Digital Repository Initiative

01 April 2014

Ray Frohlich
Director, Enterprise Systems and Infrastructure

with

Euan Cochrane

Digital Preservation Manager, Preservation

Michael Friscia

Manager of Digital Library & Programming Services, Library IT

What is a digital repository?

A combination of people, processes and technologies, which together provide the means to capture, preserve and provide access to digital objects.

What is a digital preservation service (DPS)?

A combination of people, processes and technologies, which together provide the means to preserve digital objects.

Current state

- Existing content is curated, ingested, and accessed through a variety of systems, many custom built for the specific collection or project
- Within the next year YUL will have nearly a petabyte of digital objects that will need to be preserved long term
- Digital storage infrastructure is not robust, objects are at risk of loss, and content is spread across several disparate storage infrastructures
- Digital preservation software systems and services are mostly non existent

Benefits of further investment

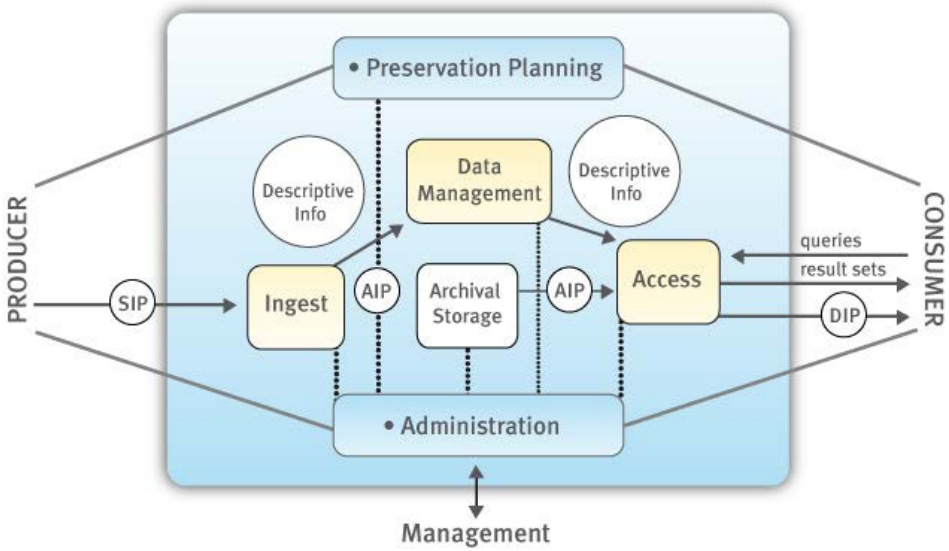
- Alignment with the Yale University Library's commitment to the stewardship of digital collections and content
- Unified, consistent, and efficient approach to long term access and retention
- Low risk of information loss
  - 4 copies of an object across 3 locations (New Haven, West Haven, Glastonbury)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

What will the repository contain?

- Digitized content
- Collections based born digital content
- Collections based metadata
- Vendor content
- Research data

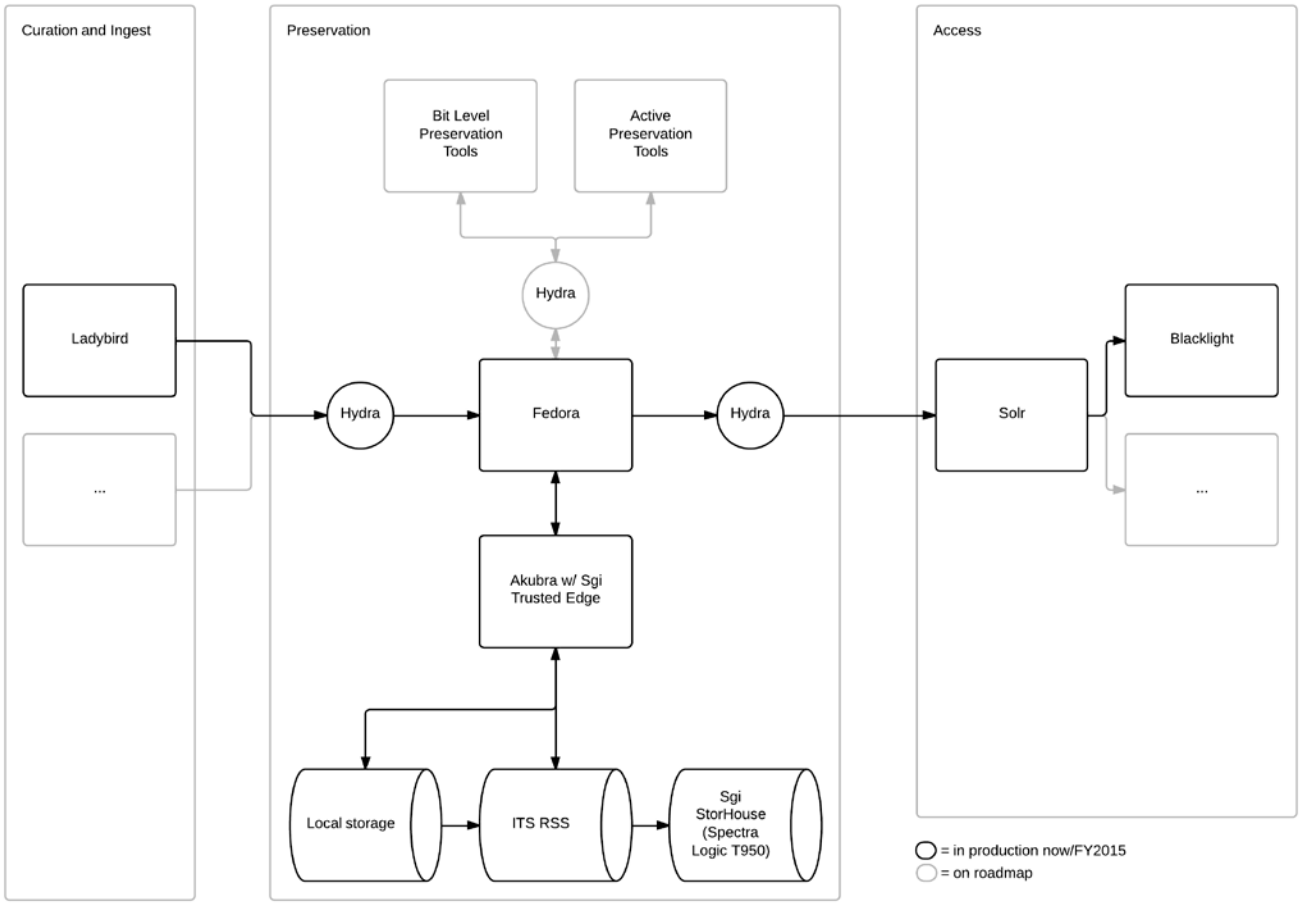
Reference architecture

The OAIS Reference Model (ISO Standard 14721)

