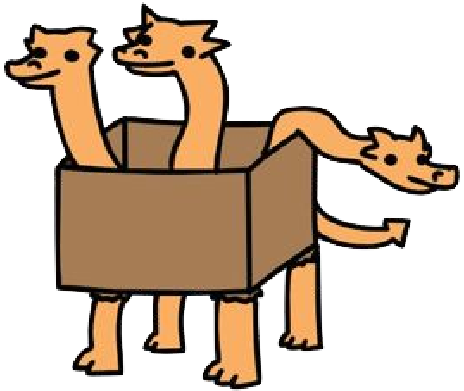
# Hydra in a Box:

## Building A Next-Generation Platform for Digital Collections

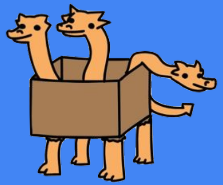
Hannah Frost, Stanford University
Gretchen Gueguen, DPLA
Mark A. Matienzo, DPLA

DPLAFest 2016 — April 14, 2016



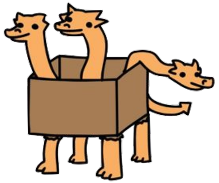
# Project Overview

- A Time for Change
- The Vision
- Project Partners
- Project Goals
- Timeline



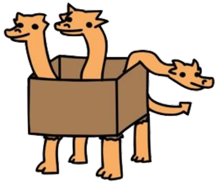
# A Time for Change

- Conversations between Stanford University, DPLA, and DuraSpace informed project design
- Current digital collections platforms originate in an earlier phase of the web, which explain current limitations
- Infrastructure needs in the DPLA Hub network
  - Legacy systems unable to leverage modern affordances of the web
  - Lack of scalable and sustainable aggregation workflows
  - Lack of support for linked data and metadata enrichment
  - Perceived lack of “obvious choices” for replacement systems for digital collections



# The Vision

- A product and service that is easy to use, easy to integrate, and that
- Reduce barriers (including cost) to DPLA contribution
- Allow digital collections to be not just ***on*** the web, but ***of*** the web
- Expand and diversify both the DPLA and Hydra communities



# Project Partners

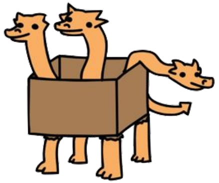
DPLA

LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

DURA SPACE™

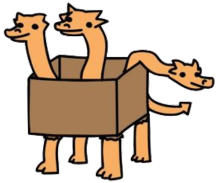
INSTITUTE of Museum and Library SERVICES

hydra



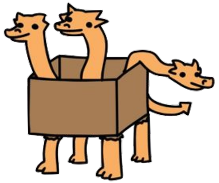
# Project Goals

- Development of turnkey, Hydra-based application that leverages and improves on core components
- Development/integration of metadata aggregation & enrichment tools
- Connect components with DPLA hubs, current Hydra partners, and prospective Hydra adopters
- Work toward a hosted service



# Timeline

- May 2015-November 2017 (30 months)
- Design process: May 2015-March 2016
- Development: March 2016-November 2017
- Service development and community engagement: throughout project



# Design Phase

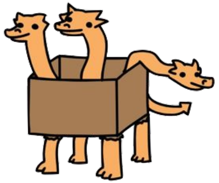
Discovery Phase (Fall 2015)

- Literature review and product/service analysis
- Surveys, interviews, and focus groups
- Community outreach

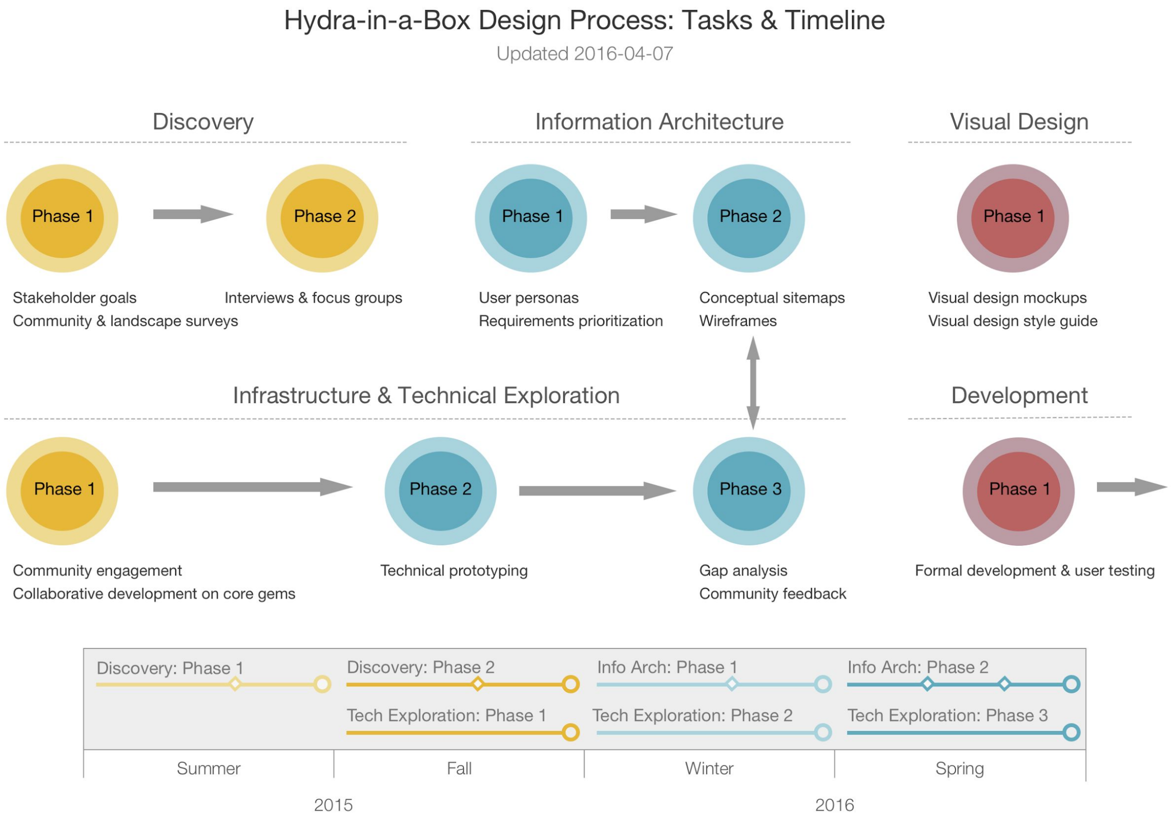
Information Architecture (Winter 2016)

