubic access and preservation services to historically significant content from cultures around the world
- Content includes maps, prints, photographs, rare books, manuscripts, journals, sound recordings, motion pictures
- Multi-lingual (7 Languages) meta-data & catalog information
- Complex content processing workflows between external (*partners, translators, hosting companies*) and internal (*catalogers, content examiners, technical development*) organizations.

The WDL Architecture Overview



Describe

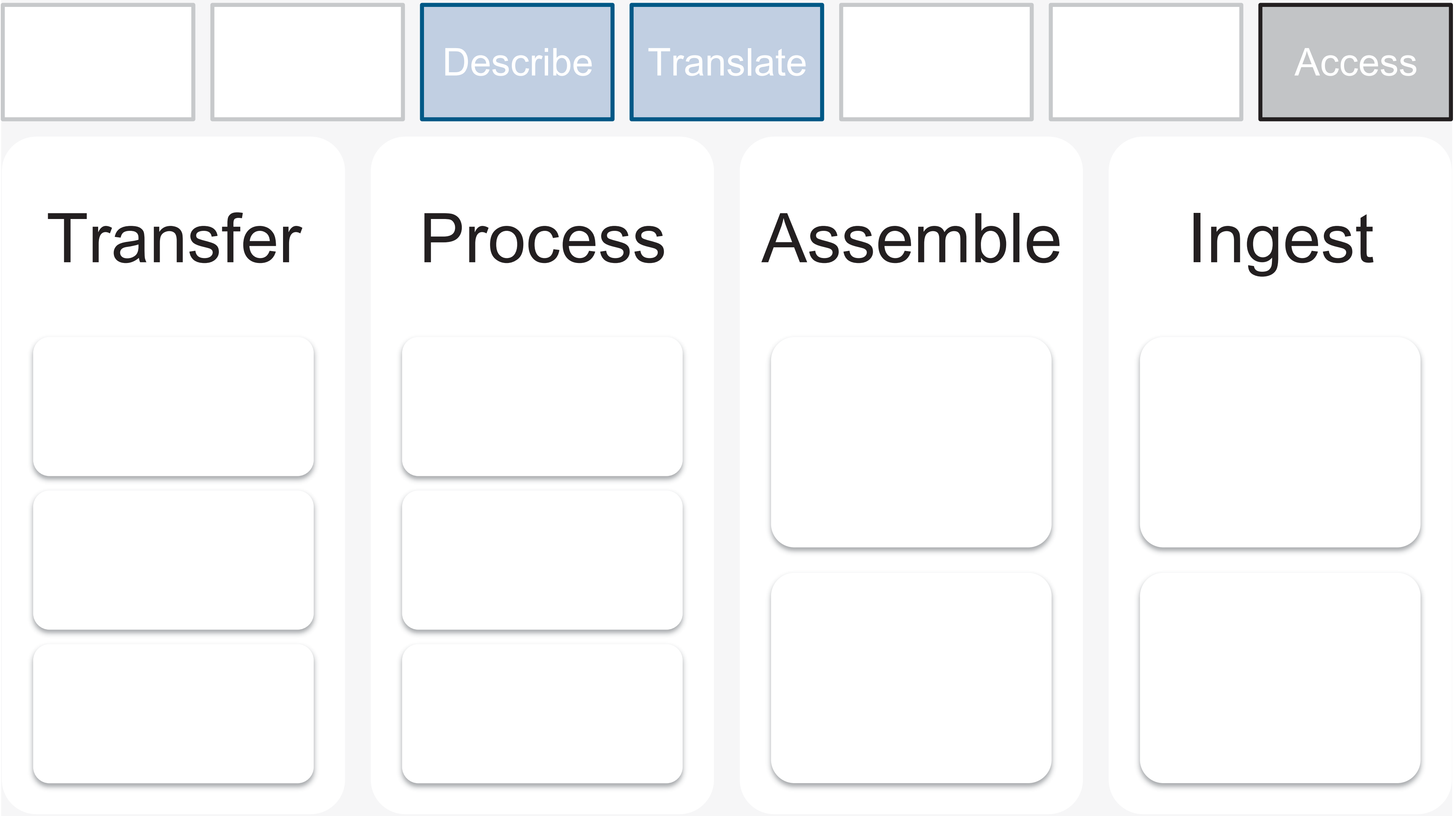