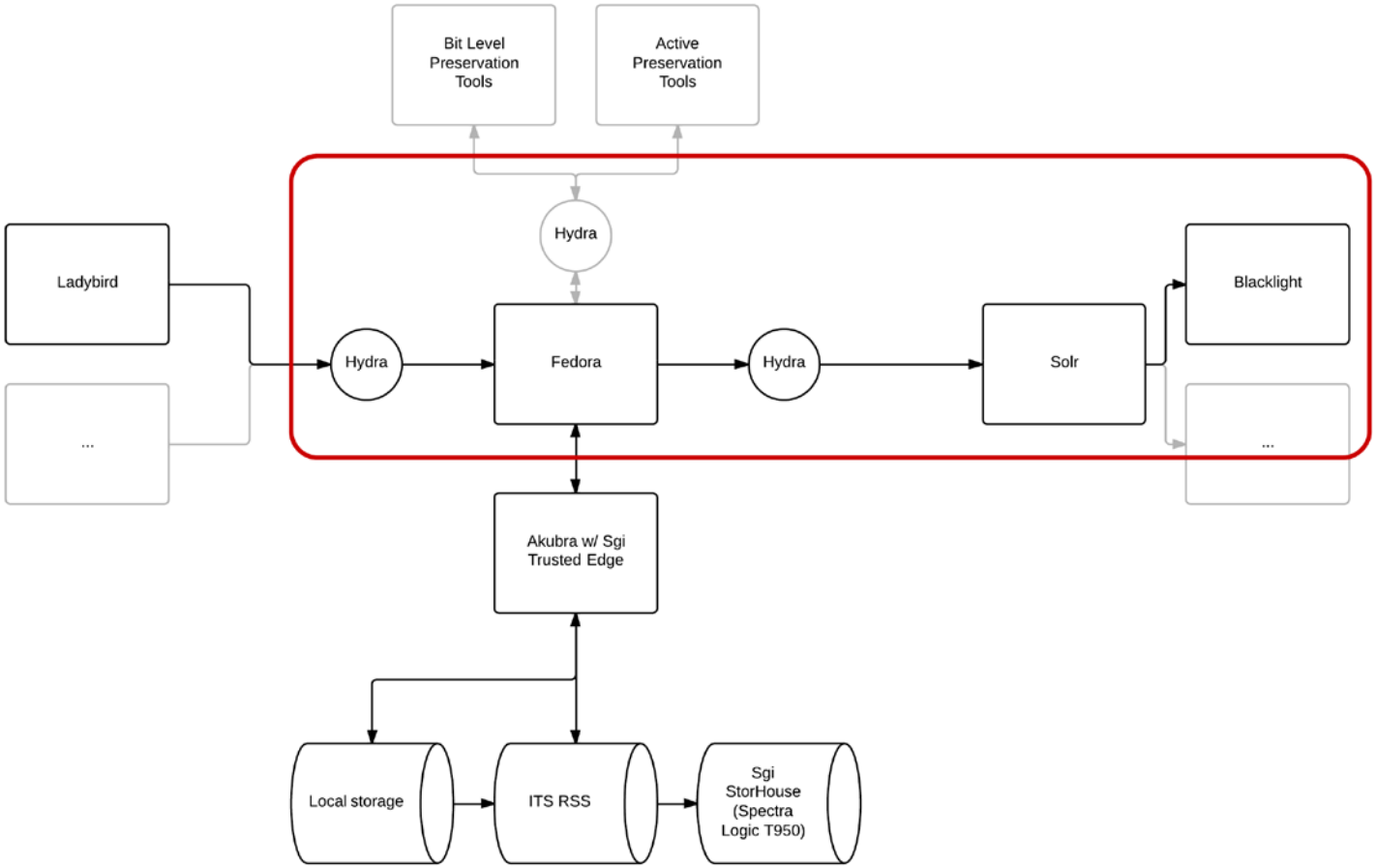


DIP: Dissemination Information Package
AIP: Archival Information Package
SIP: Submission Information Package

Current/FY2015 Implementation



Hydra Project



Hydra is...

- A Repository Solution
- A Community
- A Technical Framework
- Open Source Software
- www.ProjectHydra.org

If you want to go fast, go alone.

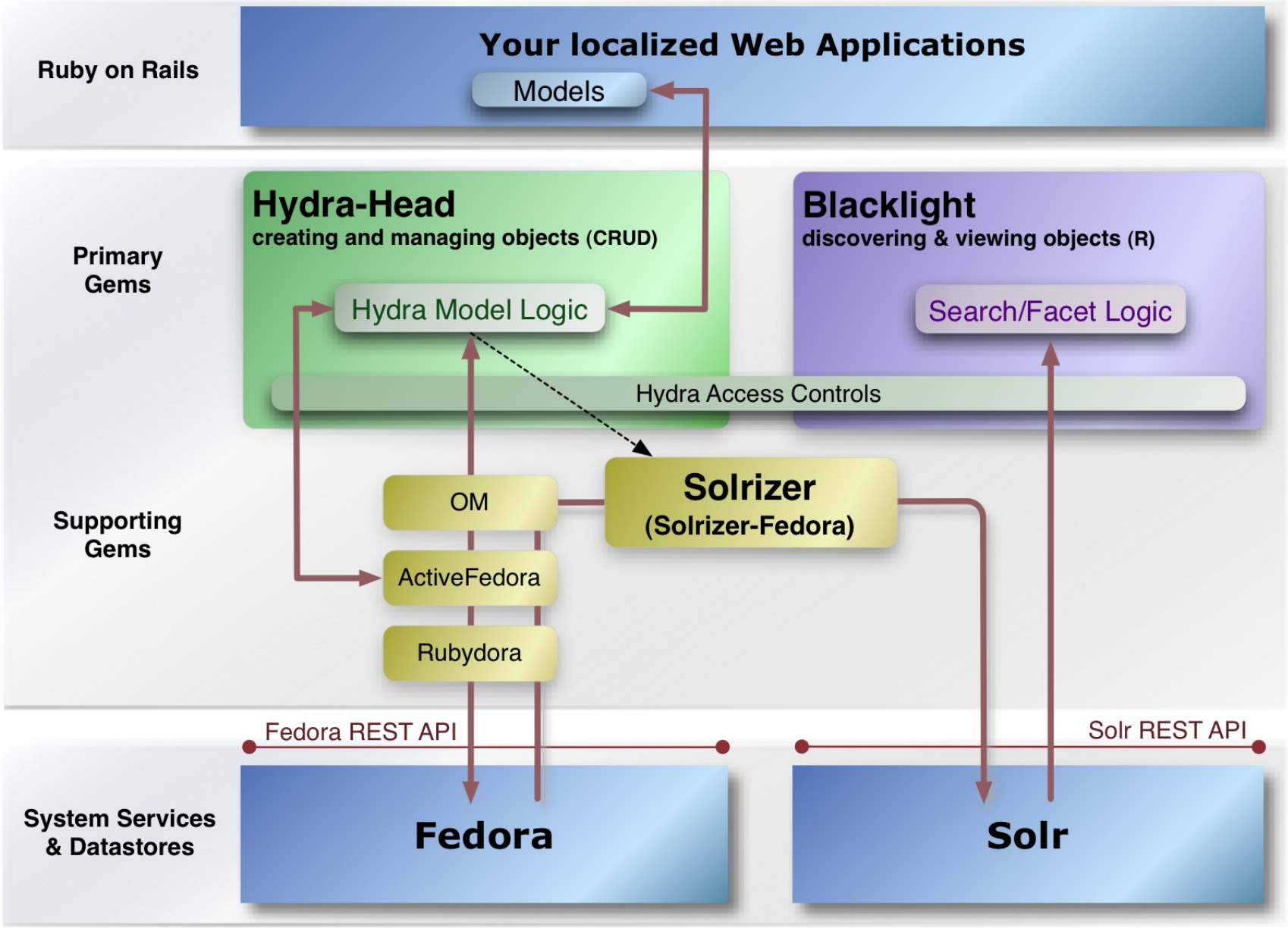
If you want to go far, go together.

Hydra Partners

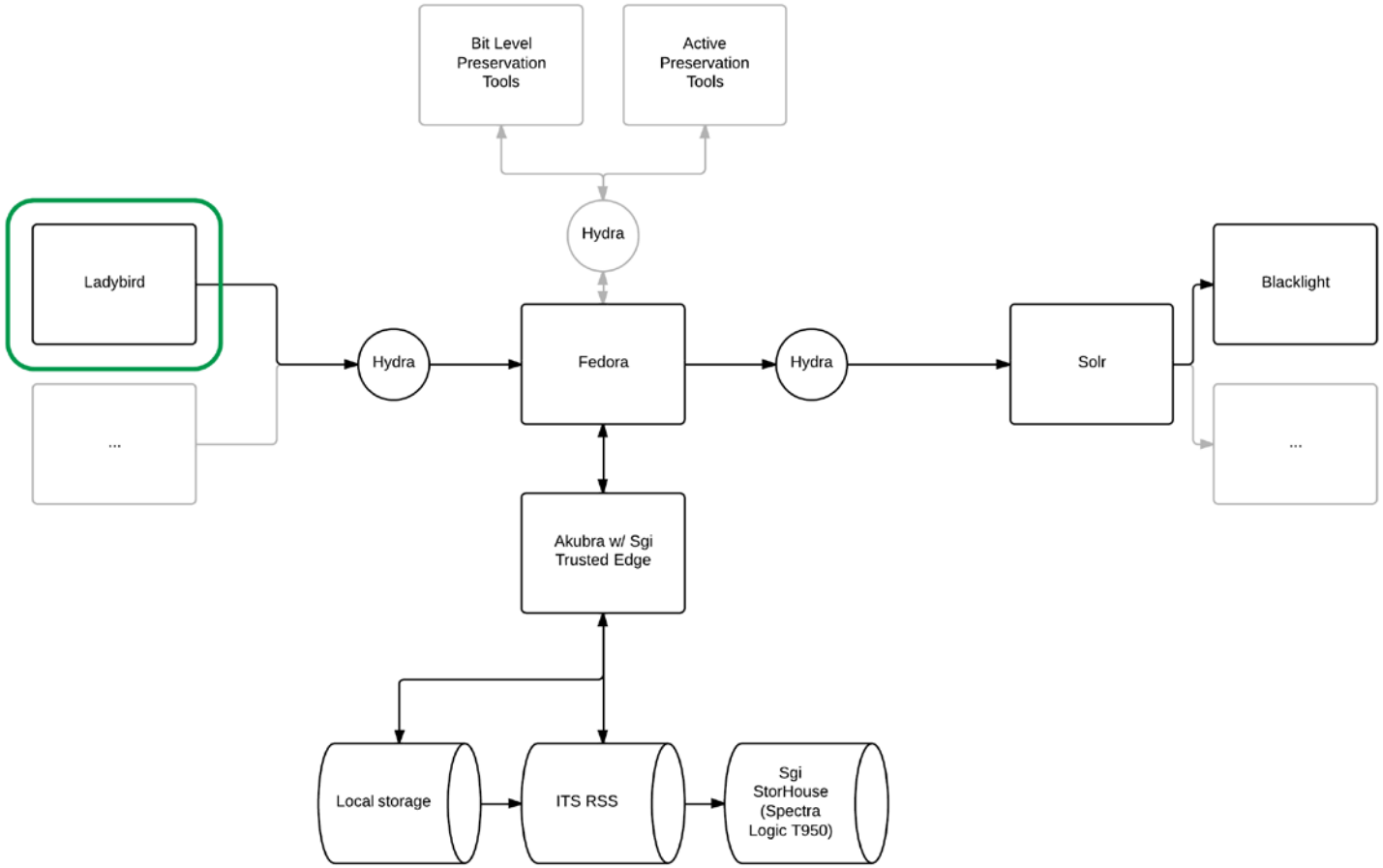
- Duraspace
- Stanford University
- University of Hull
- University of Virginia
- MediaShelf
- University of Notre Dame
- Northwestern University
- Columbia University
- Penn State University
- Indiana University
- London School of Economics
- Rock and Roll Hall of Fame
- Royal Library of Denmark
- Data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