- User requirements and personas
- Requirements - functional & technical
- Models and wireframes

Visual Design (Spring 2016)



# Design Phase



Hydra-in-a-Box Design Process: Tasks & Timeline

Updated 2016-04-07

Discovery

Phase 1 → Phase 2

Stakeholder goals
Community & landscape surveys

Interviews & focus groups

Information Architecture

Phase 1 → Phase 2

User personas
Requirements prioritization

Conceptual sitemaps
Wireframes

Visual Design

Phase 1

Visual design mockups
Visual design style guide

Infrastructure & Technical Exploration

Phase 1 → Phase 2 → Phase 3

Community engagement
Collaborative development on core gems

Technical prototyping

Gap analysis
Community feedback

Development

Phase 1 →

Formal development & user testing

Discovery: Phase 1

Discovery: Phase 2

Info Arch: Phase 1

Info Arch: Phase 2

Tech Exploration: Phase 1

Tech Exploration: Phase 2

Tech Exploration: Phase 3

Summer

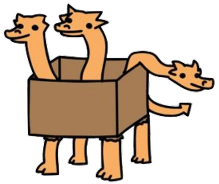
Fall

Winter

Spring

2015

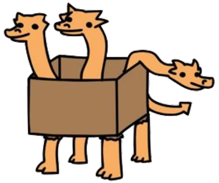
2016



# Key Areas of Progress

Design, Requirements and Specifications team:

- Community survey insights
- Analysis of user interviews, focus groups
- Content types requirements for data modeling

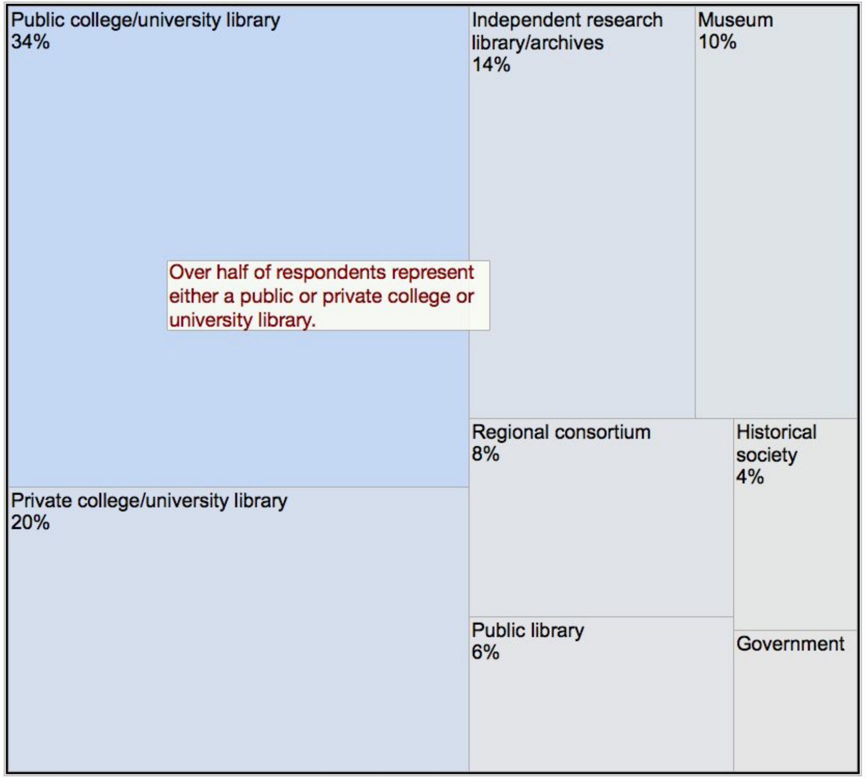


# Community Survey

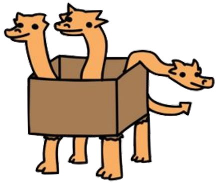
256 complete responses

311 repositories

Mostly small, US academic libraries

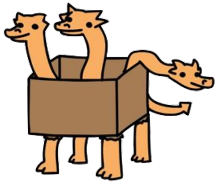


Over half of respondents represent either a public or private college or university library.