Translate

Access

WDL Catalog Application

WDL Web Application

The WDL Content Pipeline



Describe

Translate

Access

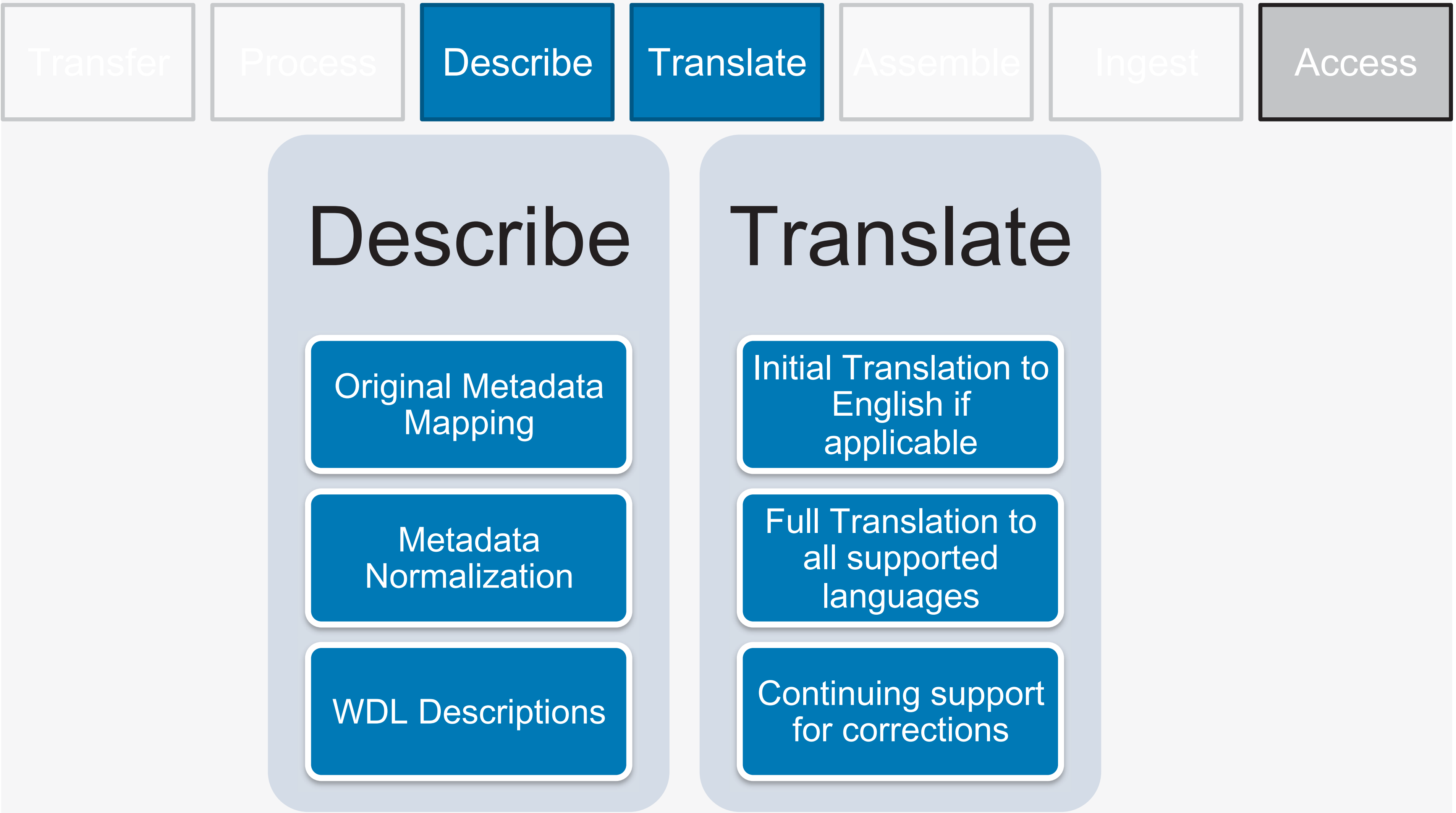
Transfer

Process

Assemble

Ingest

The WDL Catalog Application



Transfer

Process

Describe

Translate

Assemble

Ingest

Access

Describe

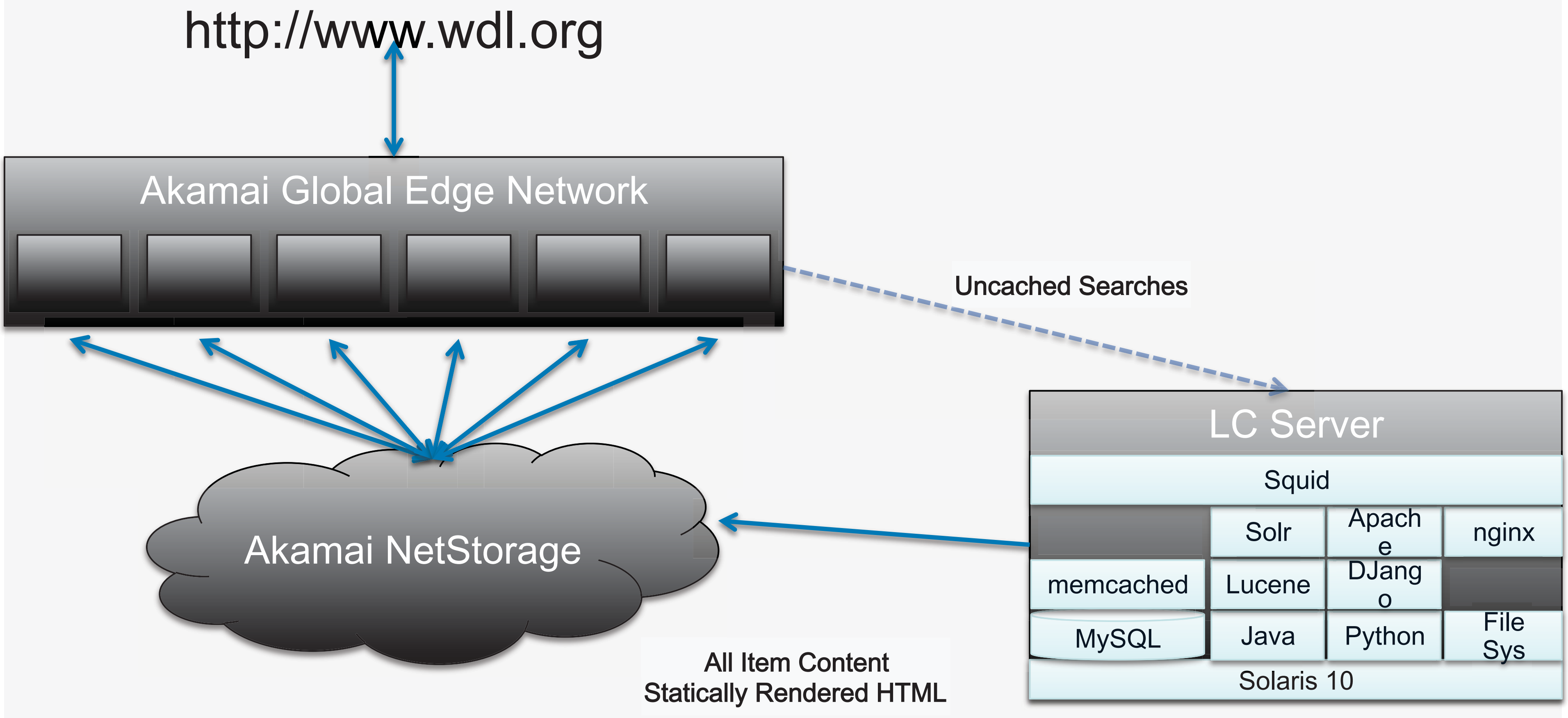
- Original Metadata Mapping
- Metadata Normalization
- WDL Descriptions

Translate

- Initial Translation to English if applicable
- Full Translation to all supported languages
- Continuing support for corrections

The WDL Web Application

Transfer | Process | Describe | Translate | Assemble | Ingest | Access



http://www.wdl.org

Akamai Global Edge Network

Uncached Searches

LC Server

Squid

Solr

Apach
e

nginx

memcached

Lucene

DJang
o

MySQL

Java

Python

File
Sys

Solaris 10

Akamai NetStorage

All Item Content
Statically Rendered HTML

STATUS

- Public launch on April 21 at UNESCO
- 1,500 items, 1,000,000 files
- 15.8 Million page views and 1.4 Million visitors on the first 2 days.
- Peak Hits/Hour: 32 Million
- 56 international partner institutions

