Upgrading? Migrating? There's a portmanteau for that!

David Wilcox, DuraSpace
Adam Wead, Penn State University
Nick Ruest, York University
Michael Friscia, Yale University

Pilot Project Motivations

Model Fedora 3 data in Fedora 4

Upgrade/migrate Hydra, Islandora, and custom implementations

Provide testing and feedback on migrations tools

Serve as examples for the broader community

Penn State: Migrating to Fedora 4

Adam Wead

PENNSTATE

Fedora 4 at Penn State

ScholarSphere

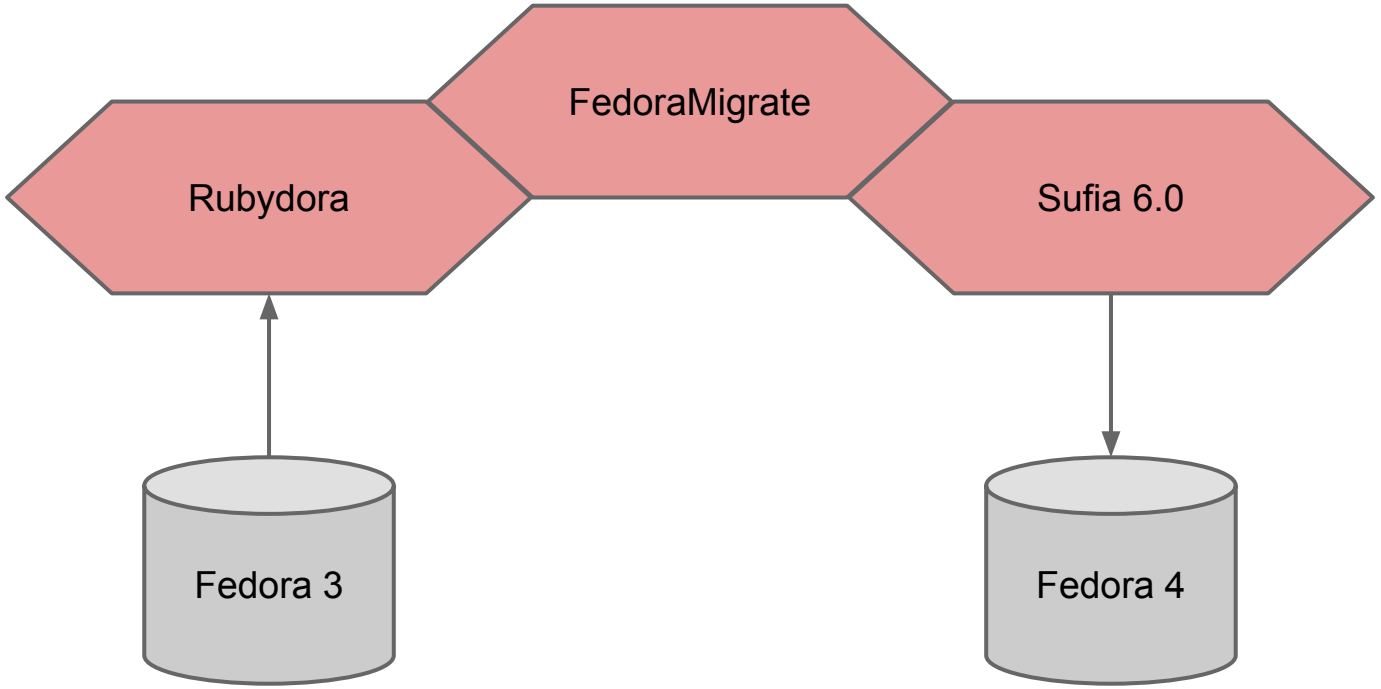
- Institutional repository built using the Sufia gem
- 4775 objects / 37GB data
- migrated to Fedora 4 in April

ArchiveSphere

- digital repository to support the University Archives
- 72262 objects / 186GB data
- migration scheduled for later this year