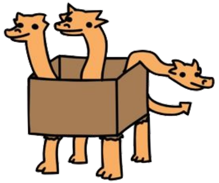
# Survey Insights

- Expectations of our project
- Satisfaction levels
  - Users of hosted services tend to be more satisfied than users of local deployments
- Strengths and weaknesses of existing repository options
- 53% plan to migrate to another system
  - Most to a Fedora-based solution
  - Rest are “not sure” what’s next

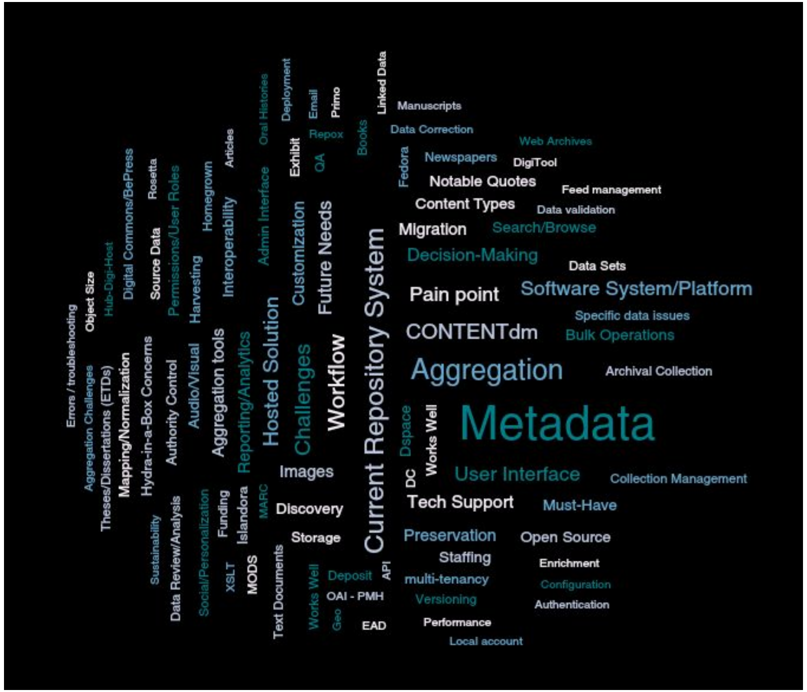


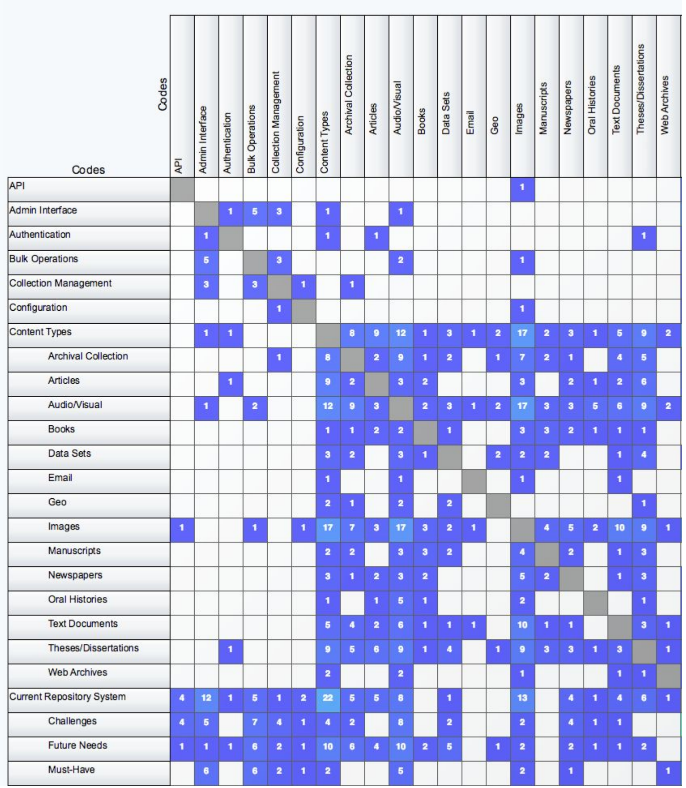
# User Interviews

- Completed 21 individual or small-group interviews and 4 focus groups
  - 55 individuals in total
  - 46 institutions in the US and Canada
  - 29 hours of recorded content
- Interviews held either in-person or through videoconference; focus groups held in-person
- Coded and analyzed process to further identify potential requirements



