Hydra/Fedora

August 4, 2014

Library IT

Brief History of YUL Digital Collections

- CONTENTdm
- Greenstone
- Custom systems (Luna, Portfolio, dbtext, Filemaker Pro, Excel, etc.)
- ODAI
- Fedora (standalone collections, e.g., AMEEL, YFAD)

Fedora is...

- Flexible Extensible Digital Object Repository Architecture
- Open Source
- Used by hundreds of organizations
- Originally developed at Cornell, now led by Fedora Project Steering Group under stewardship of DuraSpace.org
- (http://www.fedora-commons.org)
- Currently engaged in development of Fedora 4

Hydra is...

- A Repository Solution
- A Community (25 partners now, including us)
- A Technical Framework
- Open Source Software
- www.ProjectHydra.org

If you want to go fast, go alone.

If you want to go far, go together.

Hydra “Heads”

- Blacklight (for viewing)
- Ladybird (de facto)
- Avalon (A/V)
- Sufia (ScholarSphere)

Hydra Partners

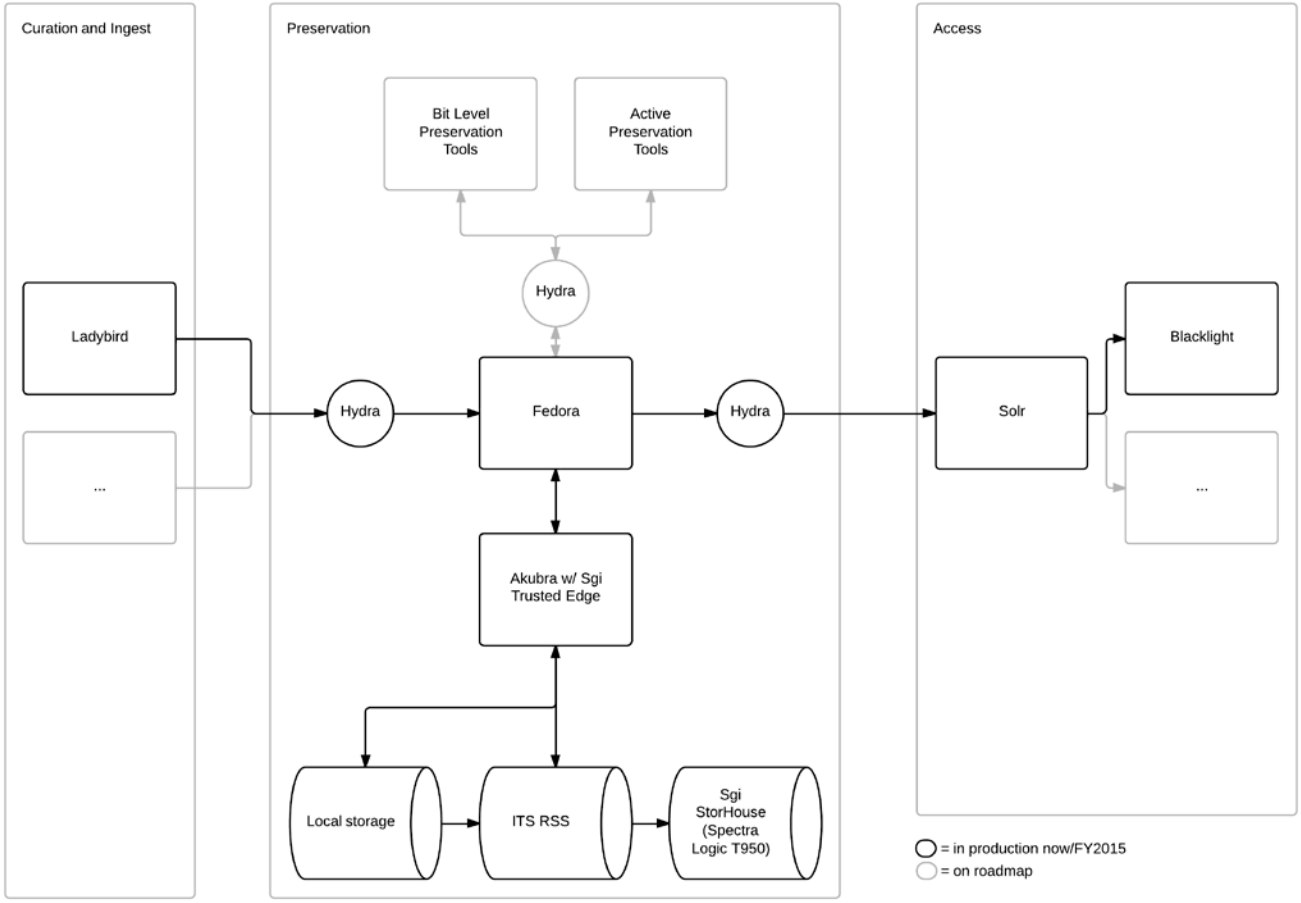
- Duraspace
- Stanford University
- University of Hull
- University of Virginia
- MediaShelf
- University of Notre Dame
- Northwestern University
- Columbia University
- Penn State University
- Indiana University
- London School of Economics
- University of Oregon
- Rock and Roll Hall of Fame
- Royal Library of Denmark
- Data Curation Experts
- WGBH
- Boston Public Library
- Duke University
- **Yale University**
- Virginia Tech
- University of Cincinnati
- Princeton University
- Cornell University
- Case Western Reserve Univ.

Benefits of ongoing investment

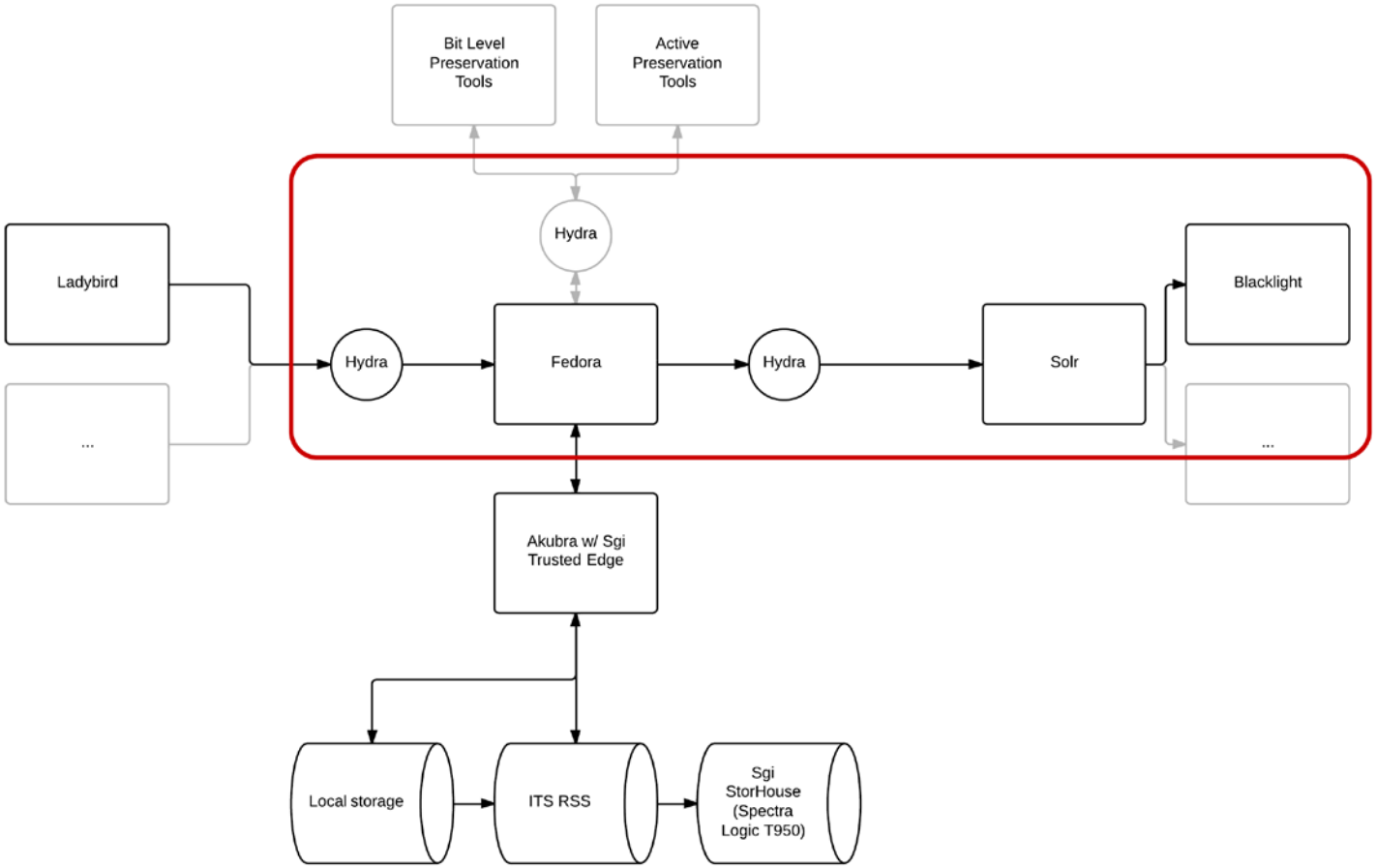
- Alignment with the Yale University Library's commitment to the stewardship of digital collections and content
- Unified, consistent, and efficient approach to long term access and retention
- Provide a consistent user experience across many collections and content types, along with discoverability
- Low risk of information loss
  - 4 copies of an object across 3 locations (New Haven, West Haven, Glastonbury) on 2 storage platforms
  - Internal integrity validation (checksum)
  - Media refreshing and replacing
- Low cost (compared to non-Yale service providers)