PENNSTATE

Migration Overview



1. Iterate over repository and migrate objects
2. Migrate relationships for each migrated object
3. Verify and test

PENNSTATE

Object Migration

- Content models are defined in Ruby
- Each object was migrated as a Ruby object
- Datastreams moved and verified
- Versions recreated as needed
- Used premis:hasDateCreatedByApplication
- Translated our existing RDF metadata into properties asserted on the F4 resource
- Rights metadata translated to W3C web ACLs
- Audit information from Fedora3 **not** migrated

PENNSTATE

Relationships & Reporting

- RELS-EXT translated to an RDF graph for each migrated object
- Progress information saved in JSON files, one per object
- Parsed JSON files for verification results
- Corroborated with independent reporting processes

PENNSTATE

Problems

- Pre-migration relationship cleanup
- Migration script required “restarts”
- Unparseable RDF
  - re-encode ISO-8859 to UTF-8
  - one “manual” field migration
  - re-edit fields with LaTeX code and content from PDF copy/paste actions
- Missed the *label* property (Doh!)

PENNSTATE

References

fedora-migrate gem

https://github.com/projecthydra-labs/fedora-migrate

Migration audits

https://github.com/psu-stewardship/scholarsphere/commit/e99bee828f4baa340019a099dba0ec6b3a069361

Penn State's process

This includes the pre-migration and post-migration audit steps

https://github.com/psu-stewardship/scholarsphere/wiki/Fedora-3-to-Fedora-4-Migration

Migration documentation for Sufia

https://github.com/projecthydra/sufia/wiki/Migrating-to-Fedora-4-with-fedora-migrate

York University -- Islandora

- Background
- Mappings
  - fcrepo object properties
  - fcrepo datastream properties
  - rels-ext
  - auditTrail
  - PCDM
- Approach

Upgration

Background

York University

Solution Packs:
Collection, Audio, Book, Compound, Large Image, Video, Web Archive

Upgration Pilot - York University

Created by David Wilcox, last modified on Feb 11, 2015

Project Overview

The York University Libraries upgration project identifies collections that cover the range of object models that the repository uses. The conservative goal is to perform an upgration on the collections listed below. The stretch goal is an upgration all of all objects in the repository.

By upgration, we mean upgrating and migrating objects and datastreams, along with security restrictions (XACML), in Fedora 3.8.0 to Fedora 4.x. Moreover, we will develop a strategy for upgrading and migrating our content models, including inline XML datastreams, managed datastreams, and external datastreams.

York University Digital Library (YUDL) is an Islandora repository that run on the HEAD version of all Islandora Foundation modules. The repository is run as close a stock/generic Islandora instance where possible. Therefore, this upgration pilot can serve as a basis for a generic Islandora Fedora 3.x to Fedora 4.x upgration.

- Collection Description(s)
- Object Models
- Fedora 3 Details
  - Storage: Legacy storage (or Akubra)
  - XML metadata : datastreams
  - XML metadata : inline
  - Content models
  - Datastream types (inline, managed, redirect, and external)
  - Identifiers
  - Indexing strategies (GSearch, RI-Search vs. F4 approaches)
  - Replication/Journaling
  - Security policies: XACML
  - OAI-PMH
  - Versions
  - Disseminators
  - Audit history
- Fedora 4 Details

Collection Description(s)

York University Digital Library contains approximately 200,000 unique digital assets.

Jean Augustine fonds

Upgration

Property mappings

fcrepo3->fcrepo4

Object properties

fcrepo3 Object properties to fcrepo4

fcrepo 3	fcrepo4	Example
PID	dcterms:identifier	yul:328697
state	fedoraaccess:objState	Active
label	fedora3model:label†	Elvis Presley
createDate	premis:hasDateCreatedByApplication	2015-03-16T20:11:06.683Z
lastModifiedDate	metadataModification	2015-03-16T20:11:06.683Z
ownerId	fedora3model:ownerId†	nruest

† The `fedora3model` namespace is not a published namespace. It is a representation of the fcrepo3 namespace `info:fedora/fedora-system:def/model`.