# Content Analysis Visualizations

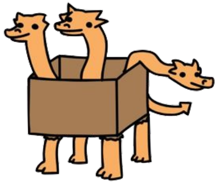




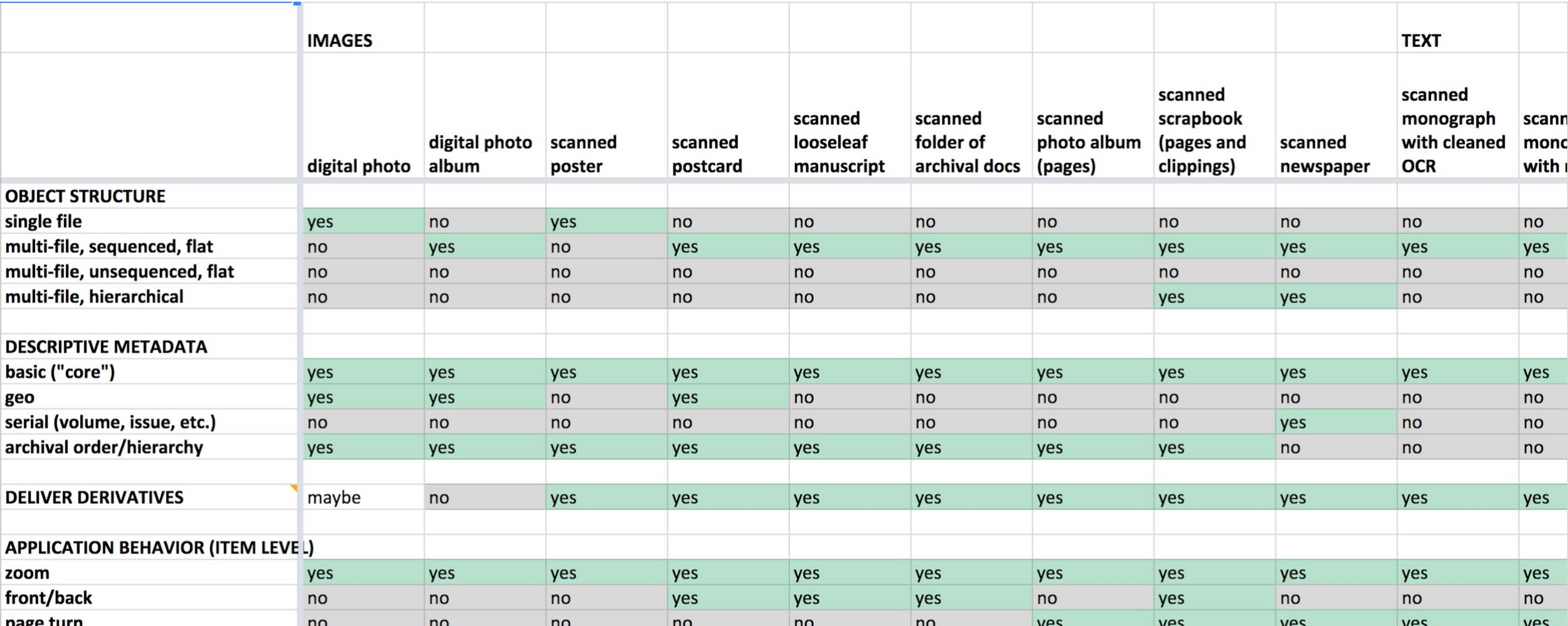
| Codes | API | Admin Interface | Authentication | Bulk Operations | Collection Management | Configuration | Content Types | Archival Collection | Articles | Audio/Visual | Books | Data Sets | Email | Geo | Images | Manuscripts | Newspapers | Oral Histories | Text Documents | Theses/Dissertations | Web Archives |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| API | | | | | | | | | | | | | | | 1 | | | | | | |
| Admin Interface | | | 1 | 5 | 3 | | 1 | | | 1 | | | | | | | | | | | |
| Authentication | | 1 | | | | | 1 | | 1 | | | | | | | | | | | 1 | |
| Bulk Operations | | 5 | | | 3 | | | | | 2 | | | | | 1 | | | | | | |
| Collection Management | | 3 | | 3 | | 1 | | 1 | | | | | | | | | | | | | |
| Configuration | | | | | 1 | | | | | | | | | | 1 | | | | | | |
| Content Types | | 1 | 1 | | | | | 8 | 9 | 12 | 1 | 3 | 1 | 2 | 17 | 2 | 3 | 1 | 5 | 9 | 2 |
| Archival Collection | | | | | 1 | | 8 | | 2 | 9 | 1 | 2 | | 1 | 7 | 2 | 1 | | 4 | 5 | |
| Articles | | | 1 | | | | 9 | 2 | | 3 | 2 | | | | 3 | | 2 | 1 | 2 | 6 | |
| Audio/Visual | | 1 | | 2 | | | 12 | 9 | 3 | | 2 | 3 | 1 | 2 | 17 | 3 | 3 | 5 | 6 | 9 | 2 |
| Books | | | | | | | 1 | 1 | 2 | 2 | | 1 | | | 3 | 3 | 2 | 1 | 1 | 1 | |
| Data Sets | | | | | | | 3 | 2 | | 3 | 1 | | | 2 | 2 | 2 | | | 1 | 4 | |
| Email | | | | | | | 1 | | | 1 | | | | | 1 | | | | 1 | | |
| Geo | | | | | | | 2 | 1 | | 2 | | 2 | | | | | | | | 1 | |
| Images | 1 | | | 1 | | 1 | 17 | 7 | 3 | 17 | 3 | 2 | 1 | | | 4 | 5 | 2 | 10 | 9 | 1 |
| Manuscripts | | | | | | | 2 | 2 | | 3 | 3 | 2 | | | 4 | | 2 | | 1 | 3 | |
| Newspapers | | | | | | | 3 | 1 | 2 | 3 | 2 | | | | 5 | 2 | | | 1 | 3 | |
| Oral Histories | | | | | | | 1 | | 1 | 5 | 1 | | | | 2 | | | | | 1 | |
| Text Documents | | | | | | | 5 | 4 | 2 | 6 | 1 | 1 | 1 | | 10 | 1 | 1 | | | 3 | 1 |
| Theses/Dissertations | | | 1 | | | | 9 | 5 | 6 | 9 | 1 | 4 | | 1 | 9 | 3 | 3 | 1 | 3 | | 1 |
| Web Archives | | | | | | | 2 | | | 2 | | | | | 1 | | | | 1 | 1 | |
| Current Repository System | 4 | 12 | 1 | 5 | 1 | 2 | 22 | 5 | 5 | 8 | | 1 | | | 13 | | 4 | 1 | 4 | 6 | 1 |
| Challenges | 4 | 5 | | 7 | 4 | 1 | 4 | 2 | | 8 | | 2 | | | 2 | | 4 | 1 | 1 | | |
| Future Needs | 1 | 1 | 1 | 6 | 2 | 1 | 10 | 6 | 4 | 10 | 2 | 5 | | 1 | 2 | | 2 | 1 | 1 | 2 | |
| Must-Have | | 6 | | 6 | 2 | 1 | 2 | | | 5 | | | | | 2 | | 1 | | | | 1 |