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

- Started June 2010
- Version 1.0 December 2013

- 17 background applications
- 4 desktop applications
- 3 web applications

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

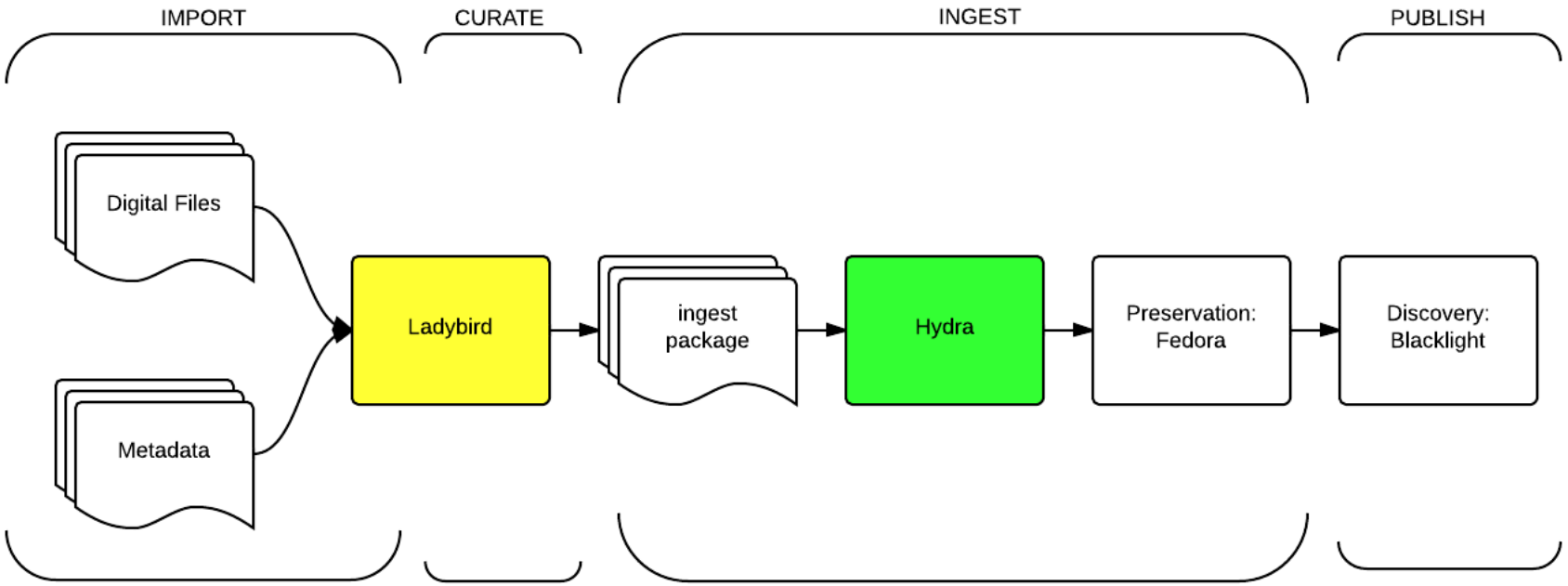
- 2,067,198 assets
- 2.5 mil on deck
- Growth: 1,500 assets per day

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

- A Jazz song by Tadd Dameron

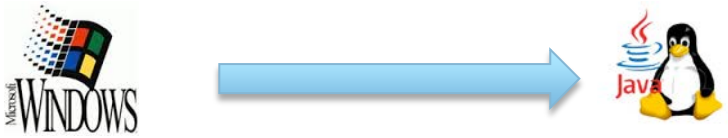
Ladybird with Hydra

Import, Curate, Ingest, Publish



Ladybird Roadmap

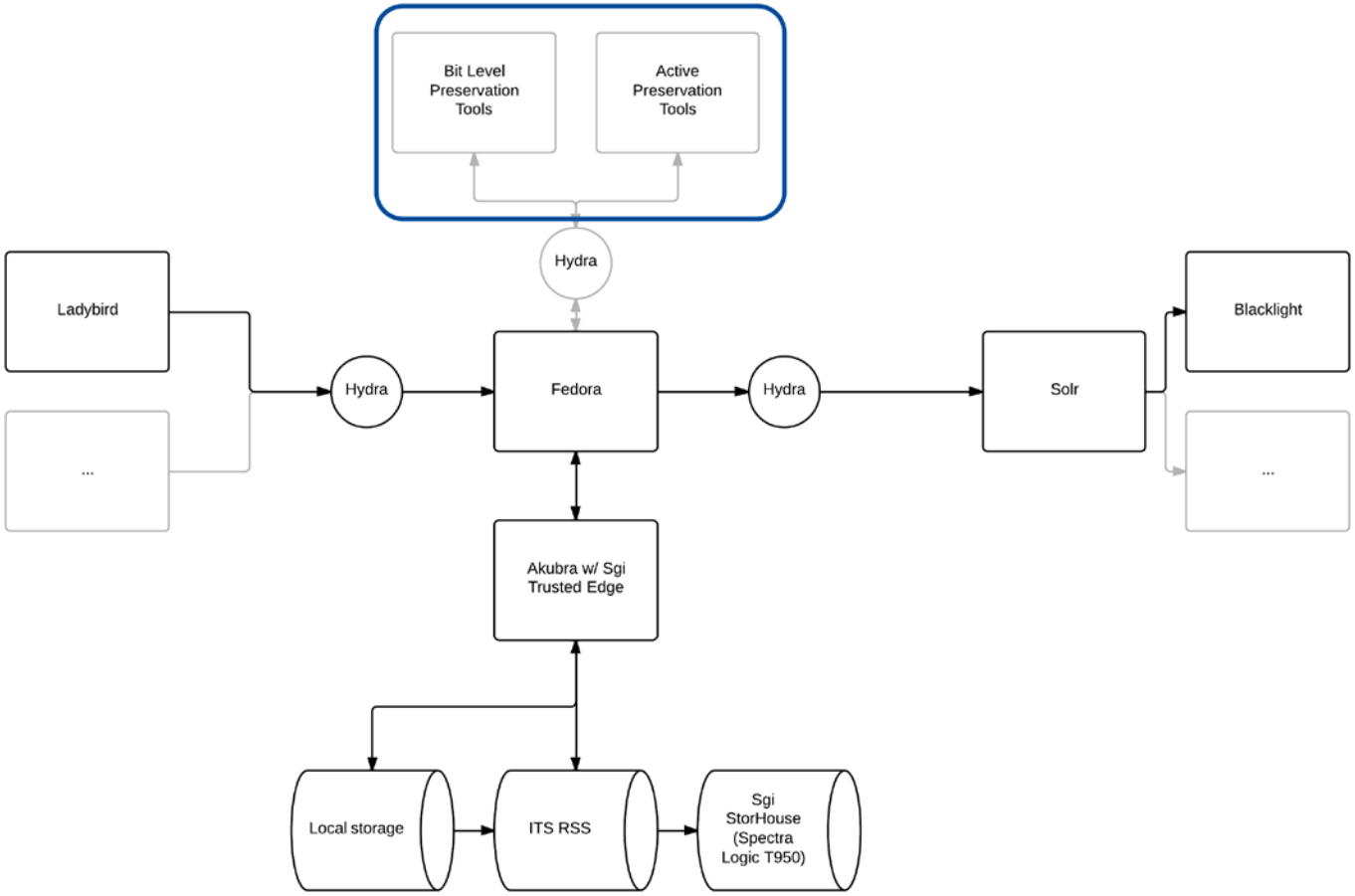
- Partnership with Columbia
- Potential partners with Princeton, MIT, Northwestern
- Release Ladybird as Hydra Head this fall
- Platform migration to Java 8, MySQL