‡ Not yet implemented

fcrepo3->fcrepo4

Datastream properties

fcrepo3 Datastream properties to fcrepo4

fcrepo3	fcrepo4	Example
DSID	dcterms:identifier	OBJ
Label	dcterms:title‡	ASC19109.tif
MIME Type	ebucore:hasMimeType†	image/tiff
State	fedoraaccess:objState	Active
Created	premis:hasDateCreatedByApplication	2015-03-16T20:11:06.683Z
Versionable	fedora:hasVersions‡	true
Format URI	premis:formatDesignation‡	info:pronom/fmt/156
Alternate IDs	dcterms:identifier‡	
Access URL	dcterms:identifier‡	
Checksum	cryptofunc:*hashalgorithm*‡	cryptofunc:sha1 "c91342b705b15cb4f6ac5362cc6a47d9425aec86"

† The `fedora3model` namespace is not a published namespace. It is a representation of the fcrepo3 namespace `info:fedora/fedora-system:def/model`.

‡ Not yet implemented

fcrepo3->fcrepo4

RELS-EXT/RELS-INT

Fedora 3.x namespace RELS-EXT predicates

```
$ grep -R "FEDORA_RELS_EXT_URI" Islandora-7.x
```

Islandora

- isMemberOfCollection
- isMemberOf

Image Annotation

- isAnnotationOf

Compound

- isConstituentOf
- isPartOf

Islandora namespace RELS-EXT predicates

```
$ grep -R "ISLANDORA_RELS_EXT_URI" Islandora-7.x
```

Book

- isPageOf
- isSequenceNumber
- isPageNumber
- isSection

Image Annotation

- targetedBy
- targets
- hasColor
- hasURN
- strokeWidth
- isEntity
- isAnnotationType

OCR

Islandora Ontology

http://islandora.ca/ontology/relsext/#
http://islandora.ca/ontology/relsint/#

Islandora Ontology

https://github.com/Islandora-Labs/islandora_ontology

fcrepo3->fcrepo4

auditTrail

Audit log migration

auditTrail mapping

fcrepo3 event	fcrepo4 Event Type
addDatastream	premis:ing‡
modifyDatastreamByReference	audit:contentModification/metadataModification‡
modifyObject	audit:resourceModification‡
modifyObject (checksum validation)	premis:validation‡
modifyDatastreamByValue	audit:contentModification/metadataModification‡
purgeDatastream	audit:contentRemoval‡

† The `fedora3model` namespace is not a published namespace. It is a representation of the fcrepo3 namespace `info:fedora/fedora-system:def/model`.

‡ Not yet implemented

migration-utils

https://github.com/fcrepo4-labs/migration-utils

Yale upgrade to Fedora 4

- Three Fedora 3 instances
  - 2 with custom front ends/FG Search
    - Roughly 12 million objects
    - Access copies only < 5TB
    - Migration to Hydra
  - 1 Hydra based
    - Roughly 2 million objects
    - Access and Masters > 120TB
    - Continues to grow

Bulk Re-Ingest & Upgration

- Fedora 4 shadow system
  - Running in parallel with production Hydra
  - Runs on different servers
  - Connects to same storage as Fedora 3
- 3 months to ingest 150TB
  - Bulk of the setup is in Metadata decisions/RDF
  - Content *should* move to PCDM easily

Auditing

- Manual audits using Tableau
  - Data from:
    - Fedora 3 & 4
    - Hydra/SOLR
- Local bulk ingest application (Ladybird)
  - Visual representations of failed audits