# Interviewee's Notable Quote

"… How many of these different systems do you need? You can have your digital collections with images and documents, you can have your IR, you can have your digital preservation system, and you can add Omeka on top of that to do exhibits. It's just too much to have four or five different systems."



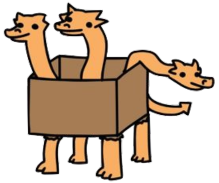
# Content Type Analysis



| | IMAGES | | | | | | | | | TEXT | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | digital photo | digital photo album | scanned poster | scanned postcard | scanned looseleaf manuscript | scanned folder of archival docs | scanned photo album (pages) | scanned scrapbook (pages and clippings) | scanned newspaper | scanned monograph with cleaned OCR | scann monc with |
| **OBJECT STRUCTURE** | | | | | | | | | | | |
| **single file** | yes | no | yes | no | no | no | no | no | no | no | no |
| **multi-file, sequenced, flat** | no | yes | no | yes | yes | yes | yes | yes | yes | yes | yes |
| **multi-file, unsequenced, flat** | no | no | no | no | no | no | no | no | no | no | no |
| **multi-file, hierarchical** | no | no | no | no | no | no | no | yes | yes | no | no |
| | | | | | | | | | | | |
| **DESCRIPTIVE METADATA** | | | | | | | | | | | |
| **basic ("core")** | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| **geo** | yes | yes | no | yes | no | no | no | no | no | no | no |
| **serial (volume, issue, etc.)** | no | no | no | no | no | no | no | no | yes | no | no |
| **archival order/hierarchy** | yes | yes | yes | yes | yes | yes | yes | yes | no | no | no |
| | | | | | | | | | | | |
| **DELIVER DERIVATIVES** | maybe | no | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| | | | | | | | | | | | |
| **APPLICATION BEHAVIOR (ITEM LEVEL)** | | | | | | | | | | | |
| **zoom** | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes | yes |
| **front/back** | no | no | no | yes | yes | yes | no | yes | no | no | no |
| **page turn** | no | no | no | no | no | no | yes | yes | yes | yes | yes |

# Early Technical Exploration

- Deploying to the Cloud
  - Leverage services for institutions without local infrastructure
- Simplifying installation and configuration
  - Users should not need to be technical to set up and maintain an instance
- Determining a starting point for application development
  - Build on existing community-based work if possible
  - Sufia 7.0 - actively under development