Hydra Roadmap

- Blacklight 5.x
- Fedora 4
- Open Archival Information System (OAIS) ingest model
- Workflow System Architecture
- Digital Preservation Interfaces

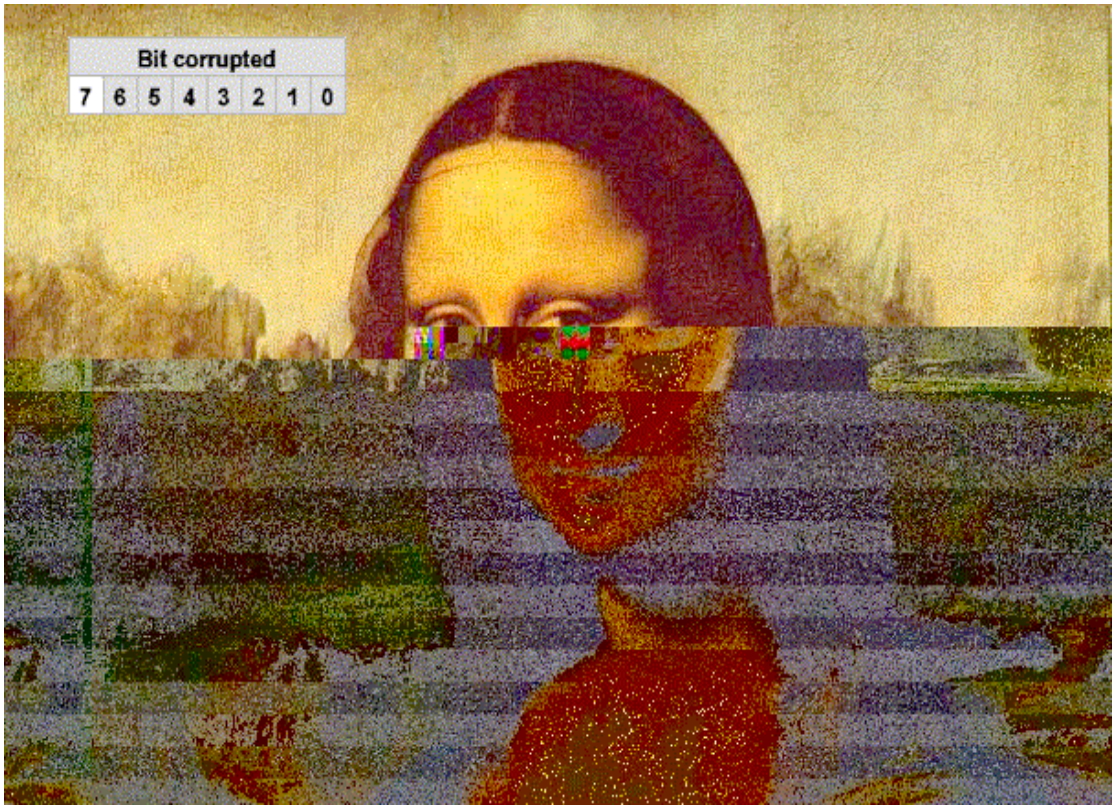
Preservation Tools



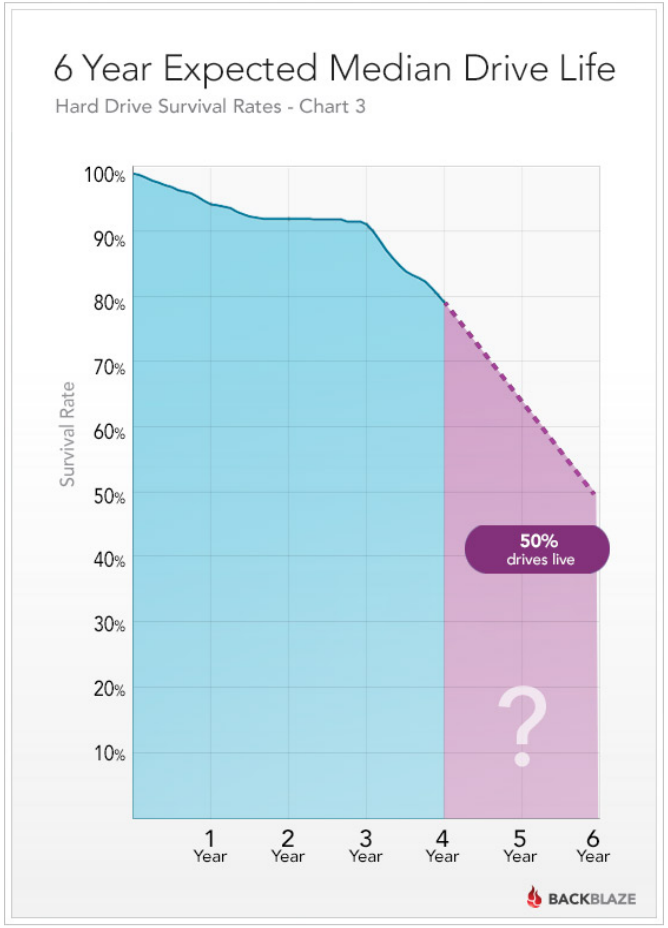
“Digital Information lasts forever or 5 years, whichever comes first”

Jeff Rothenberg. Scientific American, January 1995.