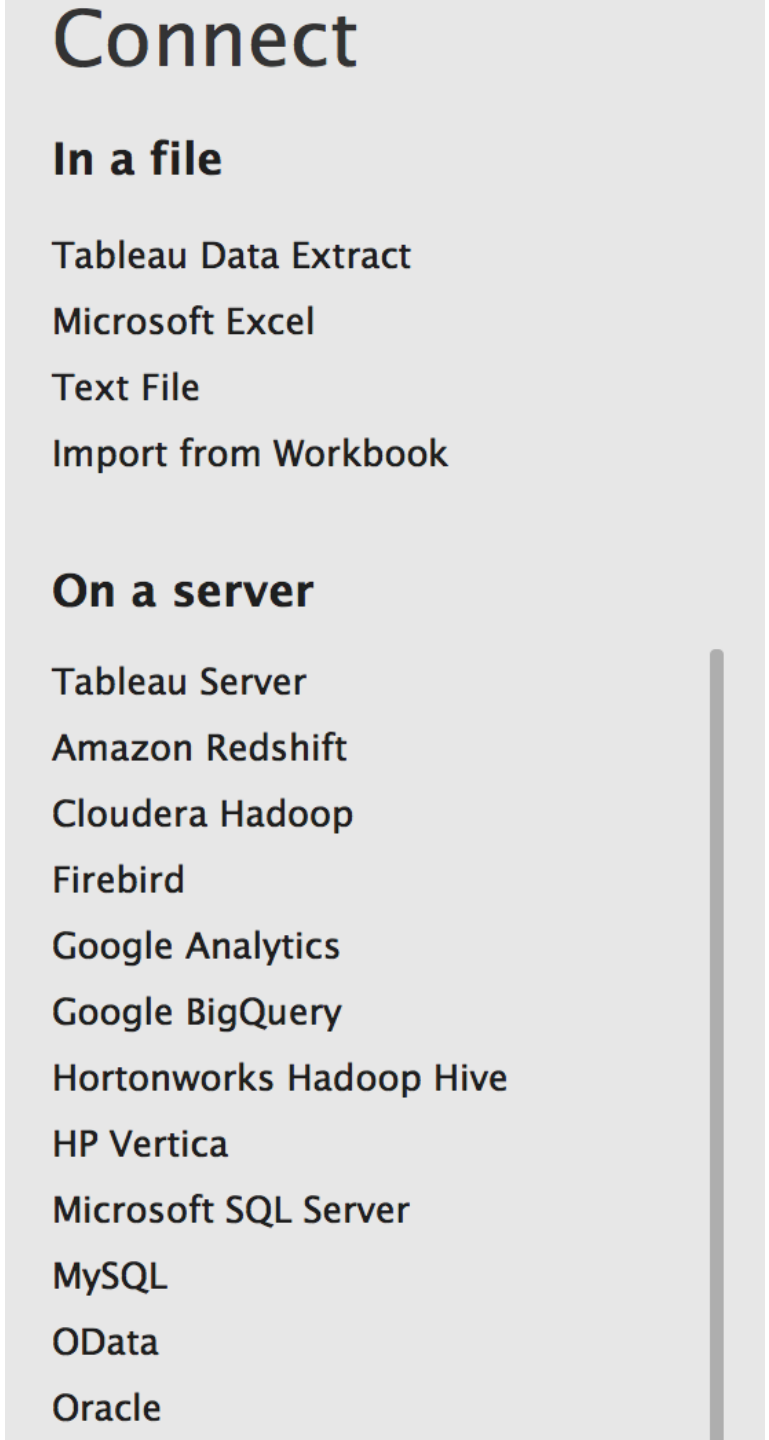