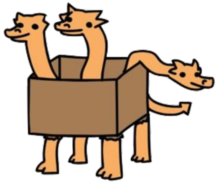


# Repository Development

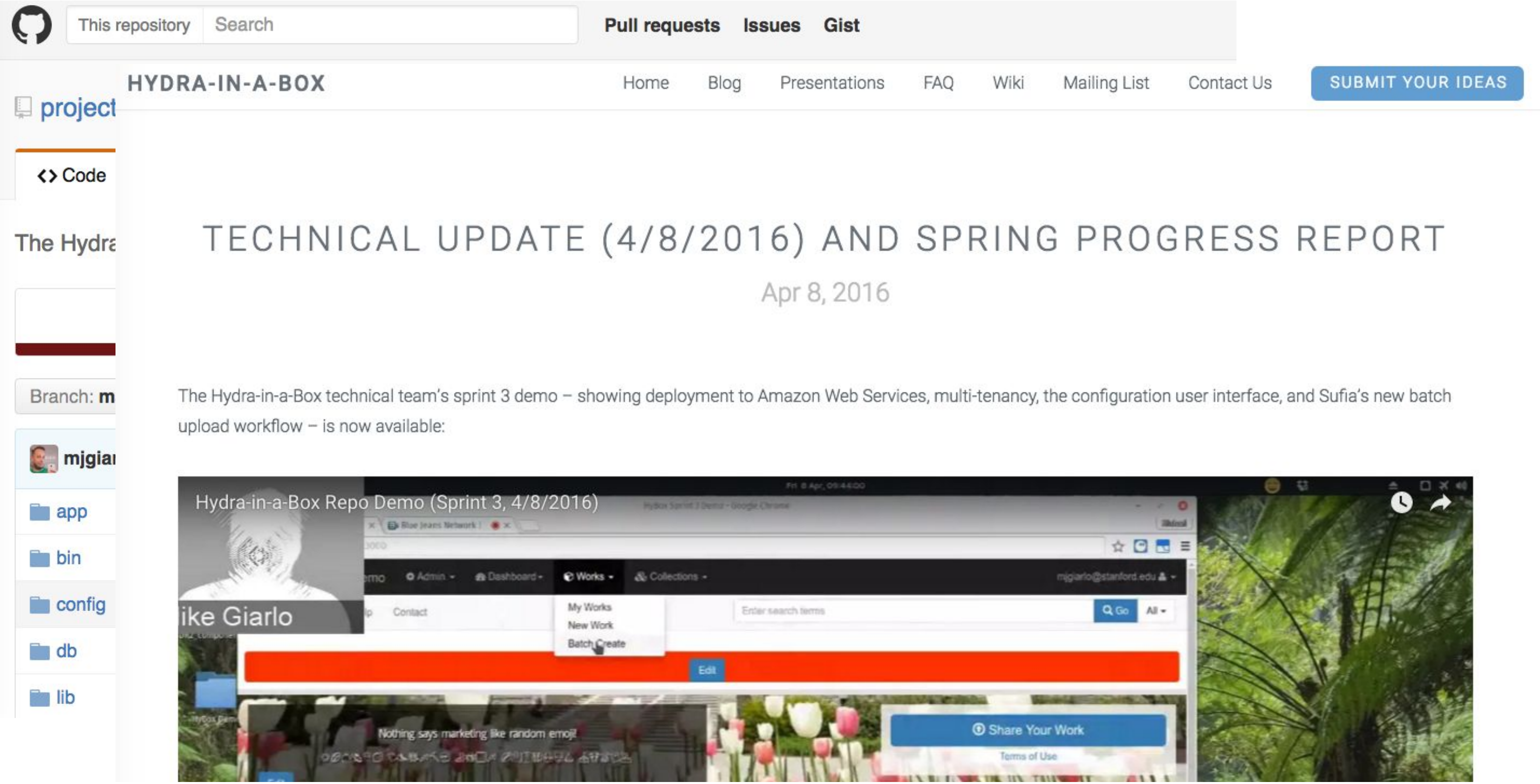
- Assembled an all-star technical team
  - 10 Engineers: software development, data modeling, development operations
  - Contributions from other institutions (Penn State, maybe others)
  - Led by Michael Giarlo

- First work cycle: March - June 2016
  - Series of one-week sprints
  - Recorded demos of iterative progress, available to the public

- First milestone: Deploy our application based on Sufia 7 to the cloud
  - Priority content types
  - Configuration UI
  - Administrative dashboard
  - Batch import



# Follow our progress

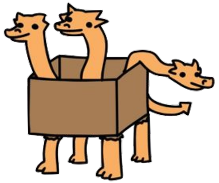


## TECHNICAL UPDATE (4/8/2016) AND SPRING PROGRESS REPORT

Apr 8, 2016

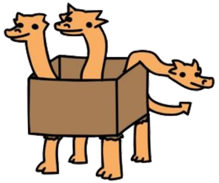
The Hydra-in-a-Box technical team's sprint 3 demo – showing deployment to Amazon Web Services, multi-tenancy, the configuration user interface, and Sufia's new batch upload workflow – is now available:

Hydra-in-a-Box Repo Demo (Sprint 3, 4/8/2016)



# Aggregator Needs

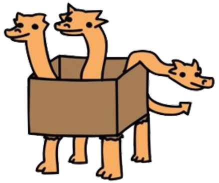
- More flexible mapping standard than XSLT
- Ability to harvest from multiple sources
- Reconciliation services that utilize linked data
- Enhanced quality control tools
- Ability to normalize and create consistencies in data values
- Easily get data in and out
- Robust enough to handle multiple feeds and multiple sources
- Processes to move data from one repository to another resemble aggregation workflows