Software Architecture

Current/FY2015 Implementation

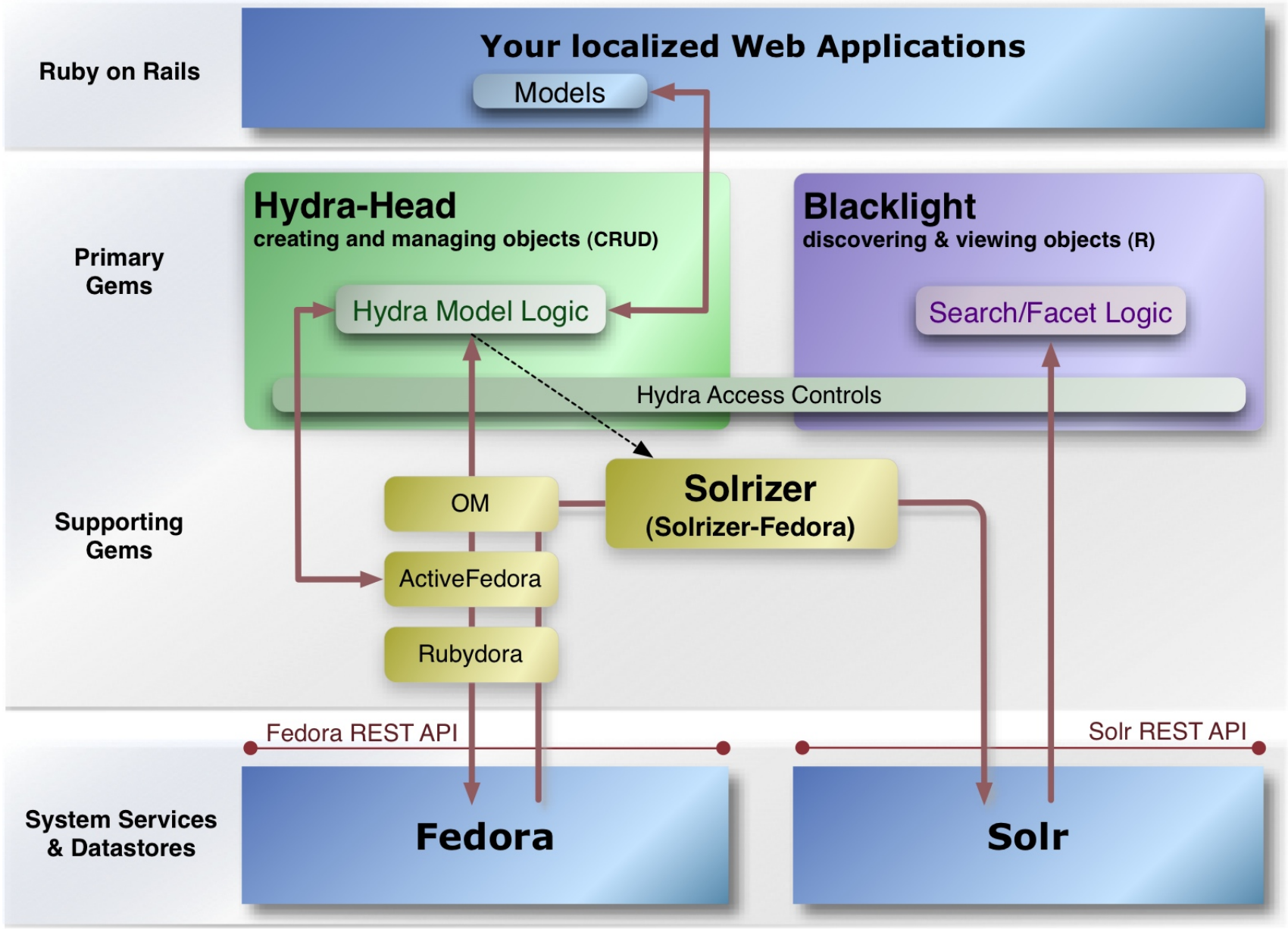


Hydra Project

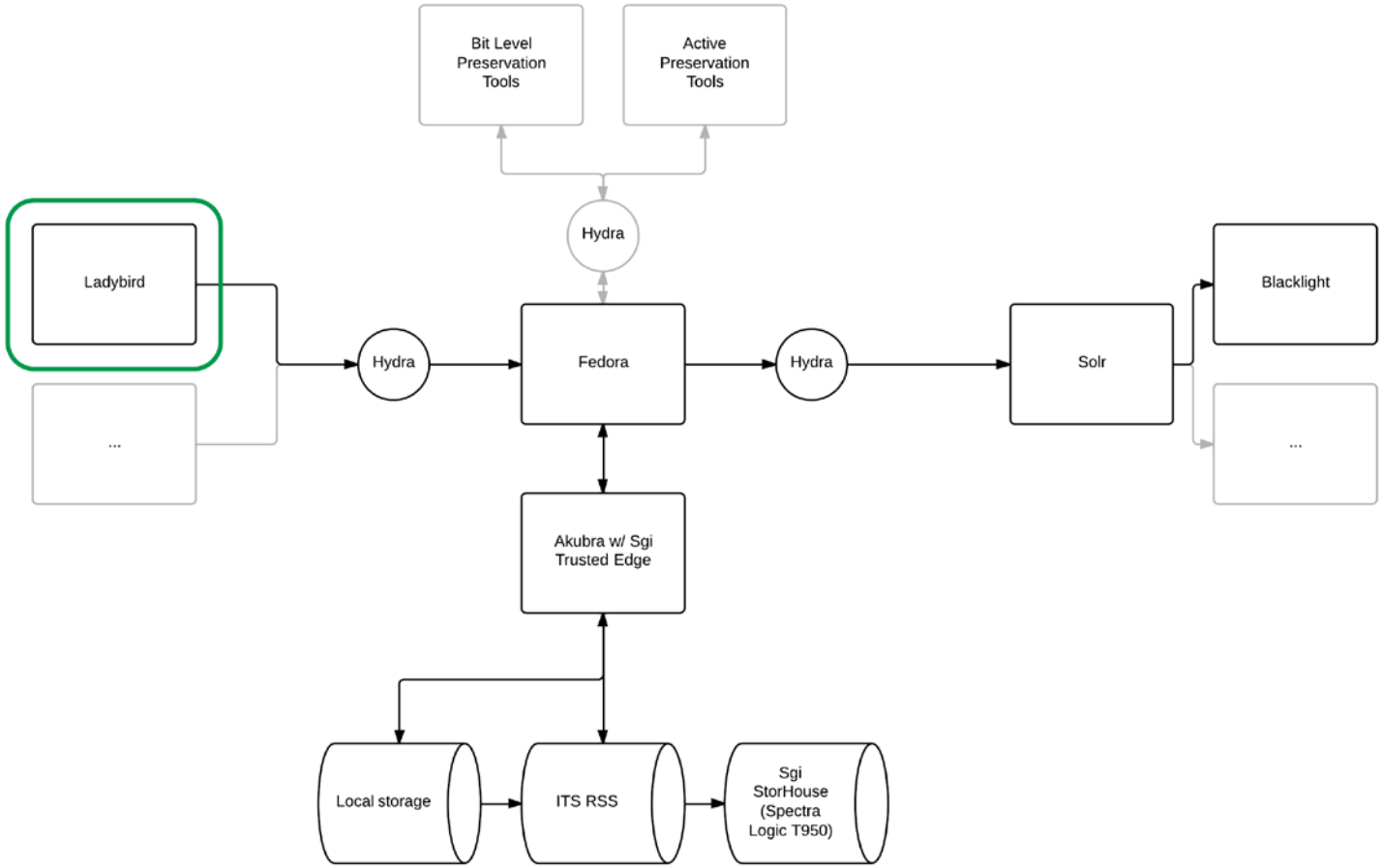


Hydra Stack

- Fedora
- Blacklight
- Ladybird
- Active Fedora
- Apache Solr
- Media Server
- Internet Archive Book Reader
- Ingest applications


Ruby on Rails
Your localized Web Applications
Models
Primary Gems
Hydra-Head
creating and managing objects (CRUD)
Hydra Model Logic
Blacklight
discovering & viewing objects (R)
Search/Facet Logic
Hydra Access Controls
Supporting Gems
OM
Solrizer
(Solrizer-Fedora)
ActiveFedora
Rubydora
Fedora REST API
Solr REST API
System Services & Datastores
Fedora
Solr

Ladybird


Bit Level Preservation Tools
Active Preservation Tools
Hydra
Ladybird
...
Hydra
Fedora
Hydra
Solr
Blacklight
...
Akubra w/ Sgi Trusted Edge
Local storage
ITS RSS
Sgi StorHouse (Spectra Logic T950)

What is Ladybird

LadyBird is a Hydra-compliant group of web-based and client applications designed to process digital collections including metadata management and digital media for both reformatted items and born-digital content across the Yale University Libraries.