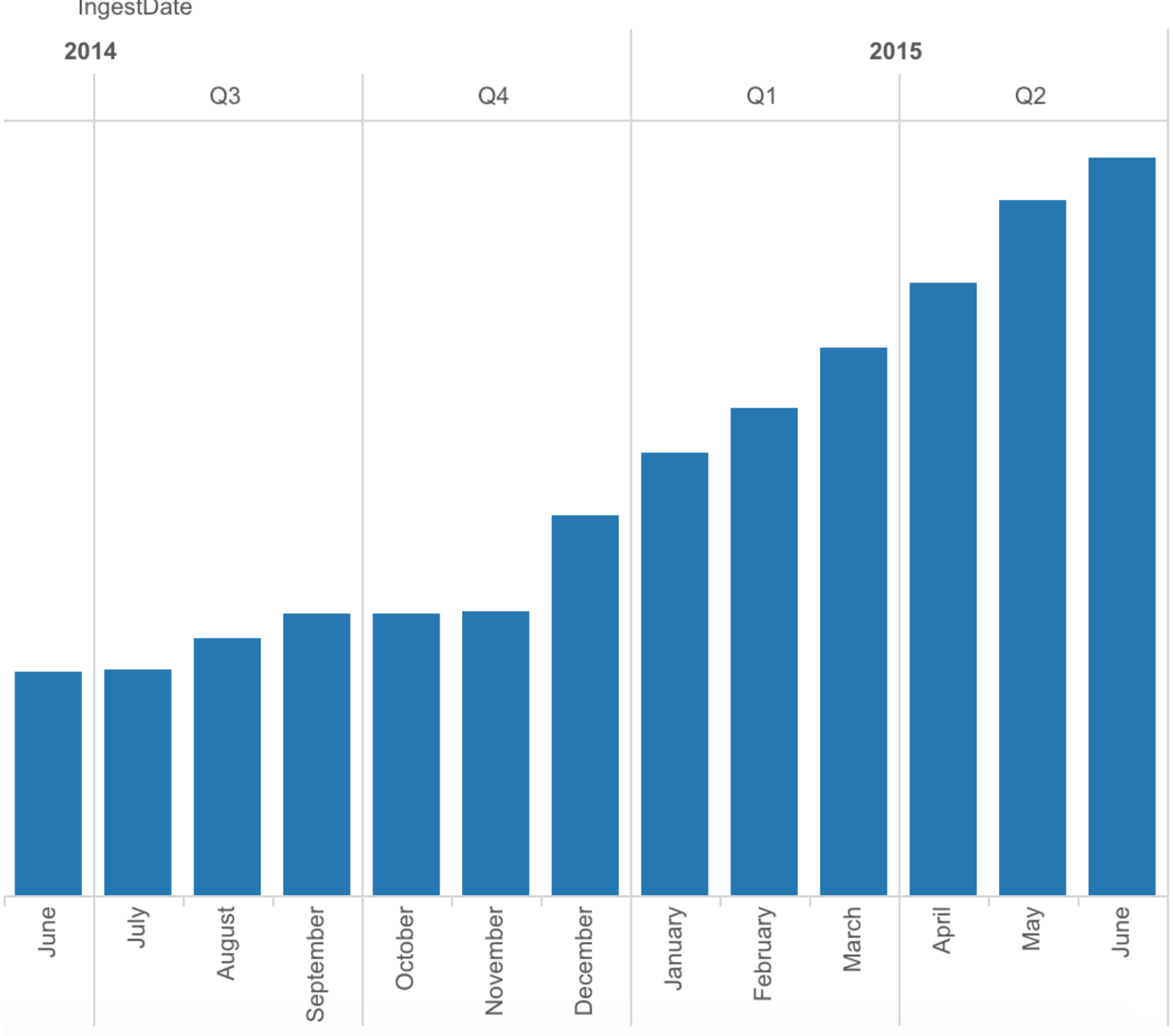