National Digital Newspaper Program (NDNP)

chroniclingamerica.loc.gov

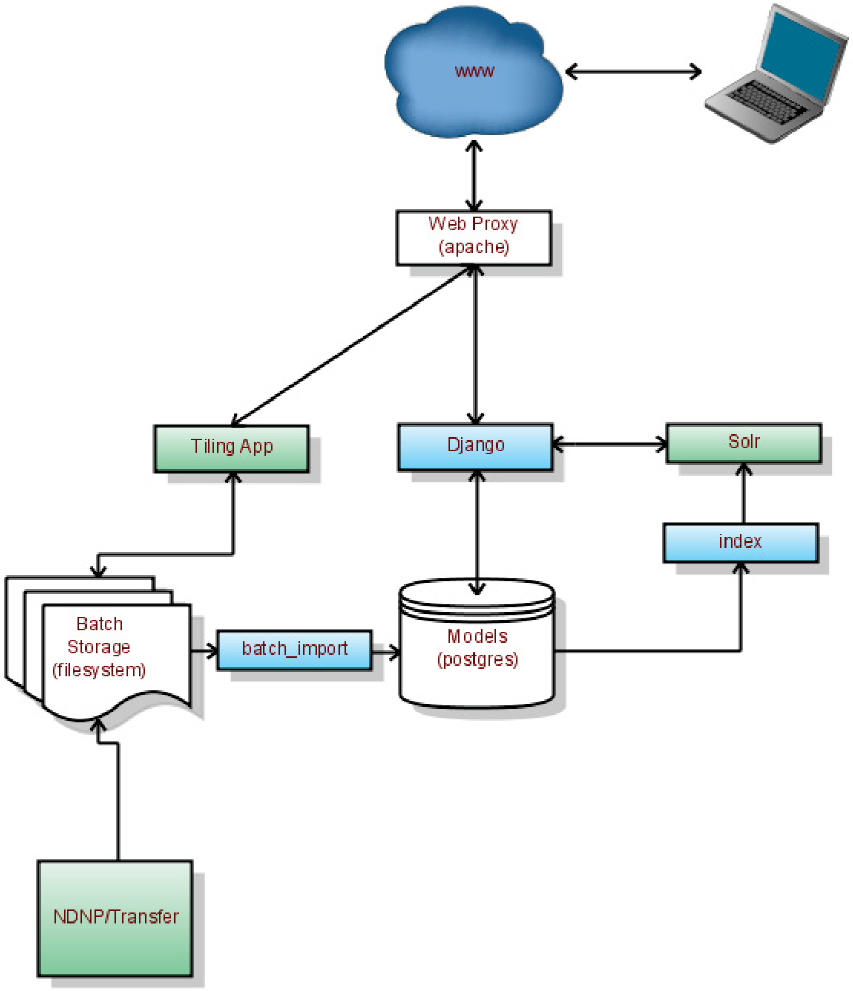
OVERVIEW

- Preservation of and access to historic U.S. newspapers
- Partnership with NEH
- Multiple content producers around the U.S.
- Content submission guidelines
- Digitization standards

FRAMEWORK

- Full-text search with hit-highlighting (Alto OCR)
- Metadata (METS, MARC, MODS)
- Uniform content submission specifications
- Validation at senders’ side (Validation Library)
- Verification upon receipt