LadyBird routes content to the Hydra/Fedora repository which in turn exposes content through our public discovery/access system, Blacklight.

Ladybird Goals

- Centralize image cataloging into a single tool
    - Luna, Portfolio, DB Text, Excel, FileMaker Pro, CONTENTdm
- Provide vocabularies that could be shared across the library
    - Potential for integrating Getty vocabularies and Linked Data
- Simplify the ingest of assets into the DAM hosted by YDC2
- Migrate content off Rescue Repository
- Simplify IT Support by having One System to manage

Ladybird

- Started June 2010
- Version 1.0 December 2013

- 20 background applications
- 4 desktop applications
- 3 web applications

- C# .Net 4.0
- 575,000 lines of source code

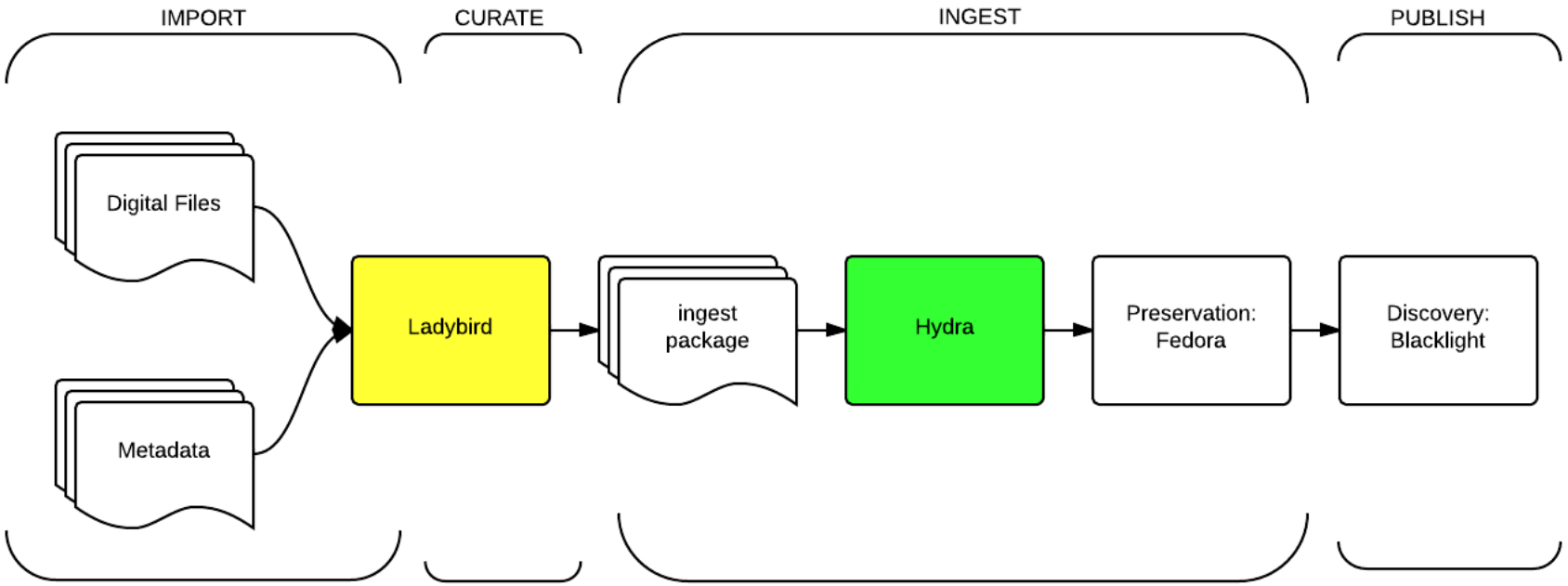
- 2,449,839 assets
- 2.5 mil on deck
- Growth: 1,500 assets per day

- 3 Microsoft SQL databases
- 360GB of raw data
- 20 TB files staged
- 40 TB to import

- A Jazz song by Tadd Dameron

Ladybird with Hydra

Import, Curate, Ingest, Publish



Ladybird Roadmap

- Potential partners with:

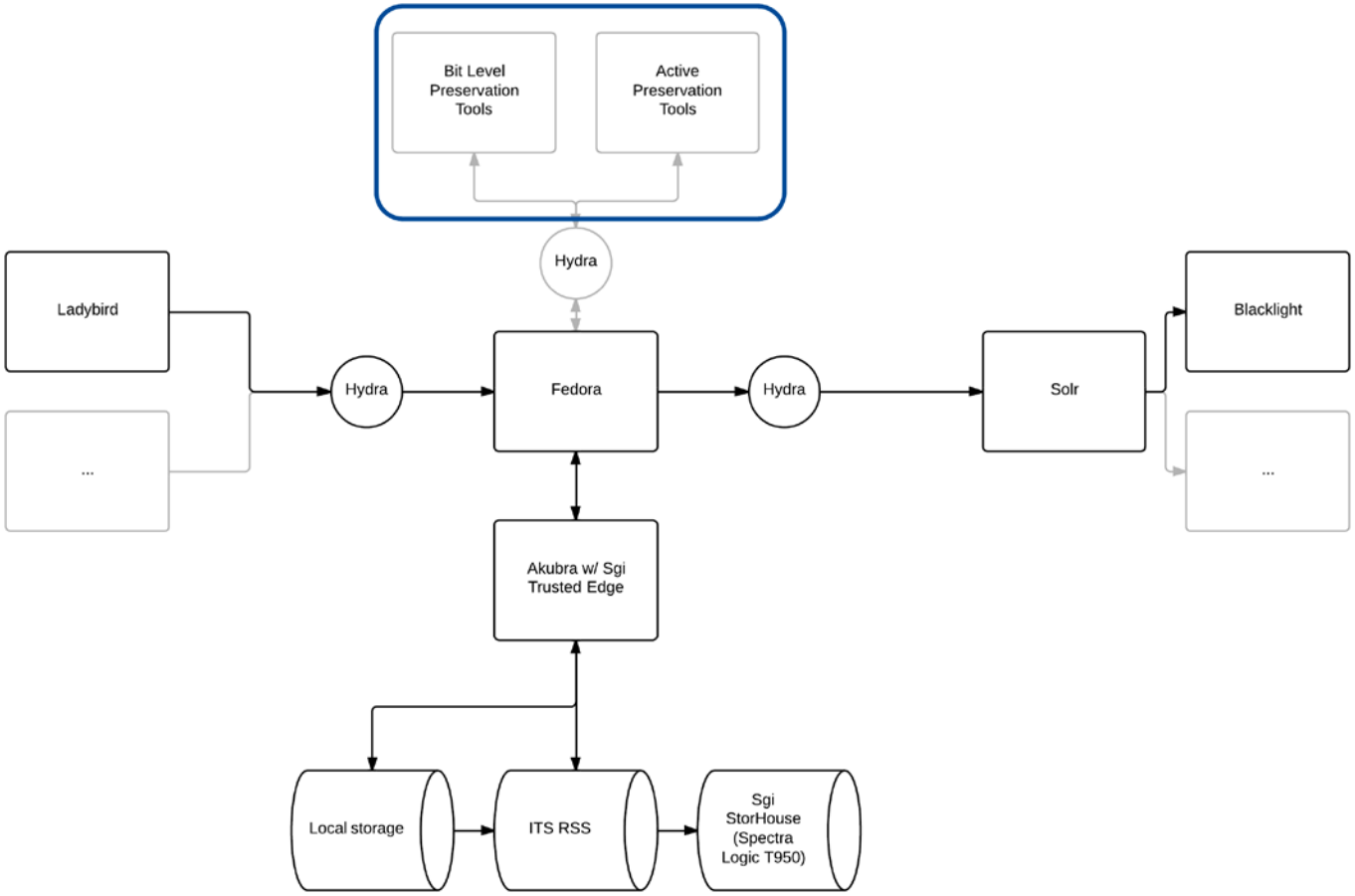
  Columbia, Princeton, MIT, Northwestern
- Release Ladybird as Hydra Head
- Collection migration this fall
- Platform migration to Java 8, MySQL

Hydra Roadmap

- Blacklight 5.x
- Fedora 4
- Open Archival Information System (OAIS) ingest model
- Workflow System Architecture
- Digital Preservation Interfaces
- Sufia – Faculty Self Archiving
- Avalon – A/V support
- Spotlight – Exhibitions
- Auditing – Statistics and Audit Trails