Tableau Datasources

- Flexible
- online and offline data
- SOLR on roadmap


IngestDate
2014
2015
Q3
Q4
Q1
Q2
June
July
August
September
October
November
December
January
February
March
April
May
June

Hydra-Ladybird Statistics

Back Workbook

Share Remember my changes Edit

Ingest Times | Hydra Growth | Hydra Growth Detail | Ladybird Growth | Ladybird Growth Detail | Yulhy Ingest Stats

IngestDate

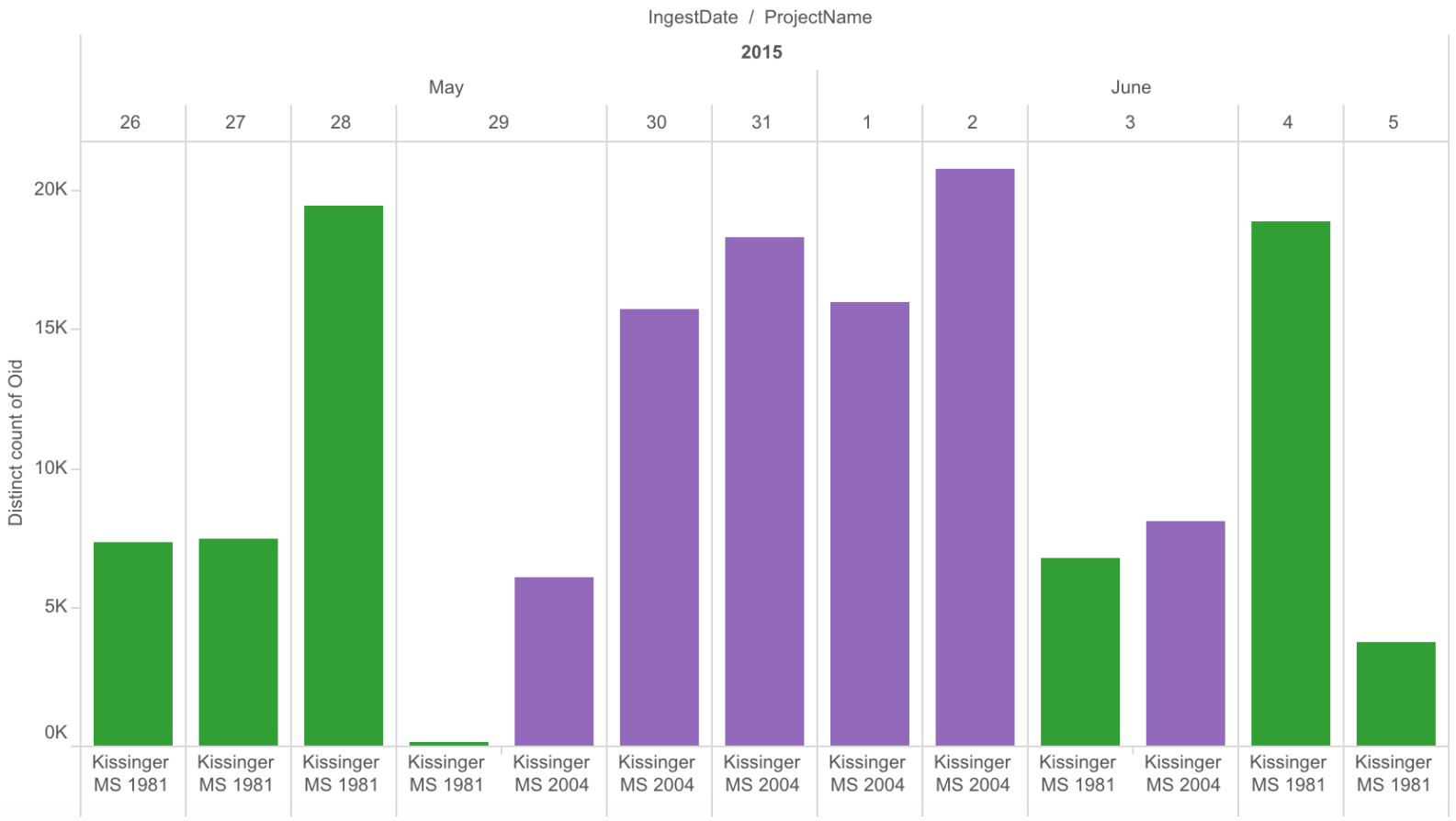
5/26/2015 12:(6/5/2015 8:37:

ProjectName

- [ ] (All)
- [ ] Day Missions Collection
- [x] Drama
- [x] Faber Birren
- [x] International Mission Phot…
- [x] Kissinger MS 1981
- [x] Kissinger MS 2004
- [x] LWL general collection
- [x] Map Department Digital C…
- [x] Maurice Durand Collection
- [x] Persian Medical Texts
- [x] Persian Philological Texts
- [x] Sack Furniture Archive
- [x] Yale Indian Papers Project