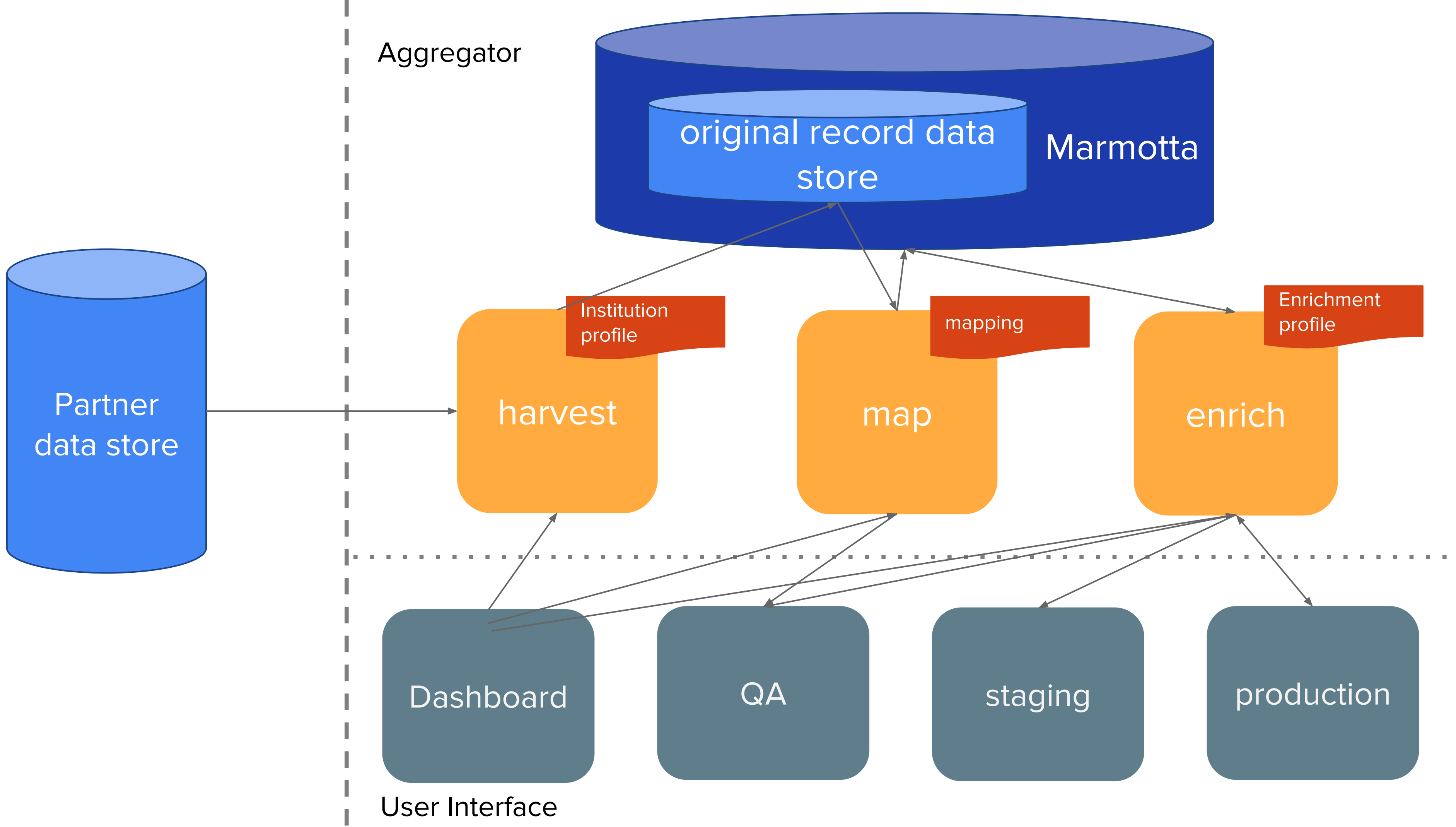


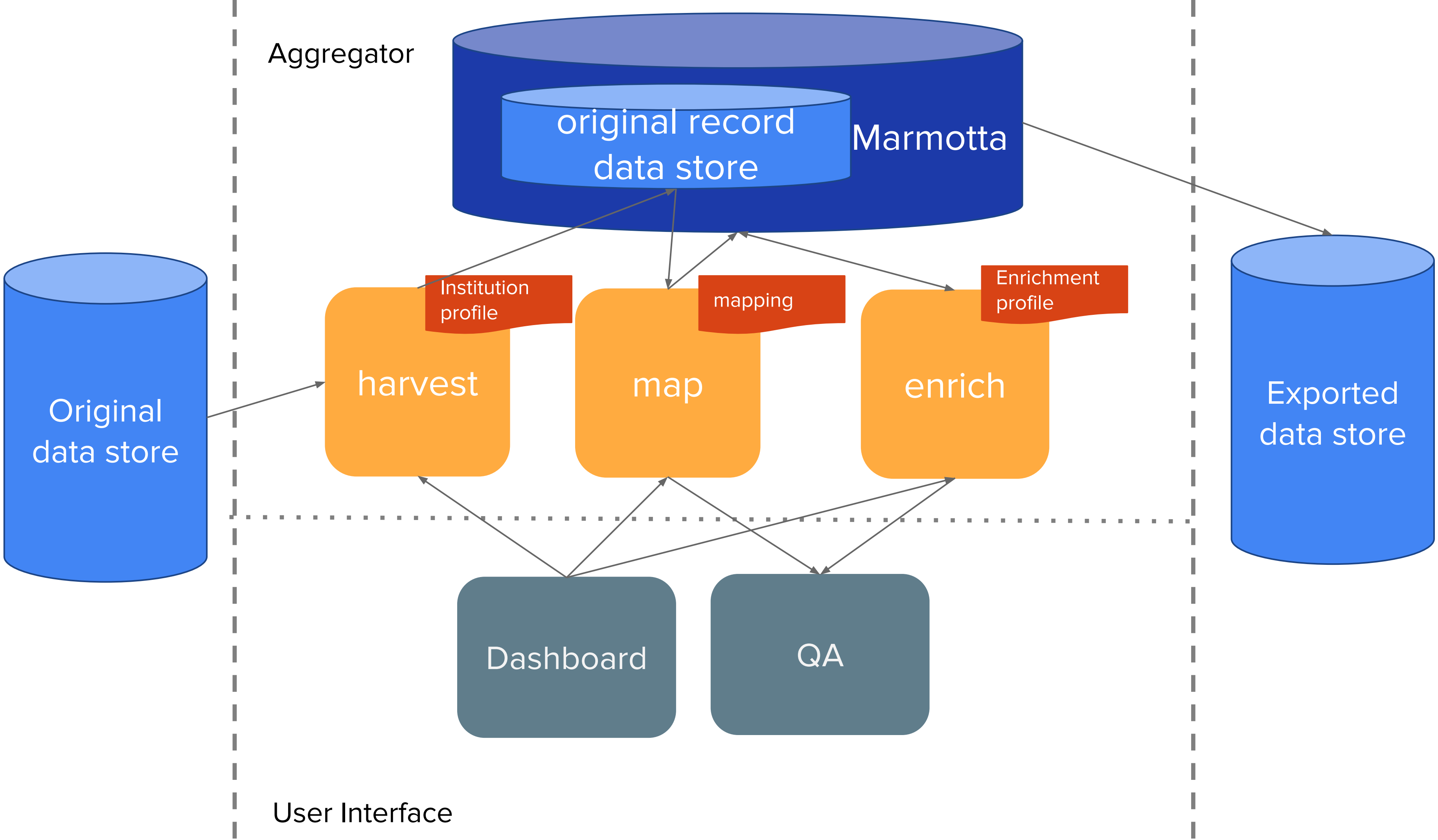
# DPLA's Aggregation System, Heiðrún

Three Main Functions

- Harvest
  - Source agnostic
- Map
  - Mapping DSL expressed in Ruby
  - Maps to RDF triples
- Enrich
  - Modular enrichments written to normalize and enhance data

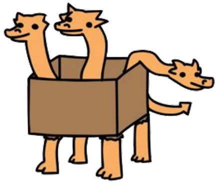



Aggregator
original record data store
Marmotta
Partner data store
Institution profile
harvest
mapping
map
Enrichment profile
enrich
Dashboard
QA
staging
production
User Interface


Aggregator
original record data store
Marmotta
Institution profile
mapping
Enrichment profile
harvest
map
enrich
Original data store
Exported data store
Dashboard
QA
User Interface

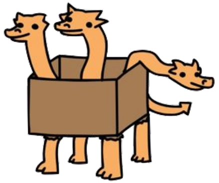
# Roadmap

- Completing requirements now
- April - July
  - Design remaining infrastructure
  - Develop user interface requirements further
- August - November
  - Develop dashboard tools
  - Analyze convergence points with repository
  - Plan for improvements to QA interface
  - Begin User Testing