Preservation

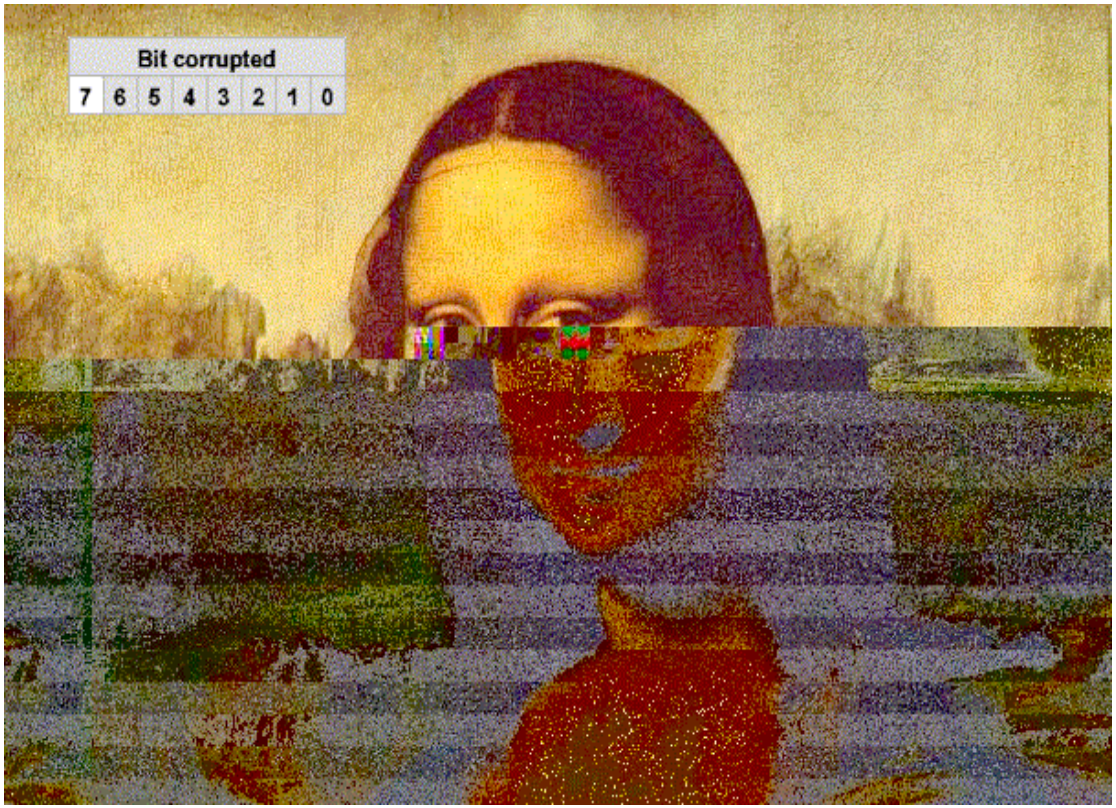
Preservation Tools



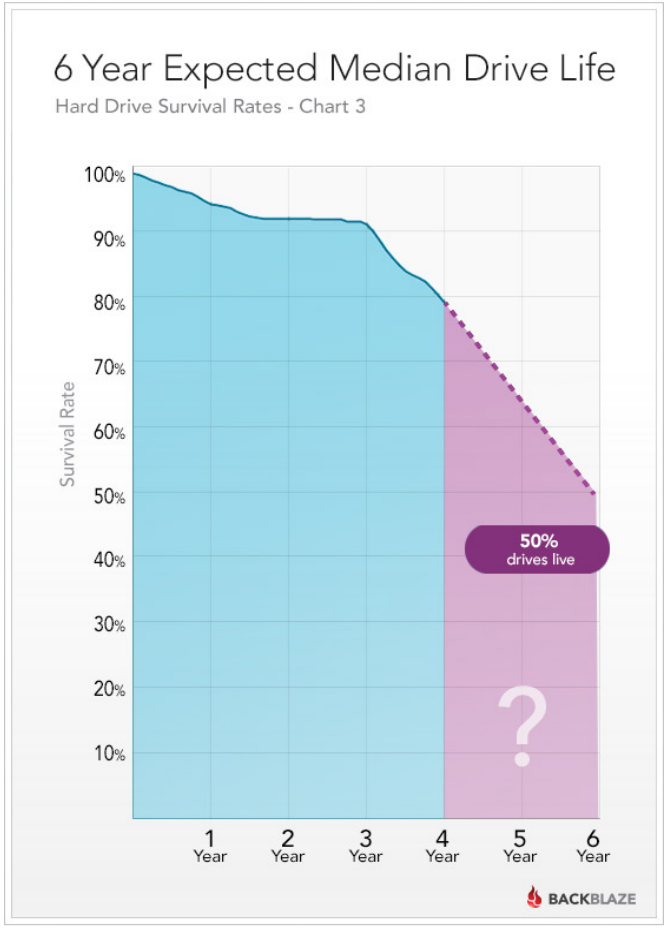
“Digital Information lasts forever or 5 years, whichever comes first”

Jeff Rothenberg. Scientific American, January 1995.

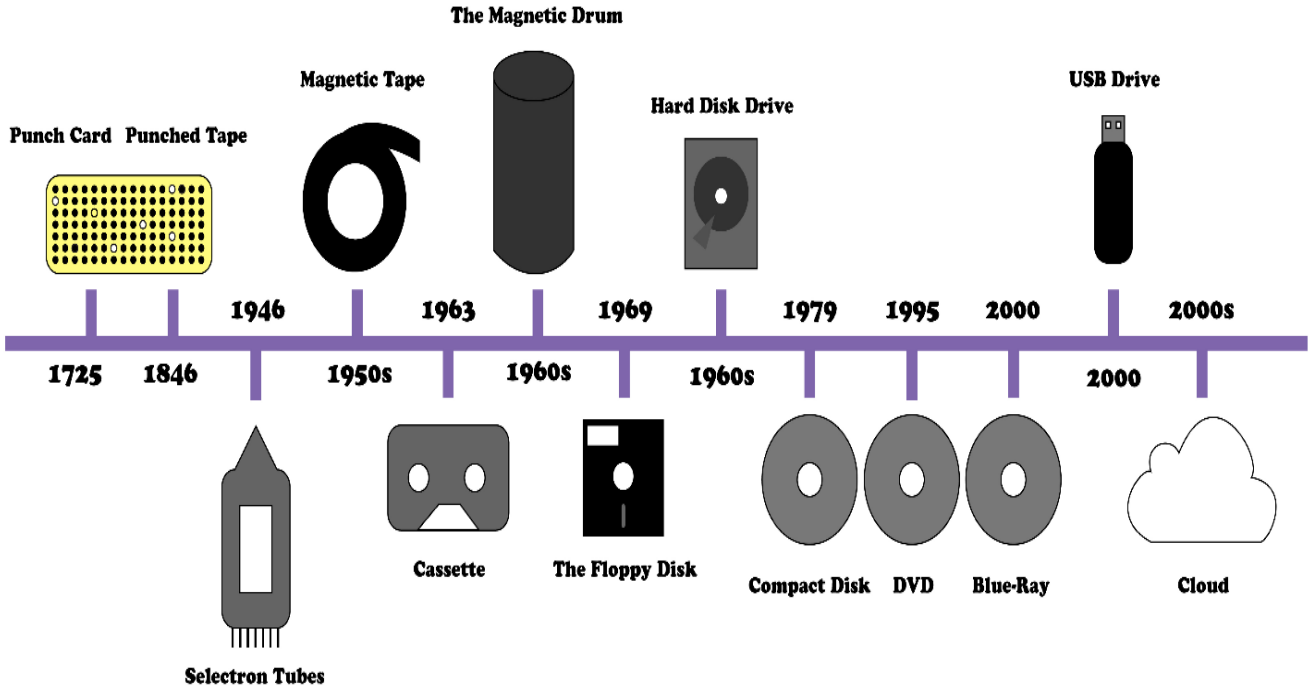
Digital Preservation Challenge: Bit Rot



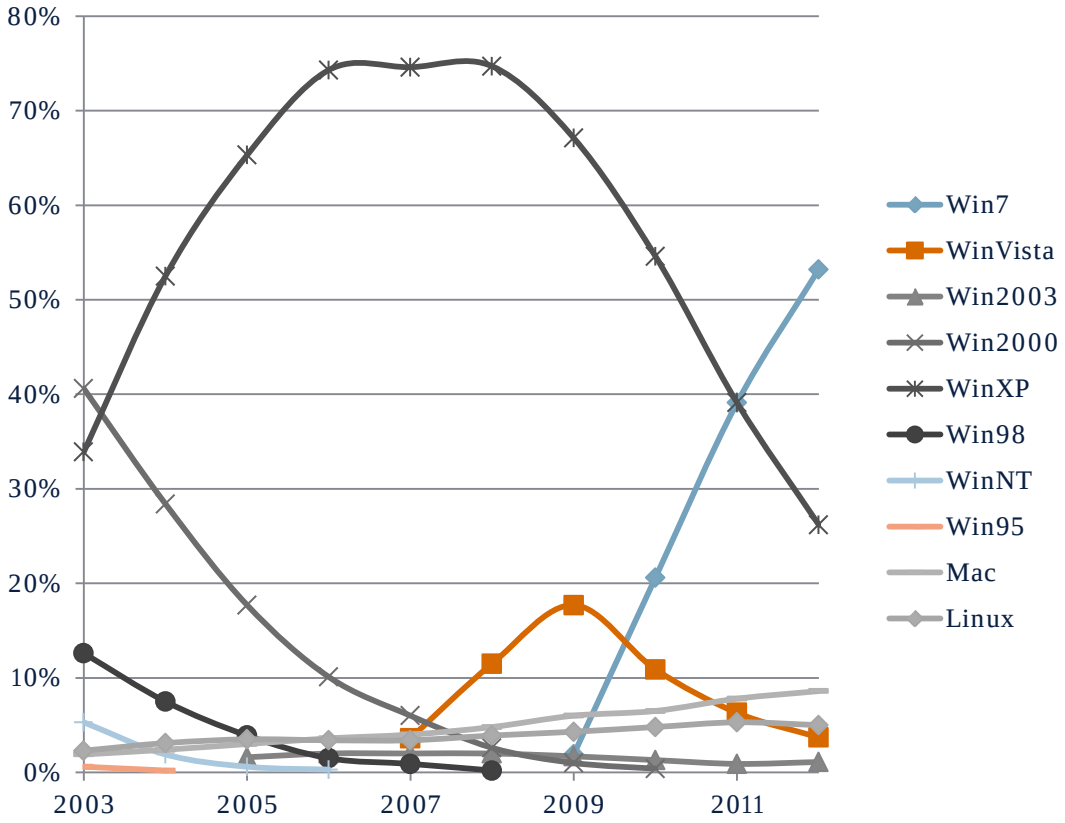
Digital Preservation Challenge: Hardware Failure