ProjectName

- Kissinger MS 1981
- Kissinger MS 2004



Hydra Issues

Back Workbook

Share Remember my changes Edit

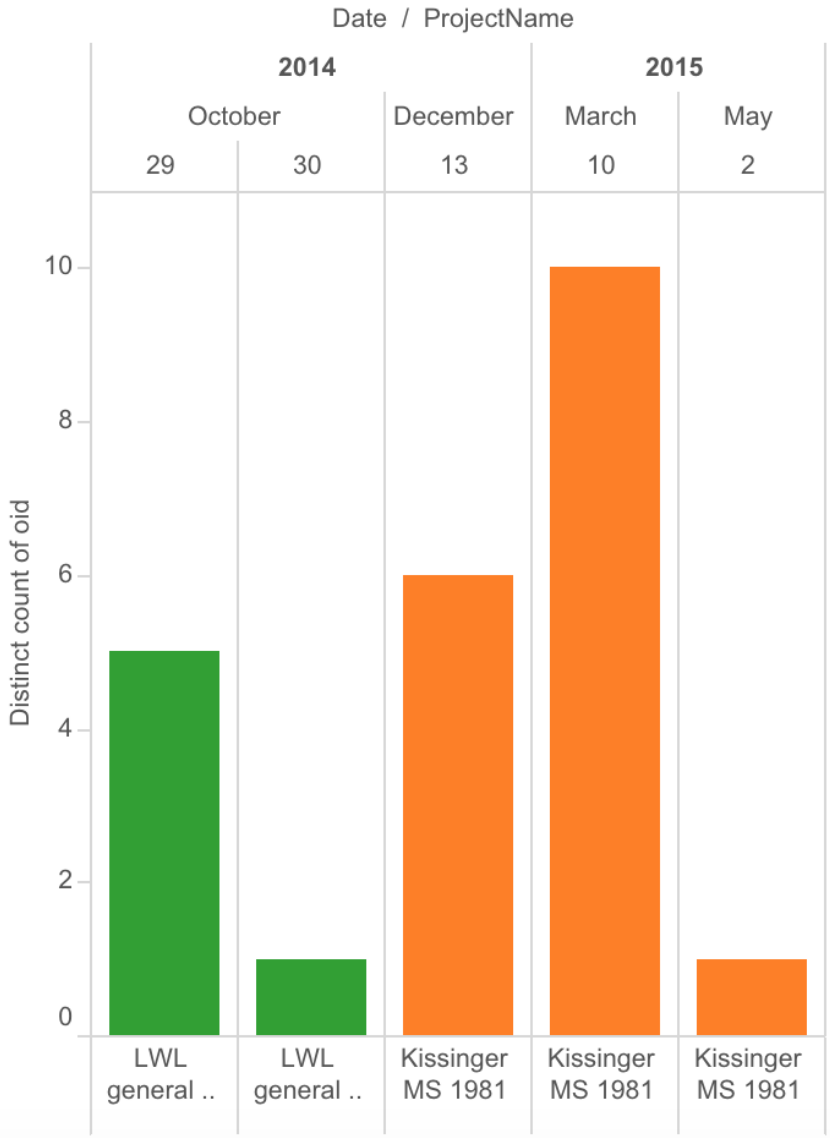
Possible Ingest Fails | Active Parent, deleted Child | State Mismatch parent/child | OID Missing in Hydra | Hy

ProjectName

- [ ] (All)
- [ ] Day Missions Collection
- [x] Kissinger MS 1981
- [x] LWL general collection