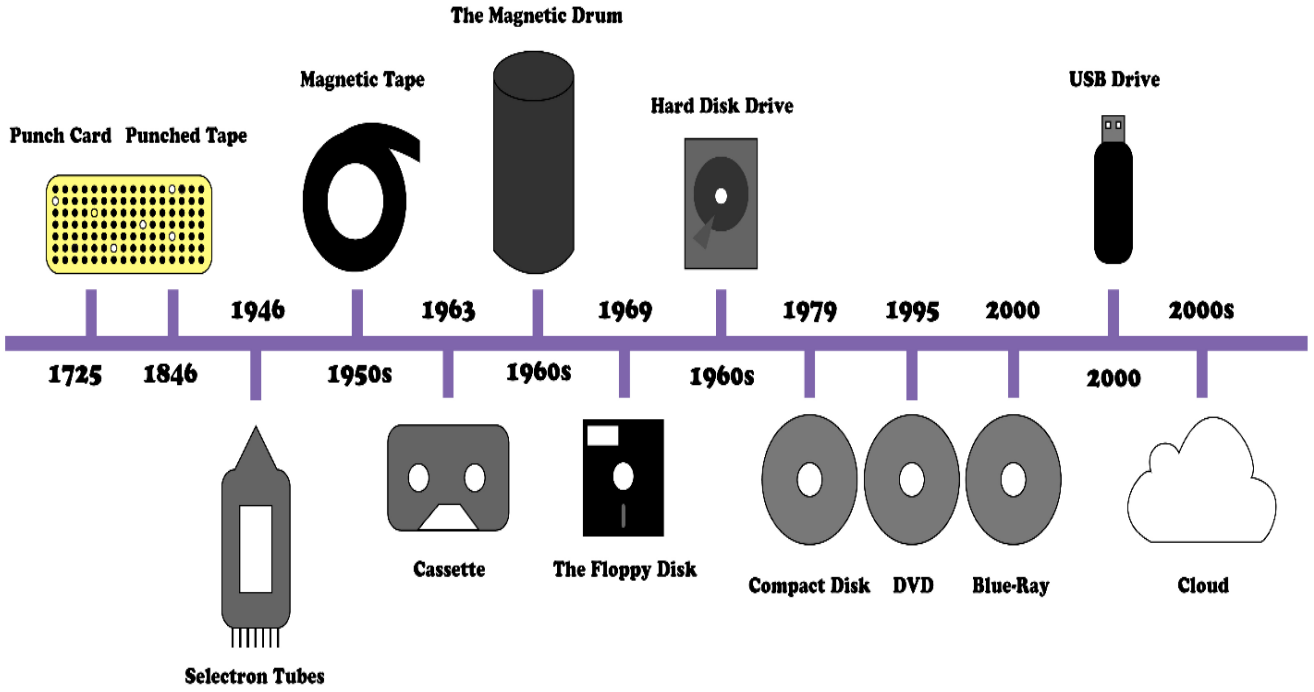
Digital Preservation Challenge: Bit Rot



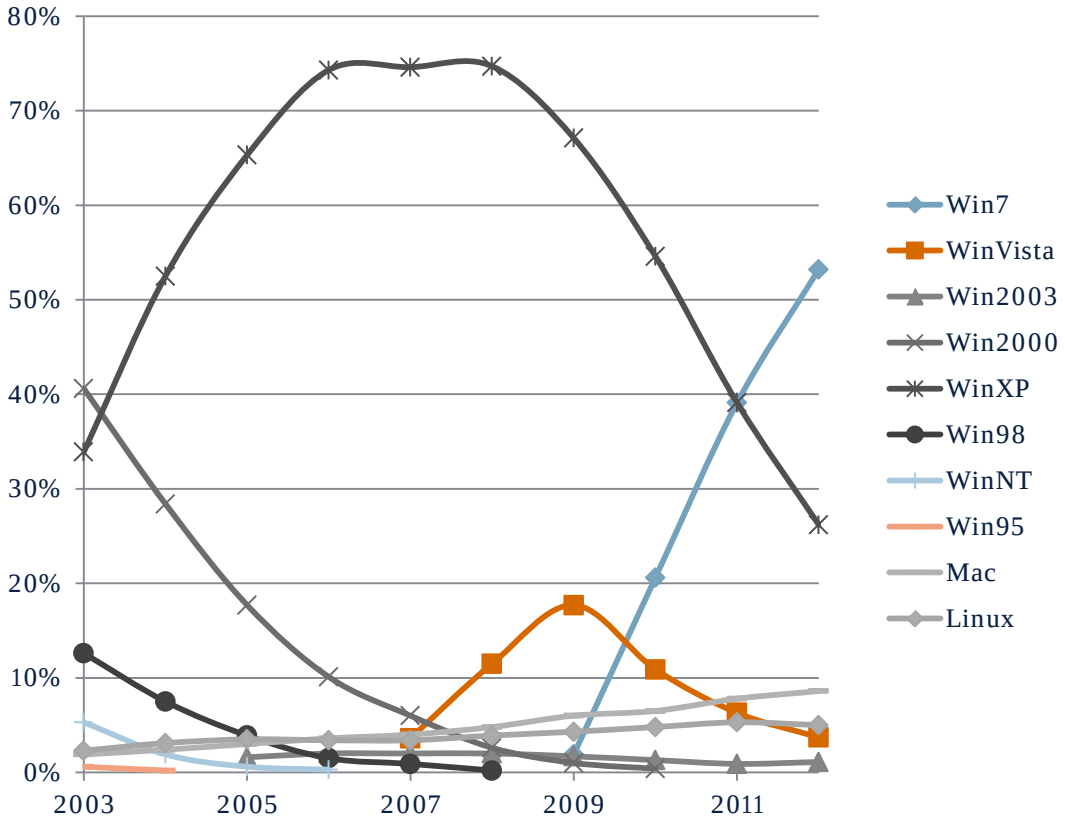
Digital Preservation Challenge: Hardware Failure



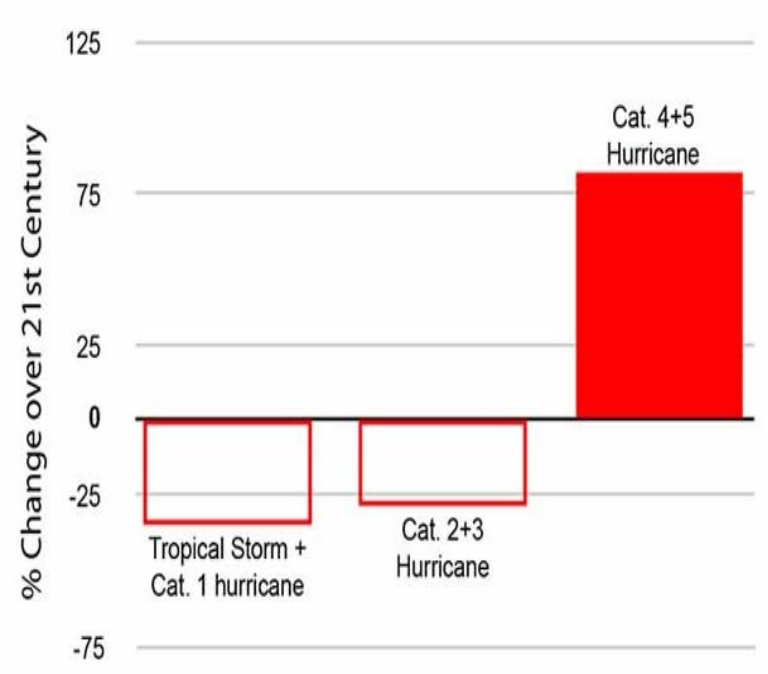
Digital Preservation Challenge: Hardware Obsolescence



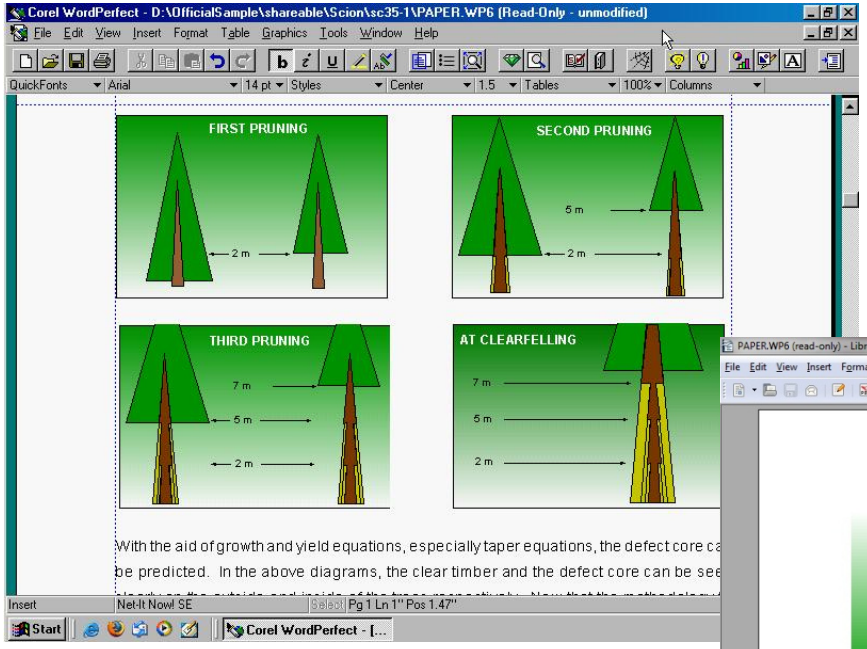
Digital Preservation Challenge: Software Obsolescence



Digital Preservation Challenge: Natural Disasters



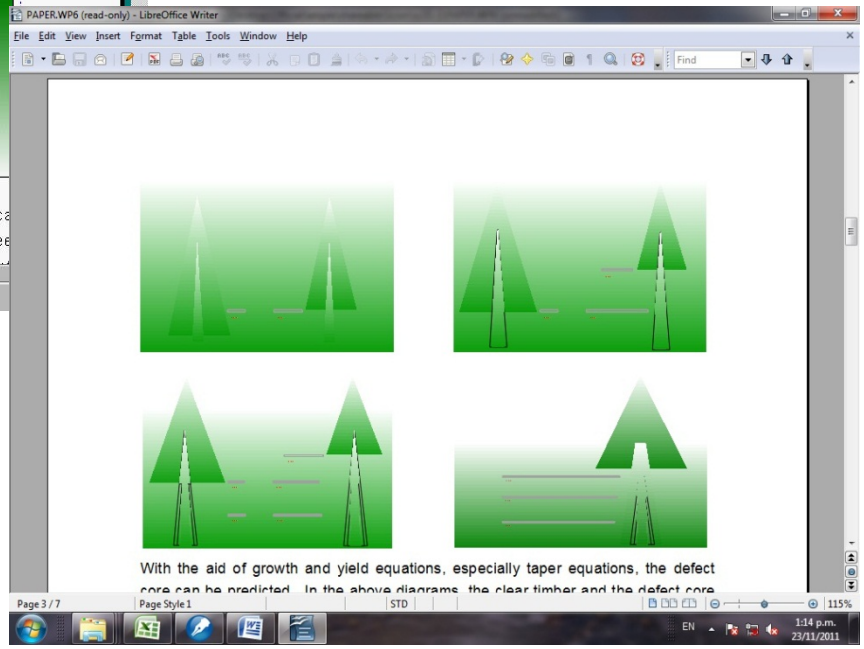
Digital Assets Degrade Without Maintenance



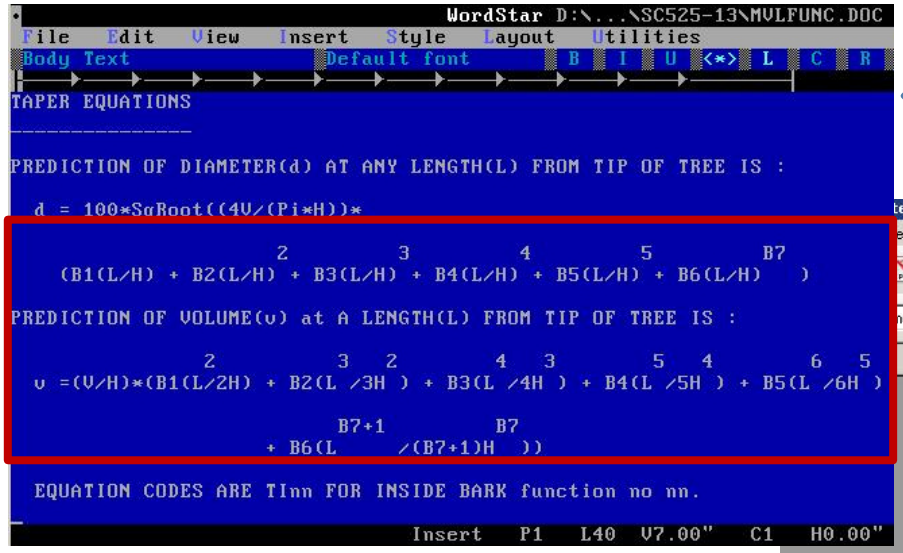
Original digital asset includes visual data