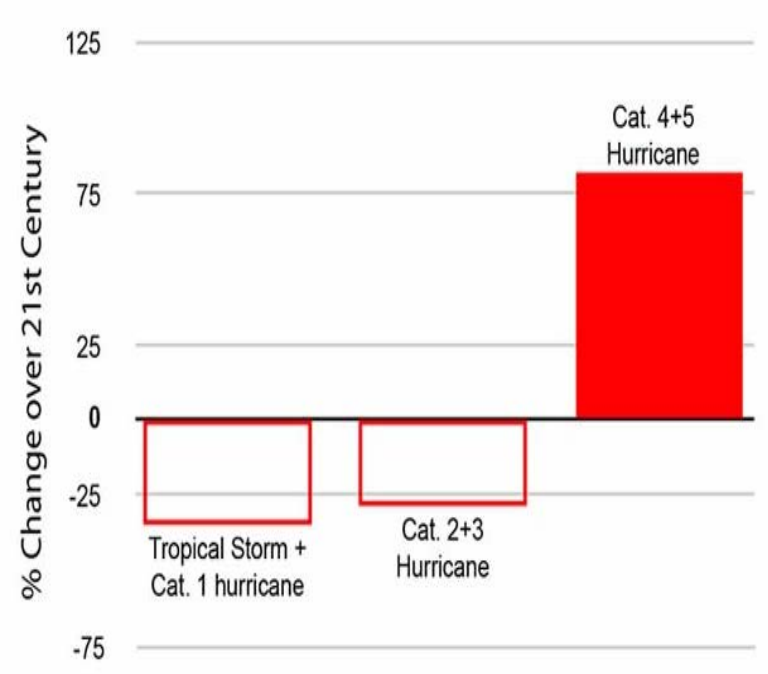
Digital Preservation Challenge: Hardware Obsolescence



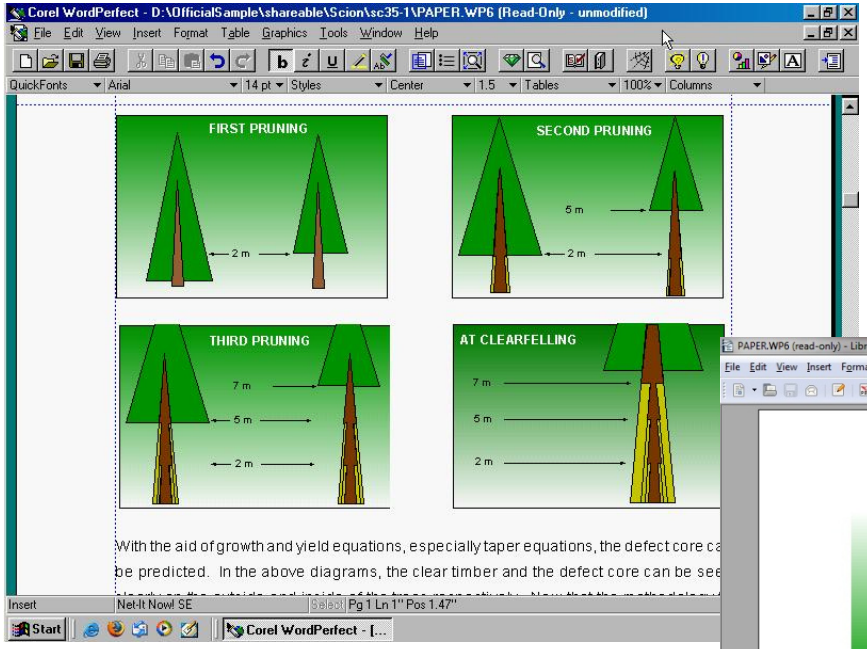
Digital Preservation Challenge: Software Obsolescence



Digital Preservation Challenge: Natural Disasters



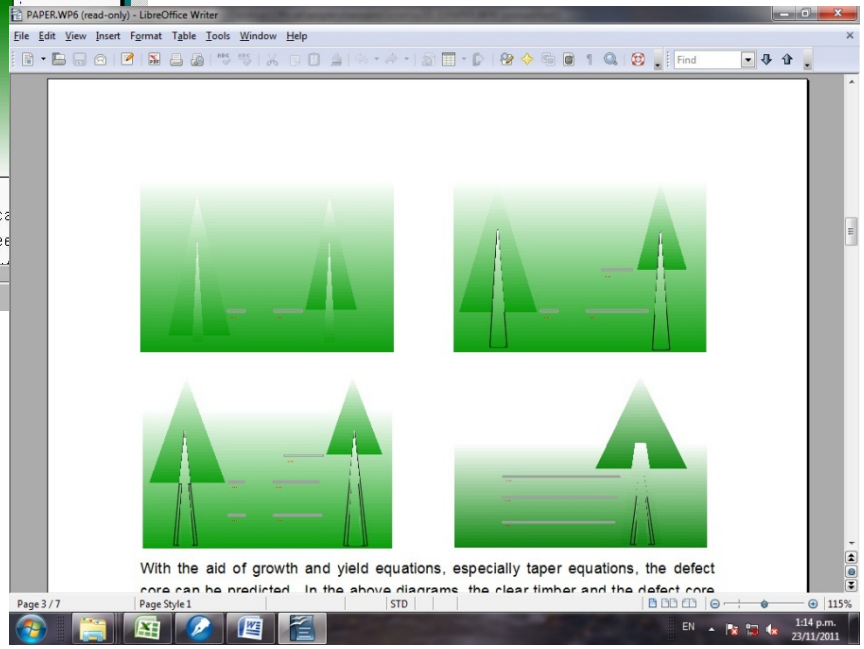
Digital Assets Degrade Without Maintenance



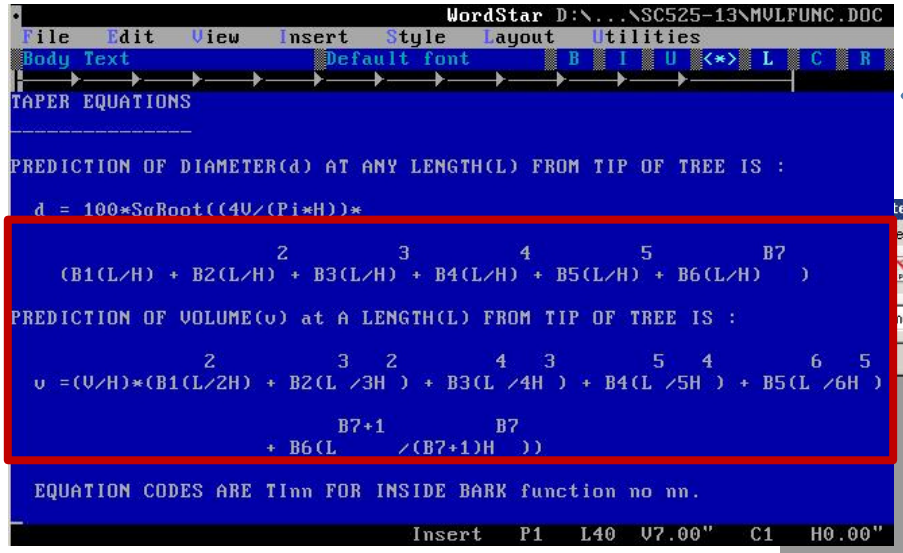
Original digital asset includes visual data

Modern software alters this data:

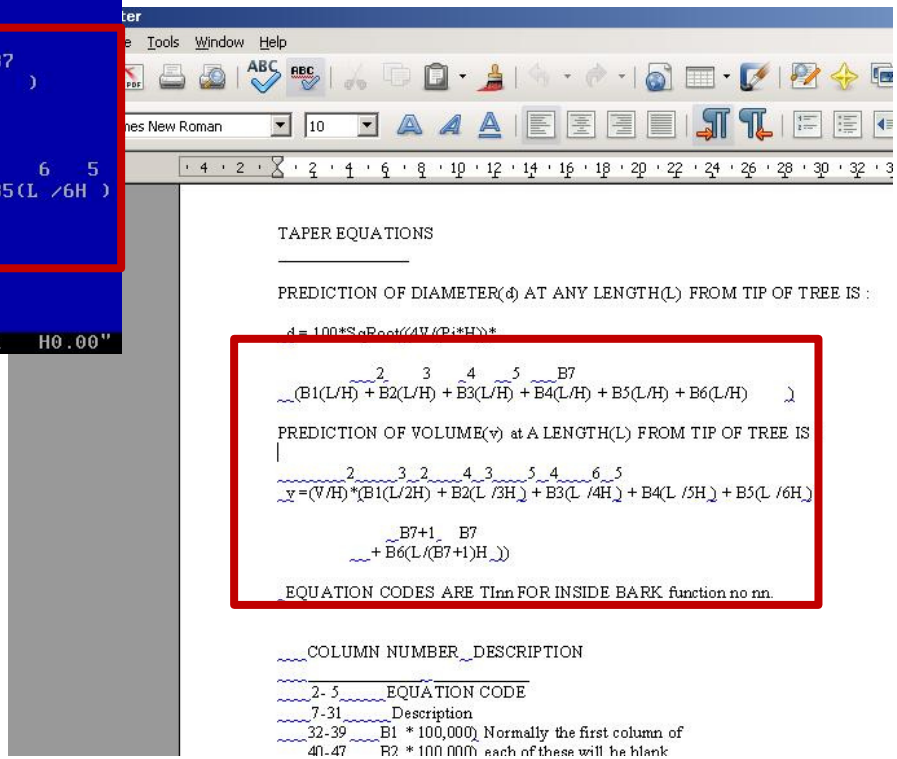
- Changing its meaning
- Reducing the asset's value



Inaction will Reduce Asset Value



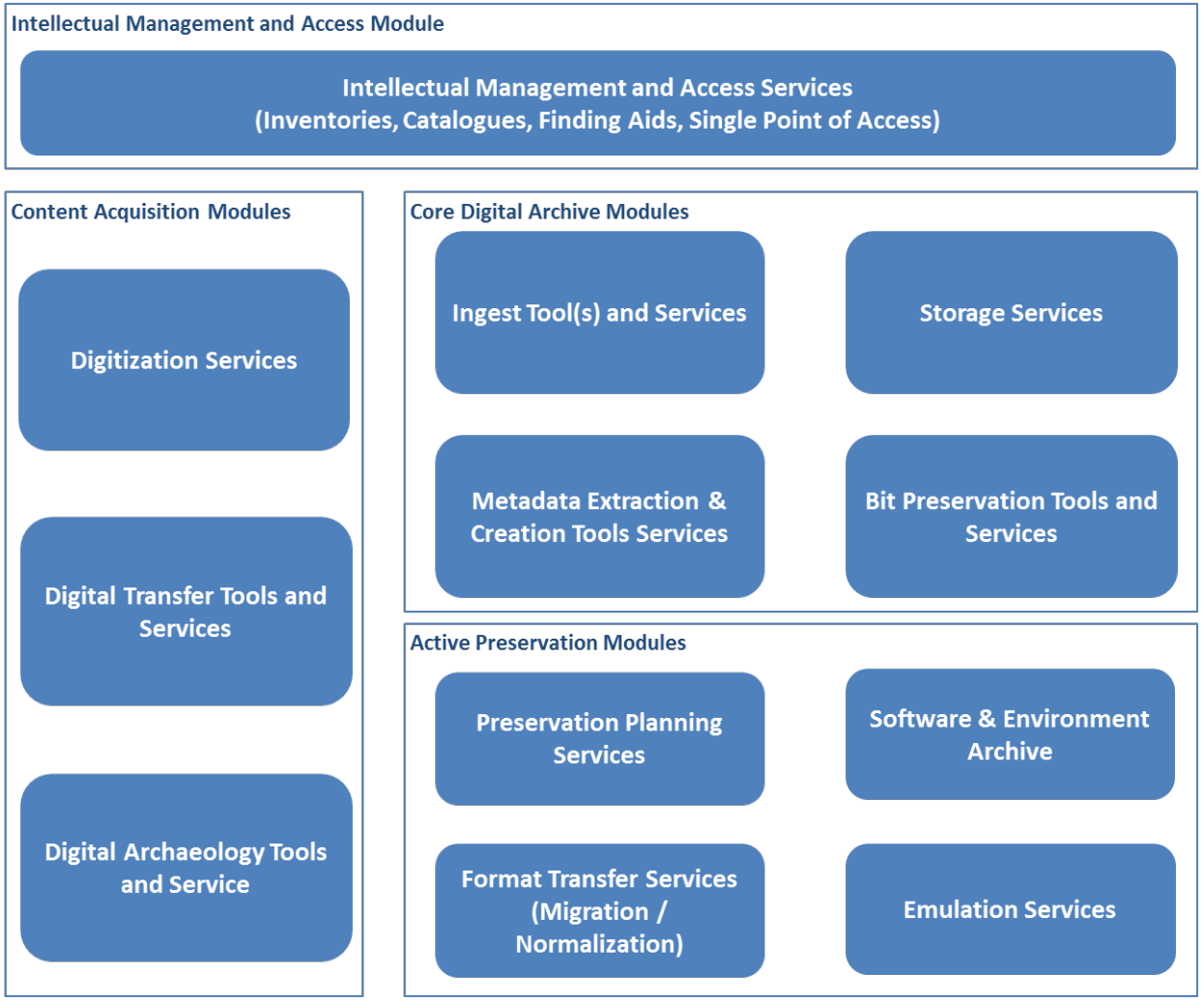
Original digital asset includes important equations



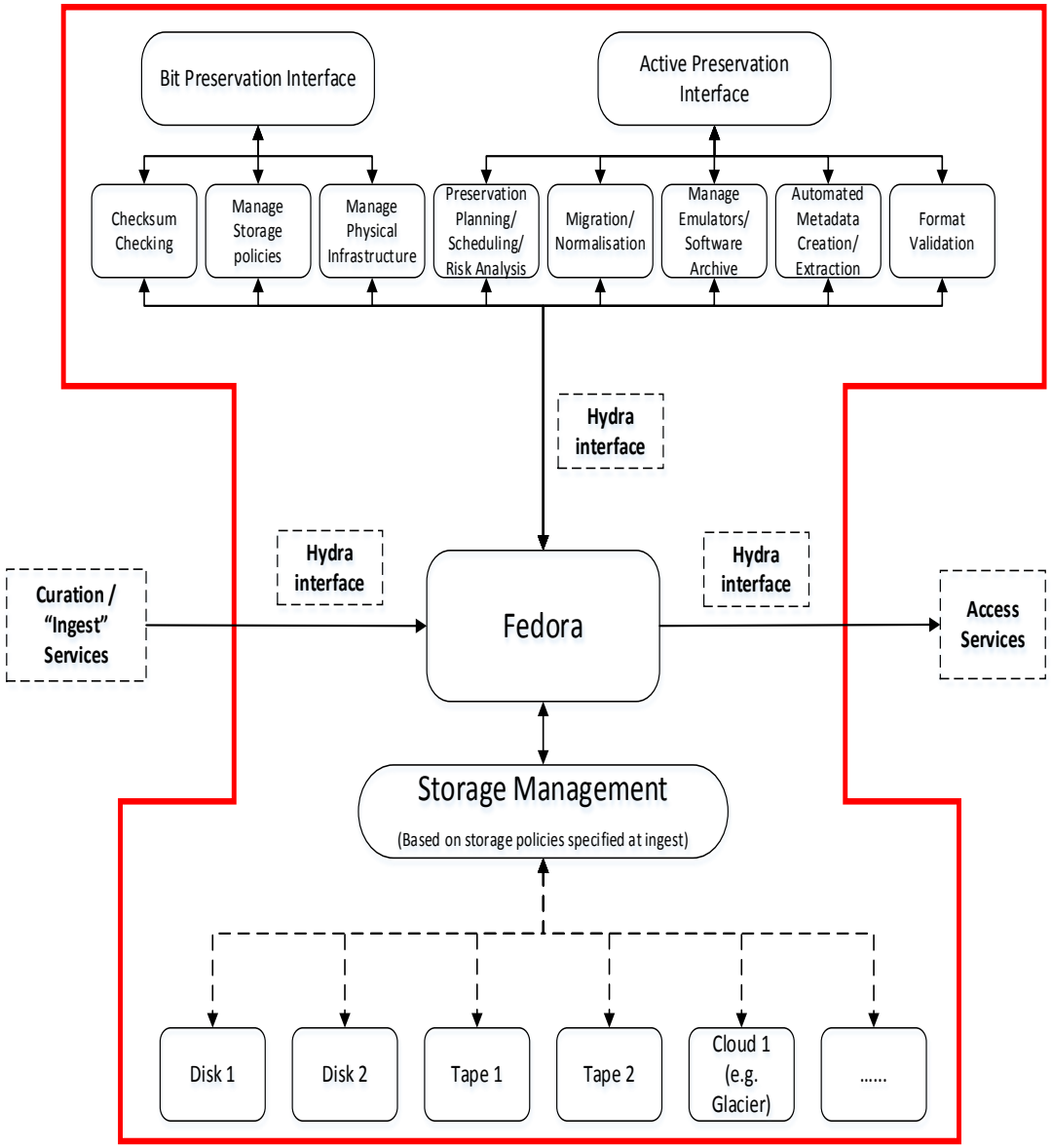
Modern software alters these equations:

- Changing their meaning
- Removing trust in information
- Destroying the asset's value

Digital Preservation Tools & Services



Proposed Digital Repository and DPS Architecture



Proposed Basic Digital Preservation Services

Bit Preservation

At least 4 copies, stored in at least 3 locations with different risk profiles, regularly monitored, with seamless media & software management (refreshment, replacement, etc)

Secure Storage with Managed Access

Audited secure storage with authorized, timely access and clear exit strategies

Obsolescence Monitoring

Identify technichal characteristics of files, associate with interaction software and hardware, software and hardware obsolescence monitoring, informing content owners when content is becoming inaccessible

Provenance and Authenticity Assurance

Logging & preserving all provenance events, ability to report on history of activities, checksum creation, independent storage and regular validation

Standards Compliance

Compliance with ISO 14721:2012: Open archival information system (OAIS) Reference model & with ISO 16363:2012: Audit and certification of trustworthy digital repositories

Digital Preservation Tools Roadmap

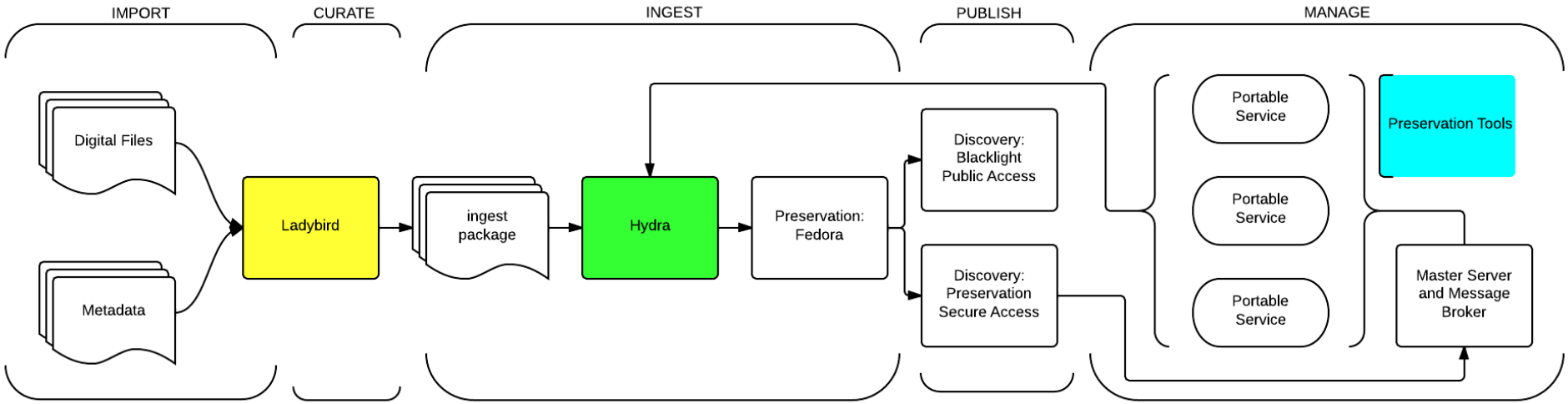
- Programming team formed
- Gathering use cases and user stories
- Platform selection

Simplest use case:

- Validate file: 17 sec average
- Validate current repository: 883 days

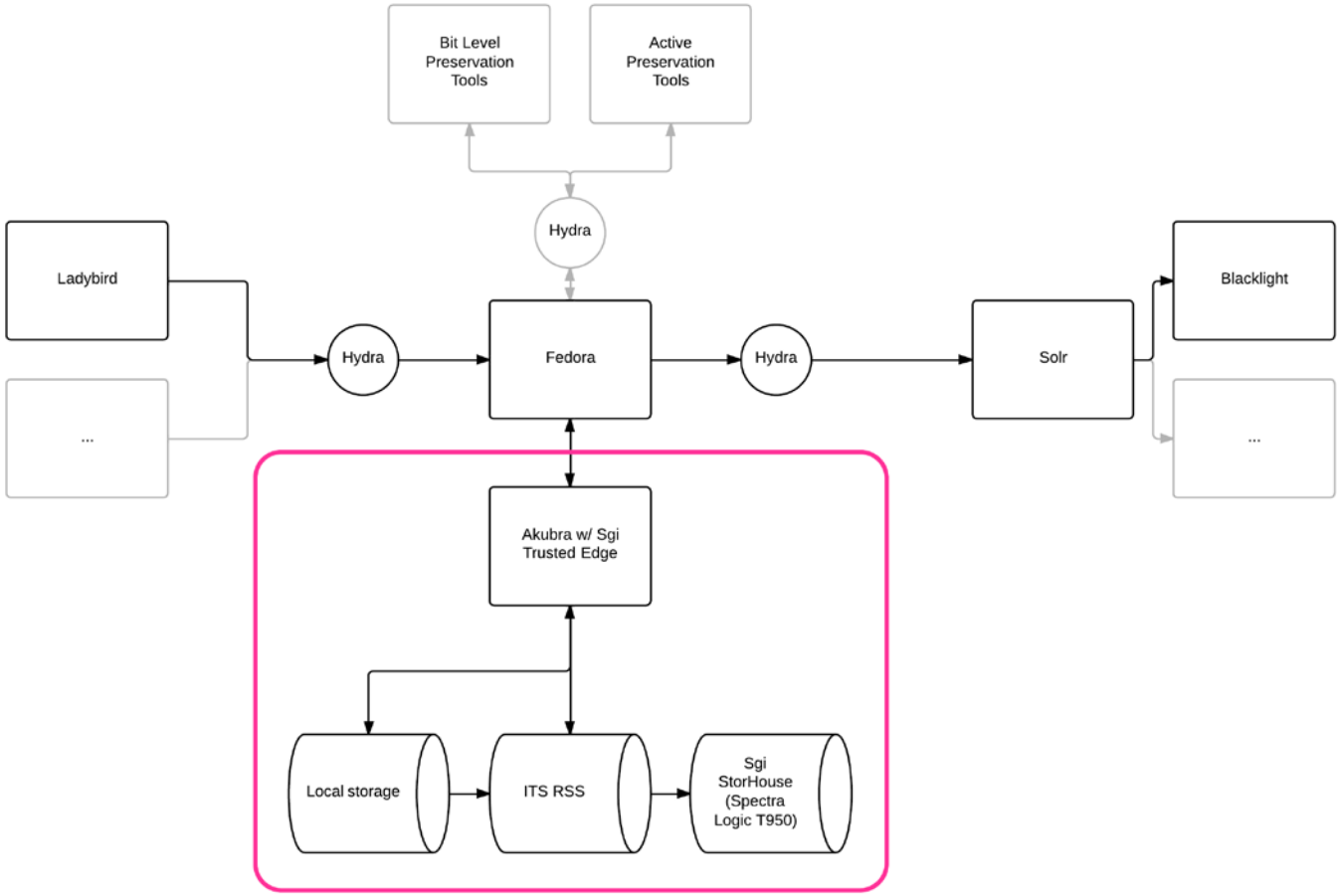
- Target: 1 day

Digital Preservation with Hydra

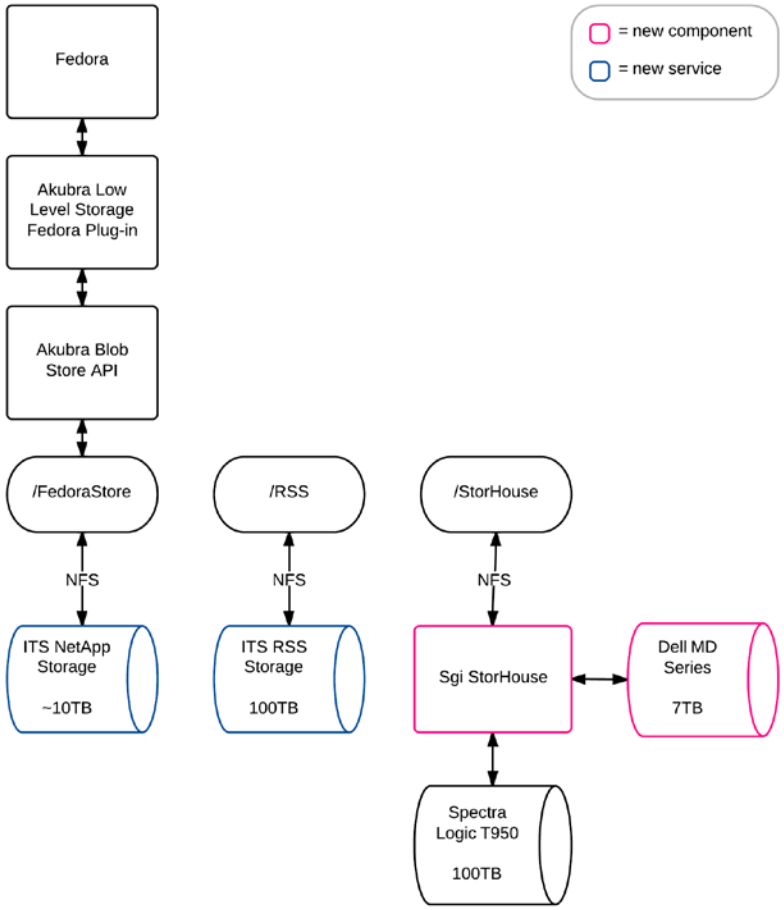


Infrastructure

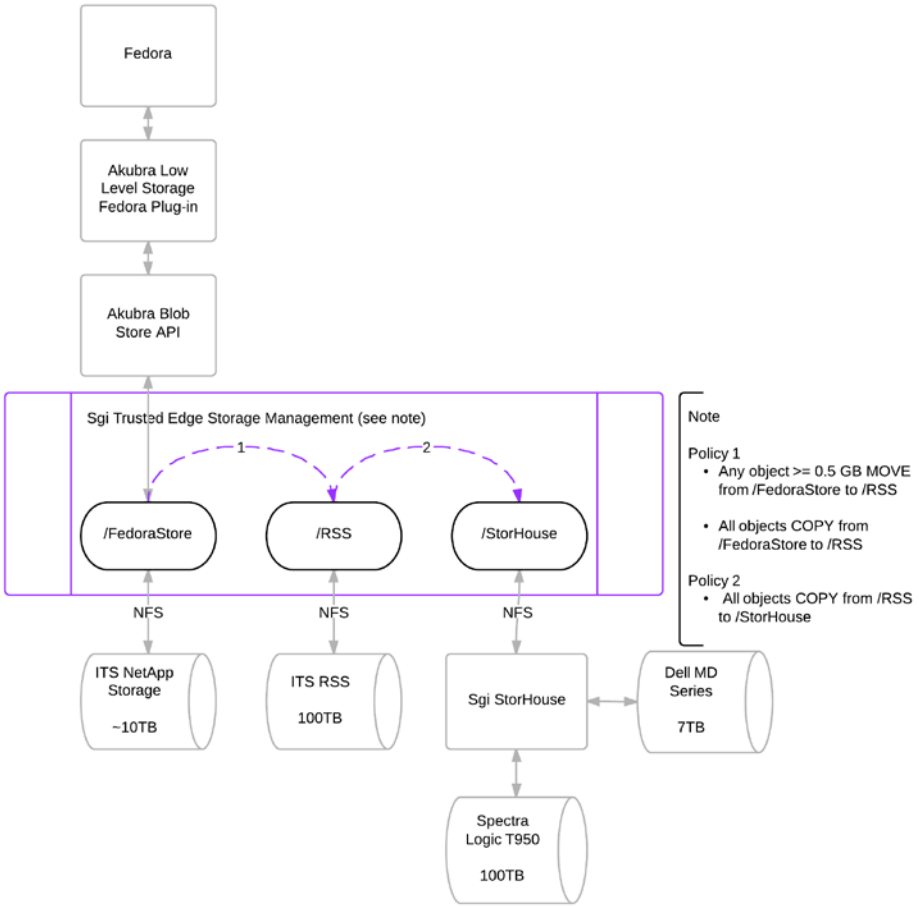
Storage Infrastructure



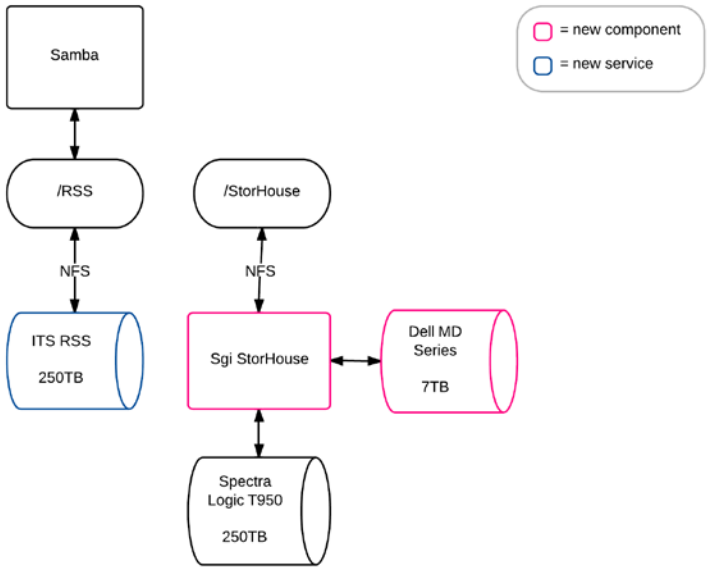
Proposed FY2015



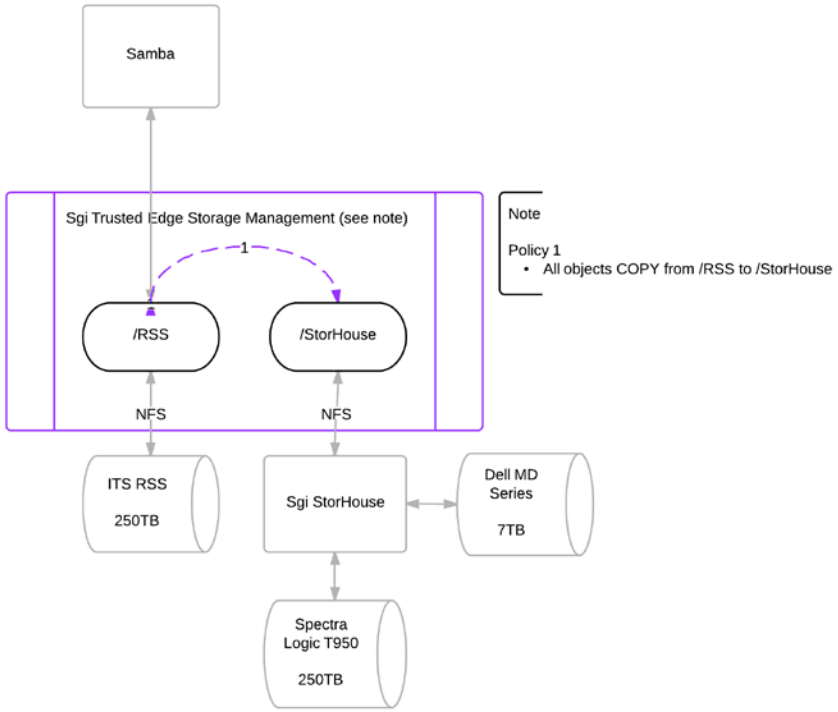
Proposed Trusted Edge Policy



Proposed FY2015 Staging



Proposed Staging Trusted Edge Policy



Storage Roadmap

Fall 2014

- Transition from NetApp storage to ITS RSS 2
- Stand-up Fedora 4 for testing. Configure and exercise new storage management layer (ModeShape/Infinispan).

Opportunities to explore

- Migration to Yale ITS Sgi StorHouse implementation
- ITS RSS 2 and/or HPC storage
- Out-of-region location for data replication
- Continue exploring external storage providers

A note about external storage providers

Service Provider	Cost per GB/Year	Endowment cost	Endowment Period	Content types accepted	# of Copies	Bit preservation?	Active Preservation?	Curation?	Access?
Chronopolis	$2.15	N/A	N/A	all	3	Y	N	N	N
Digital Preservation Network (DPN)	$0.83	$4.88/GB	20 years	all	3	y	N	N	N
Dspace Direct	$33.00	N/A	N/A	Limited	2 - 4	y	N	P	Y
DuraCloud	$1.11	N/A	N/A	all	2 - 4	y	N	N	P
HathiTrust	N/A	N/A	Permanent	Limited	3	y	P	N	Y
LOCKSS	N/A	N/A	N/A	Limited	N/A	y	P	N	P
OpenICPSR	$6	$60/GB	10 years	Limited	6	y	P	P	Y
Portico	N/A	N/A	N/A	Limited	"multiple"	y	P	P	Y
Preservica (Tessella)	$2.74	N/A	N/A	all	"multiple"	y	P	N	Y
DPS - Steady Growth	**$0.97**	**TBD**	**TBD**	**all**	**4**	**Y**	**Y**	**N**	**N**
DPS - Medium Growth	**$0.82**	**TBD**	**TBD**	**all**	**4**	**Y**	**Y**	**N**	**N**
DPS - High Growth	**$0.72**	**TBD**	**TBD**	**all**	**4**	**Y**	**Y**	**N**	**N**

Possible Future Paths

- Research Data support
- Support for A/V via Avalon
- Support for self-archiving of materials via Sufia (and later via Hydramata project)
- Active preservation tools
- Embedding content in LMS systems via LTI
- Support for exhibitions via Spotlight
- GeoBlacklight
- ORCID support
- Fedora 4 – active storage management, migration path