ProjectName

- Kissinger MS 1981
- LWL general collection



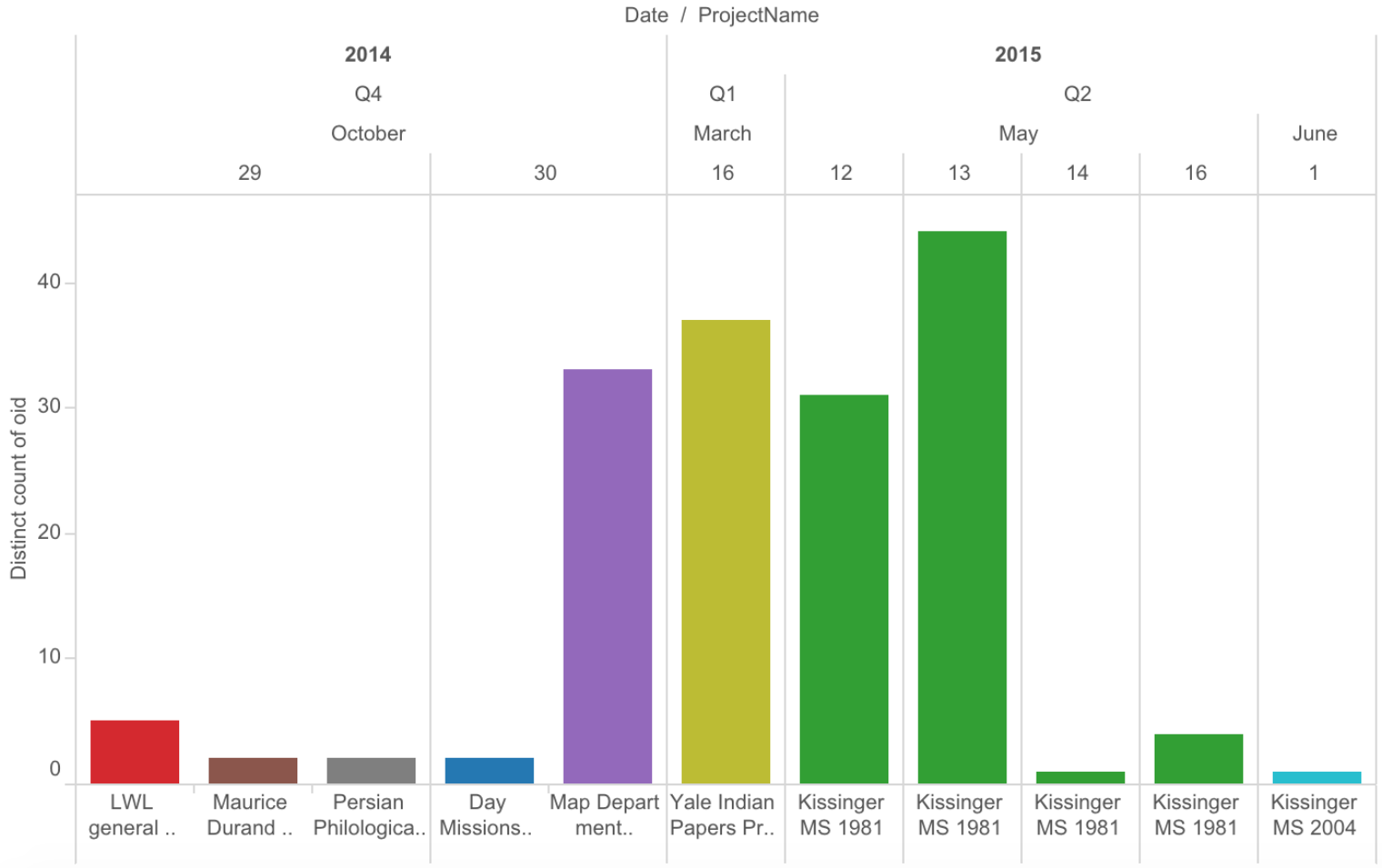
ProjectName

- [x] (All)
- [x] Day Missions Collection
- [x] Kissinger MS 1981
- [x] Kissinger MS 2004
- [x] LWL general collection
- [x] Map Department Digital C…
- [x] Maurice Durand Collection
- [x] Persian Philological Texts
- [x] Yale Indian Papers Project

ProjectName

- Day Missions Collection
- Kissinger MS 1981
- Kissinger MS 2004
- LWL general collection
- Map Department Digital …
- Maurice Durand Collection
- Persian Philological Texts
- Yale Indian Papers Project

Objects who failed a datastream audit which indicates a file could be missing in Hydra.



Completion

- Goal:
  - All assets moved and audited by March 2016
  - All Fedora 3 instances shut down Summer 2016