Modern software alters this data:

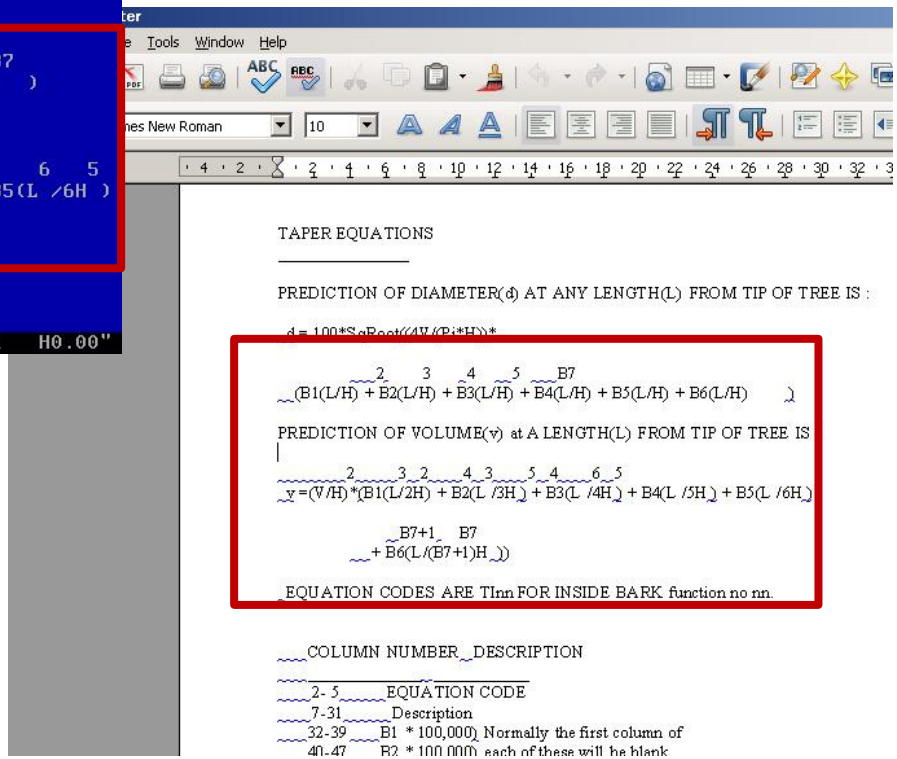
- Changing its meaning
- Reducing the asset's value



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

Addressing the Challenges

		Mitigations						
		Create & Maintain Multiple Copies	Store Copies in Risk-Diversified Locations	Employ Diverse Systems & Vendors	Preserve Original Software	Preserve Original Hardware	Implement Emulation solutions	Migrate Content
Risks	Natural Disasters	✓	✓					
	User Error	✓	✓	✓				
	Media Failure	✓		✓				
	Bit Rot	✓						✓
	Hardware Evolution				✓	✓	✓	✓
	OS Evolution				✓	✓	✓	✓
	Software Evolution				✓	✓	✓	✓

Long Term Digital Preservation Uses Tools and Techniques from Related Disciplines

How is Digital Preservation different to Digital Asset Management?

- Implementing Digital Preservation processes, tools and services enables trust in the integrity and authenticity of digital assets throughout long-term technological change.
- Unlike Digital Asset Management, Long Term Digital Preservation requires a comprehensive understanding of the formats and software dependencies of the objects being preserved and more comprehensive storage and access policies than DAMSs usually require.

How is Digital Preservation Different to Data Migration?

- Data Migration techniques can form a part of more comprehensive digital preservation workflows. Some digital archives transfer content from owners using systems migration processes.

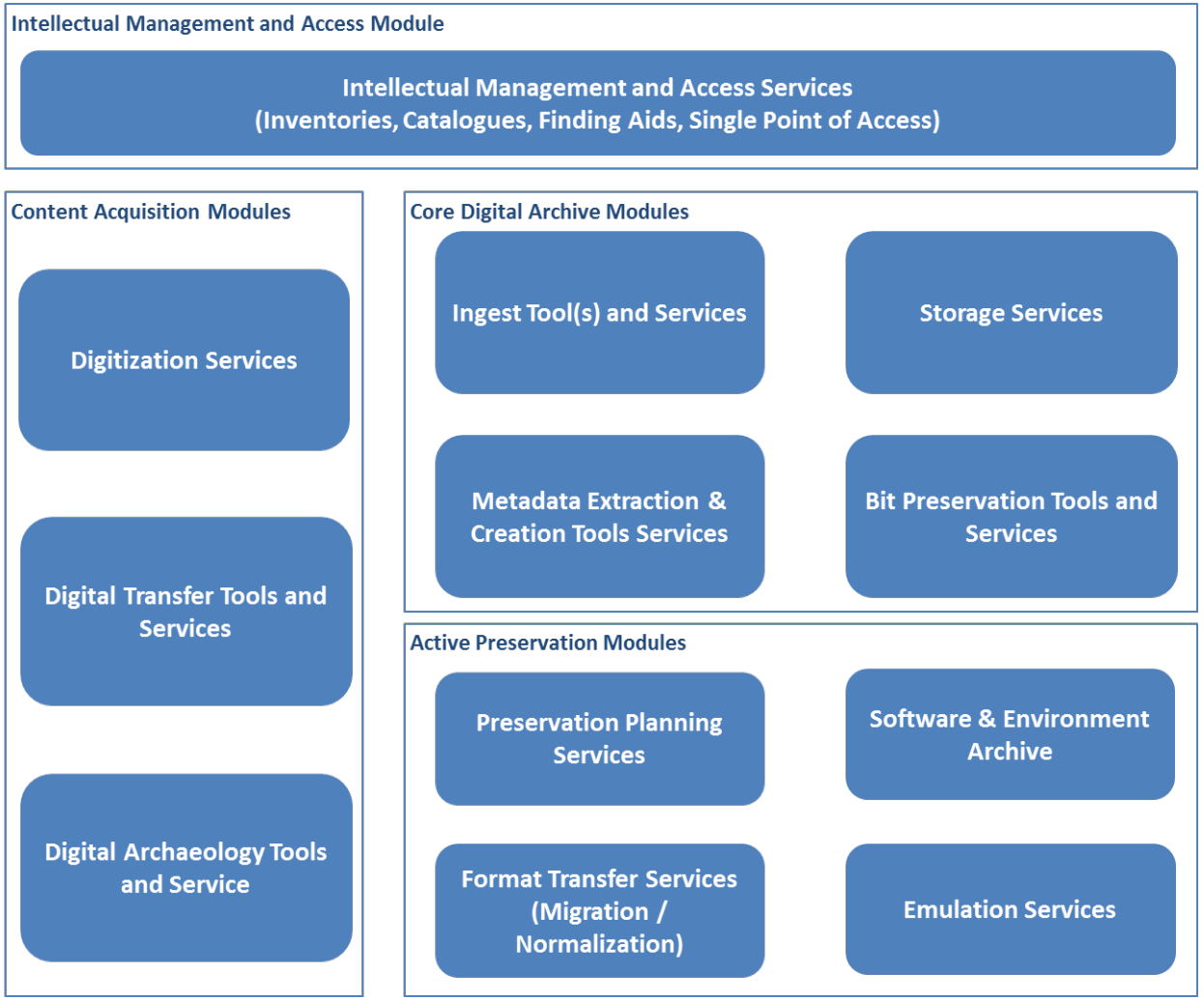
How is digital Preservation different to Disaster Recovery?

- Digital Preservation storage solutions have a unique risk profile and normally require a lower risk of loss of information than standard disaster recovery solutions allow for.
- Digital Preservation storage solutions often have access profiles that can be significantly different to disaster recovery requirements.

How is digital Preservation different to Digital Forensics?

- Digital Preservation uses the tools of digital forensics to achieve different aims, such as using disk imaging tools to snapshot data in place and using forensic analysis tools to identify and characterise the technical properties of digital files.