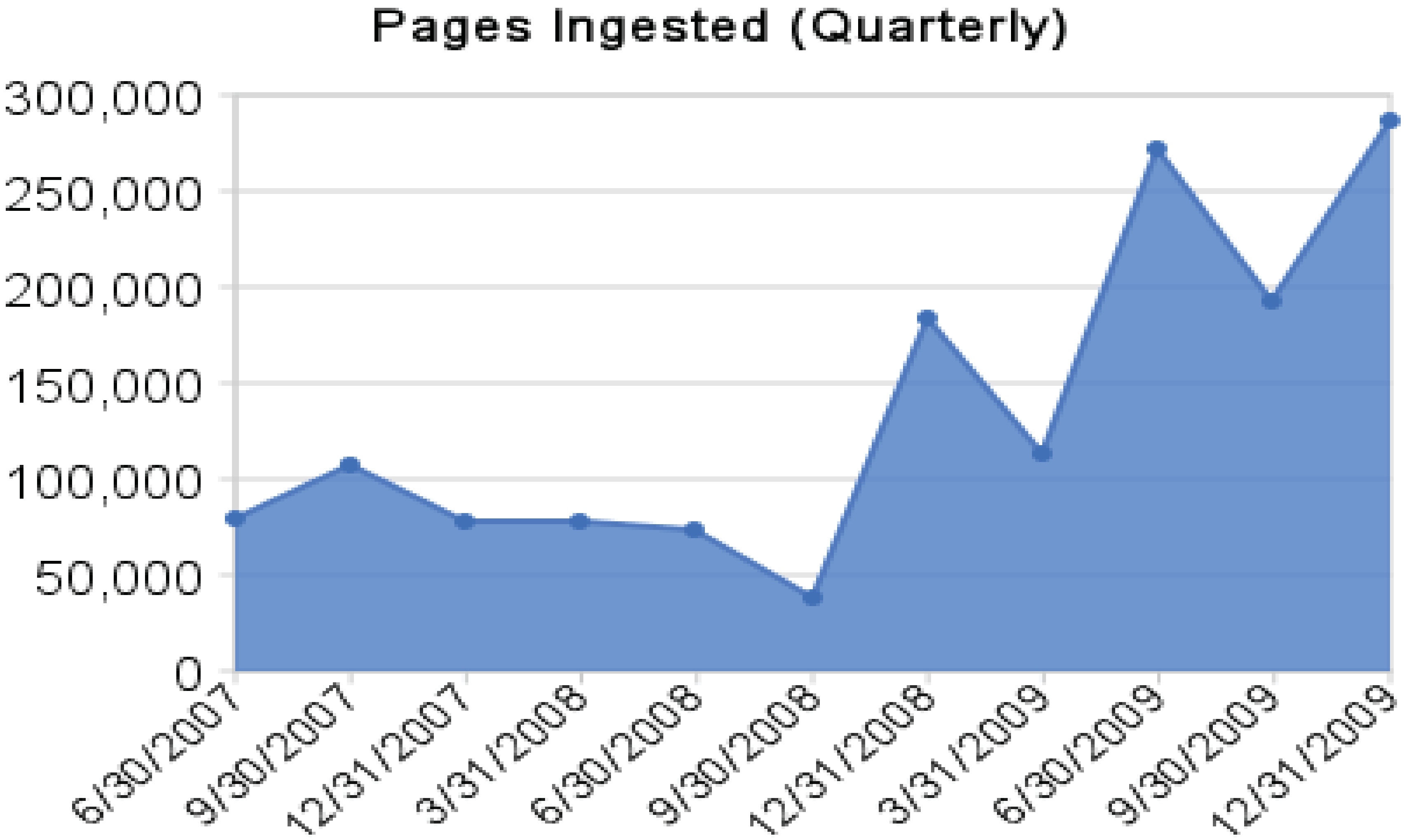
TECHNICAL ARCHITECTURE



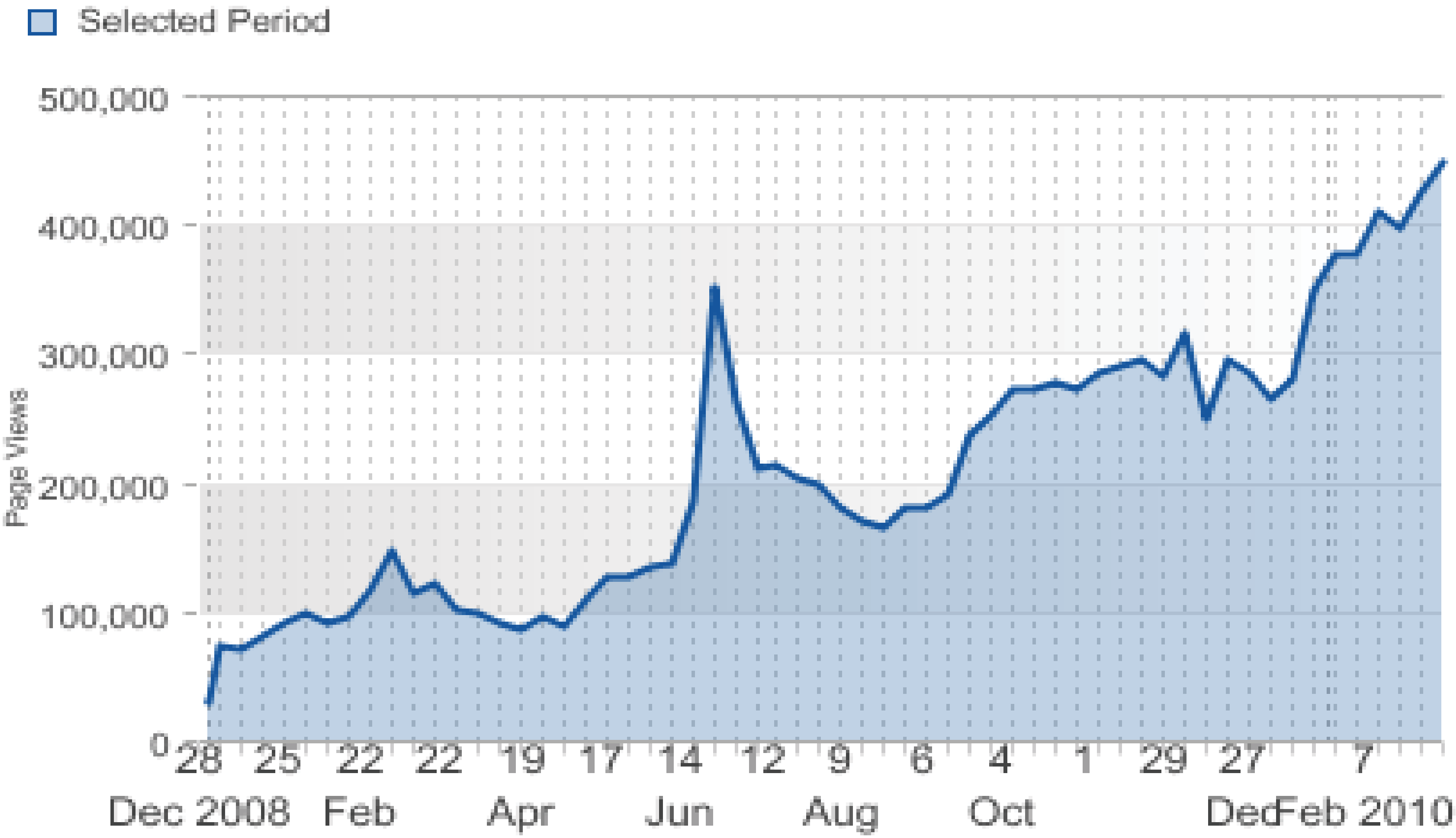
STATUS

- 20 U.S. State partners
- 1,700,000 newspaper pages ingested.
- 140,000 holding records.
- 3.6 million digital objects
- Automated ingest
- 50 Tb of content indexed and made available in few hours.
- 100,000 newspaper pages transferred and ingested per month
- Persistent identifiers and locators
- Enhanced discoverability: Open to crawlers & search engines
- Scalability and performance of access
- Content use: Flickr, Mashups, NSF's Digging into data

Ingest Throughput



Site Traffic



eDeposit

Overview

- Content in various formats from multiple sources (starting with eJournals)
- Capture and transfer content through Copyright Office
- Content accessible through LC Catalog Systems
- Integrated with various divisions' workflows
- Automated, scheduled transfer and ingest
- Does not require high technical capability from senders and system operators

STATUS

- Prototype successfully developed and tested
- Parts of curatorial access features transferred to NDNP
- Transfer system deployed and tested in production.
- LS workflows & system interfaces developed.
- Copyright workflows & system interfaces developed.
- New regulation on demand deposit published!

Next steps!

- Preliminary ingest services
  - Based on Bags
  - Semantic mapping of Bag files to digital objects (items)
- Bit preservation services
  - Applied to Bags initially
  - Applied to ingested files thereafter
- Access services
  - Item-level access
  - Persistent URL’s
  - Repository API’s

Challenges

- Managing expectations
- Repository infrastructure vs. content projects
- Software development process
- Resources
  - Priorities
  - Stability
- New technologies