- November - March 2017
  - Develop QA improvements
  - Refine interfaces and infrastructure
  - Implement job scheduling



# Developing a Hosted Service

- Project partners collaborating to develop requirements for a cloud-hosted service based on the repository product under development
- Market research underway, starting with analysis of information discovered during the design phase
- Evaluating tiered service models depending on needs of potential adopters
- Significant technical work to focus on develop a shared, maintainable, and scalable service



# More Information

Visit our website and blog: **http://hydrainabox.org/**

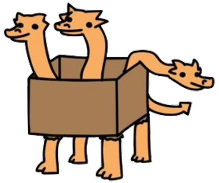
Follow us on Twitter: @HydraInABox

Public information list

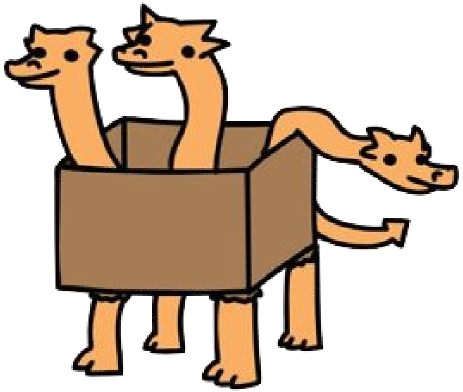
hybox-info@googlegroups.com

Contact us

hybox-contact@googlegroups.com



# Thank You!



Hannah Frost
hfrost@stanford.edu
@feefifofannah

Gretchen Gueguen
gretchen@dp.la
@G_AmSpinnrade

Mark A. Matienzo
mark@dp.la
@anarchivist

http://bit.ly/dplafest-hybox