Digital Preservation Tools & Services



Digital Preservation Tools Roadmap

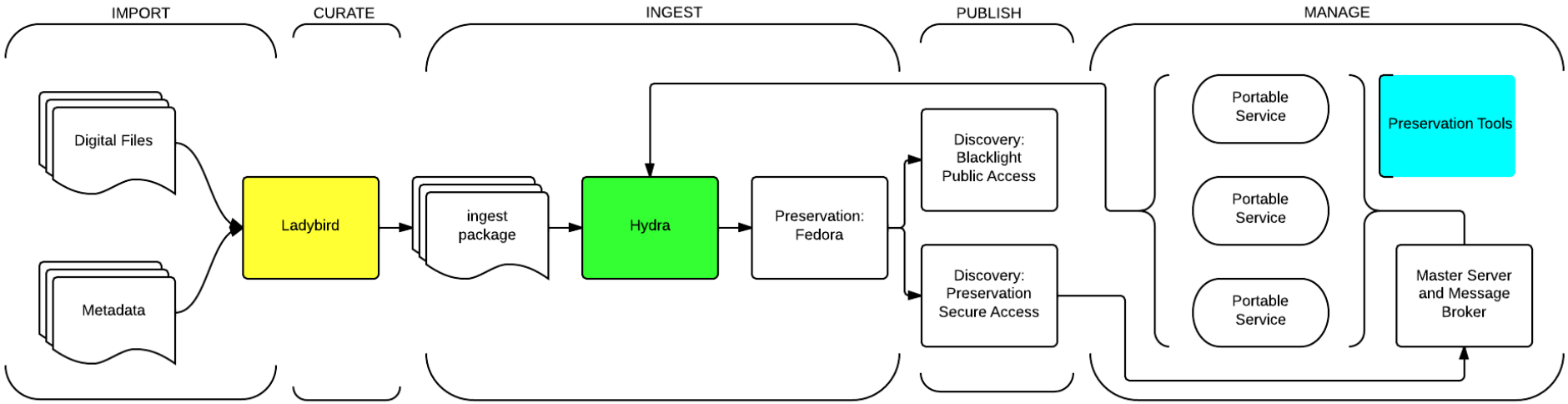
- Programming team formed
- Gathering use cases and user stories
- Platform selection

Simplest use case:

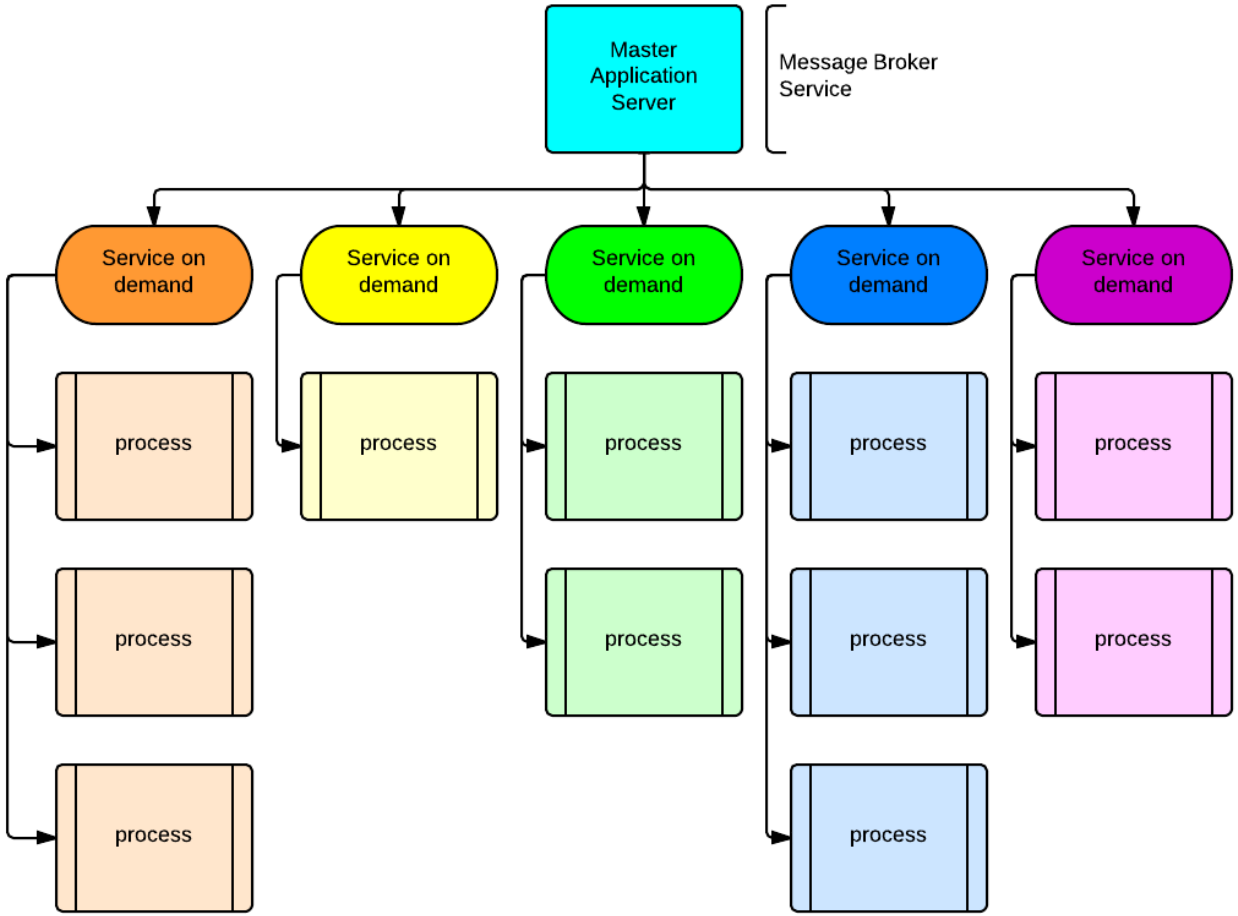
- Validate file: 17 sec average
- Validate current repository: 883 days

- Target: 1 day

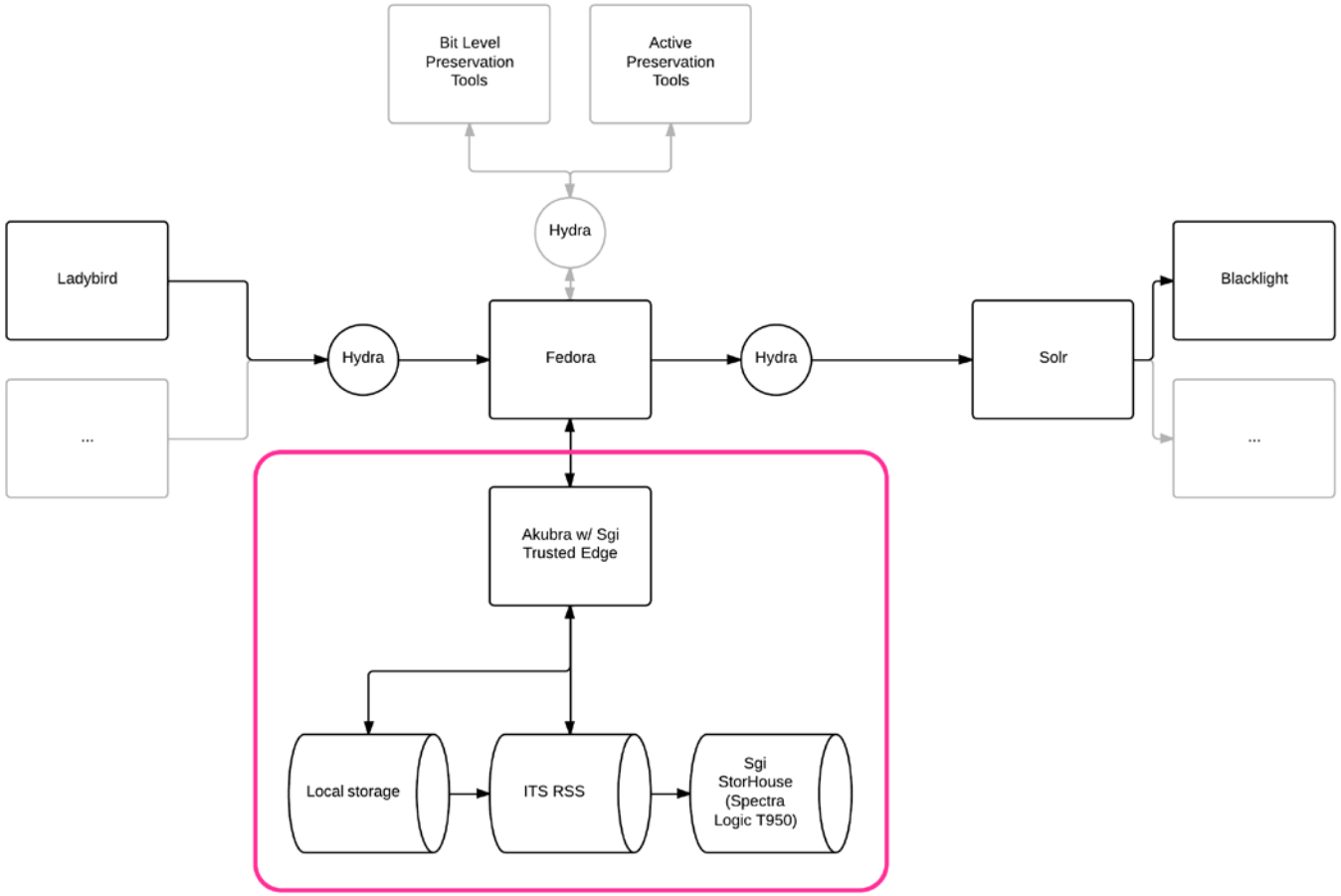
Digital Preservation with Hydra



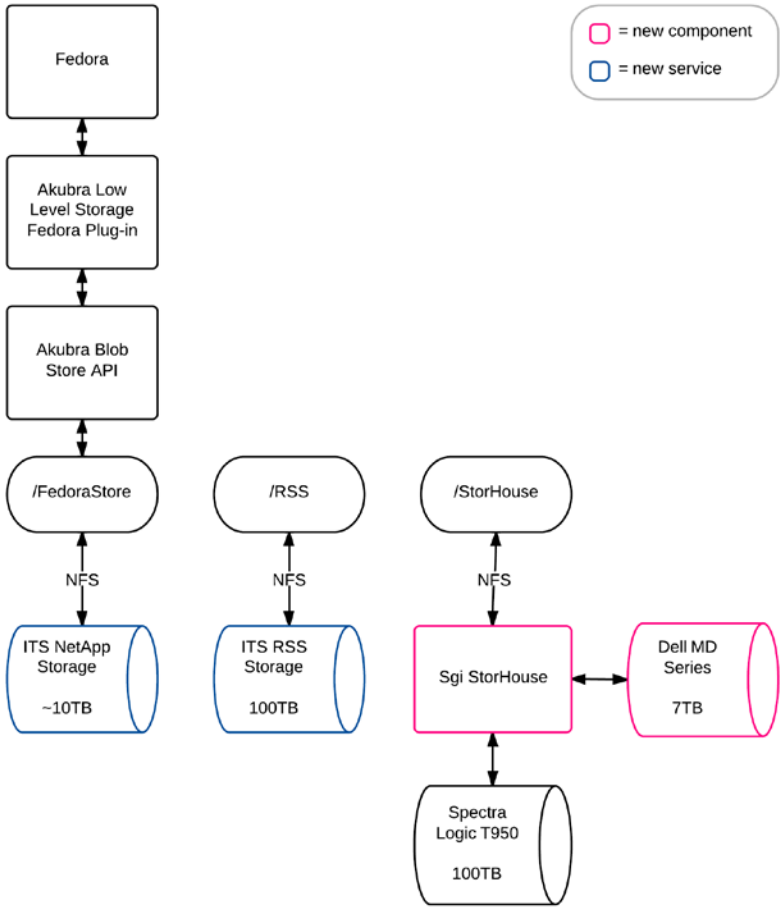
Parallel, Multi-Threaded Applications Spanned Across Virtual Servers



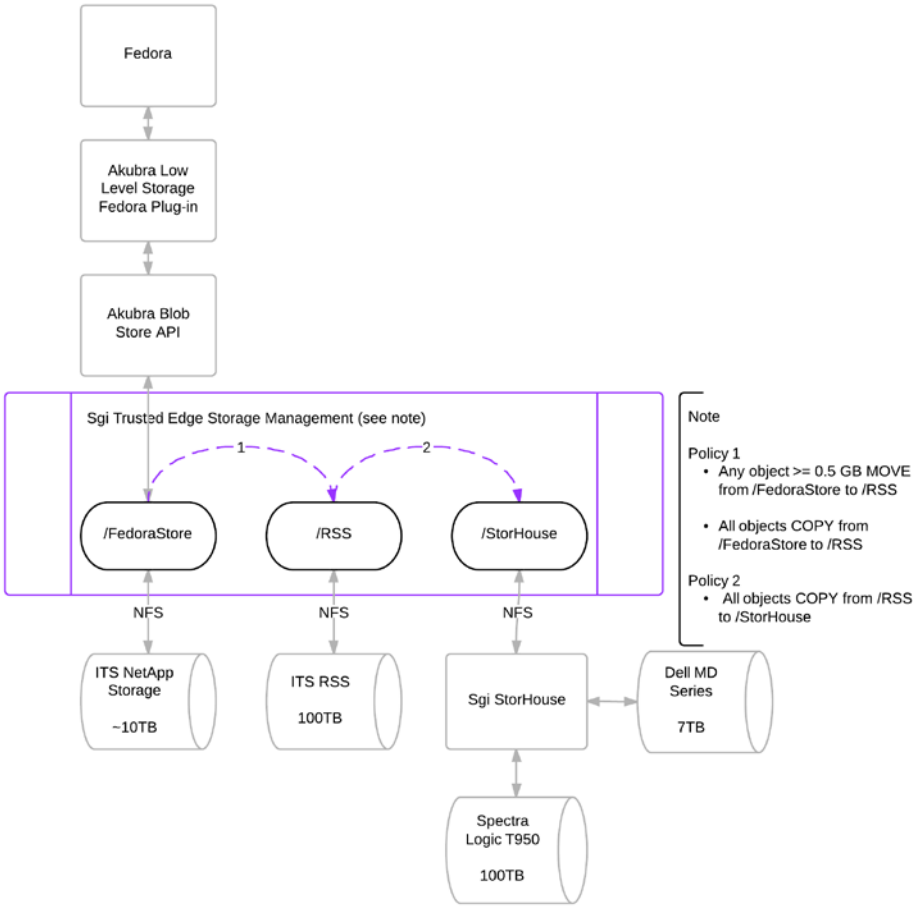
Storage Infrastructure



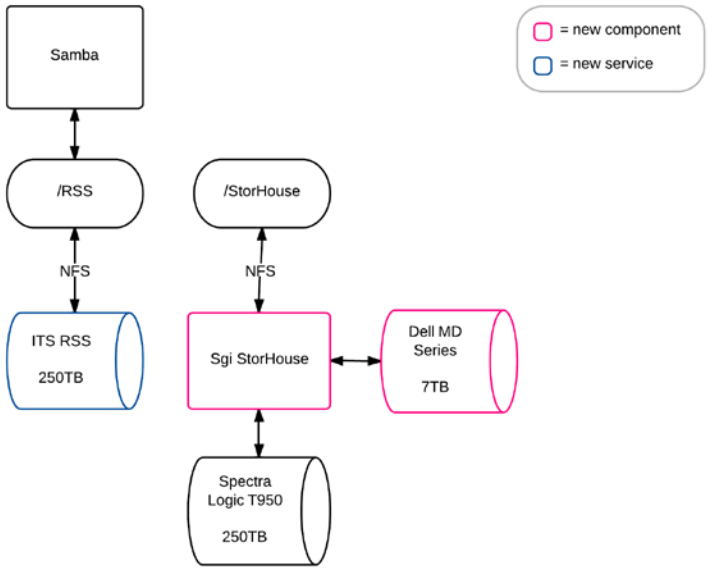
Proposed FY2015



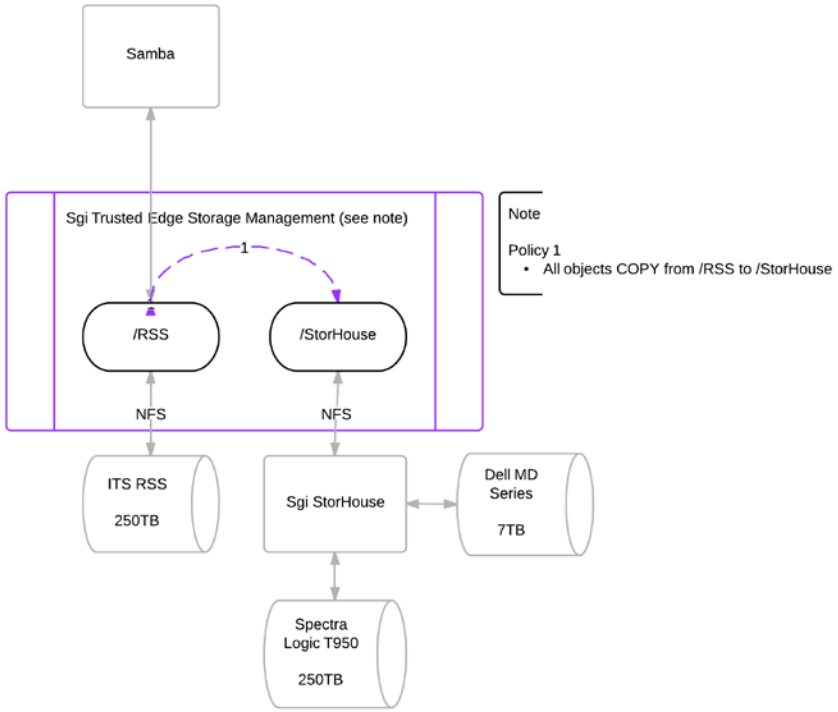
Proposed Trusted Edge Policy



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**