iRODS Technology Applied to the DataNet Federation Consortium

iRODS User Group Meeting 2015

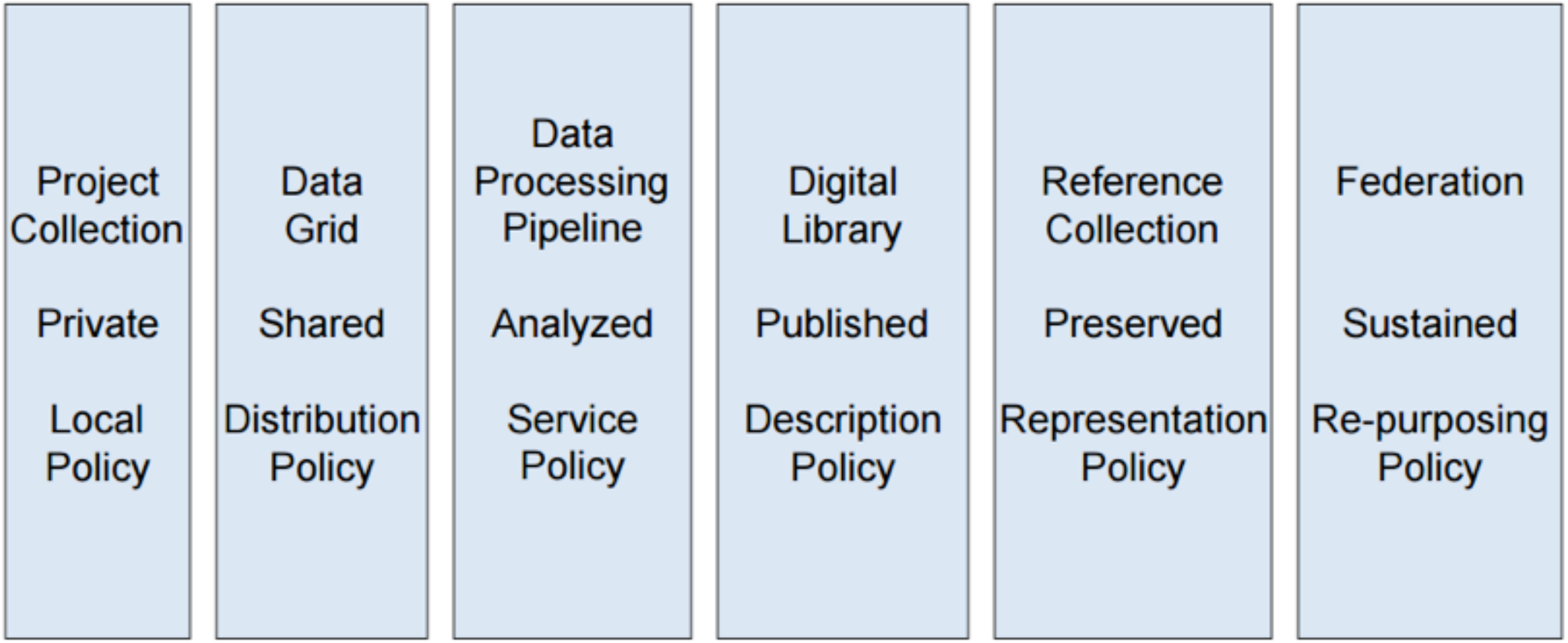
Mike Conway
Hao Xu

This material is based upon work supported by the National Science Foundation under Grant Number OCI 0940841

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation

DFC Goal

To create national scale cyberinfrastructure to support collaborative research



Stages correspond to addition of new policies for a broader community
Virtualize the stages of the collection life cycle through policy evolution

Research Data Life-Cycle

Why we do what we do

"Virtualization of the data life cycle". R. Moore [Online]. Available: https://conferences.tdl.org/tcdl/index.php/TCDL/2010/paper/view/97. [Accessed: 30-May-2015].

Cyberinfrastructure is… (Gannon's 5 + 1)

1. Data search and discovery tools.
2. ~~Security.~~ Policy Management.
3. ~~User private data storage.~~ Ubiquitous data access from private collections to published reference collections
4. Tools for designing and conducting computational experiments.
5. Data provenance tracking.

+1. Ubiquitous access through frameworks, protocols, and human interfaces.

DFC Activity Areas

- ☐ Metadata and discovery
- ☐ Data and computation
- ☐ Ubiquitous access

Metadata for Collection Formation

Project Collection	Data Grid	Data Processing Pipeline	Digital Library	Reference Collection	Federation
Private	Shared	Analyzed	Published	Preserved	Sustained
Local Policy	Distribution Policy	Service Policy	Description Policy	Representation Policy	Re-purposing Policy

- ☐ Curate by humans, or generate automatic metadata via policy + computation
- ☐ Index and organize through projection into various indexes and metadata catalogs for discovery
- ☐ Form new collections via metadata relationships or queries

Metadata Templates

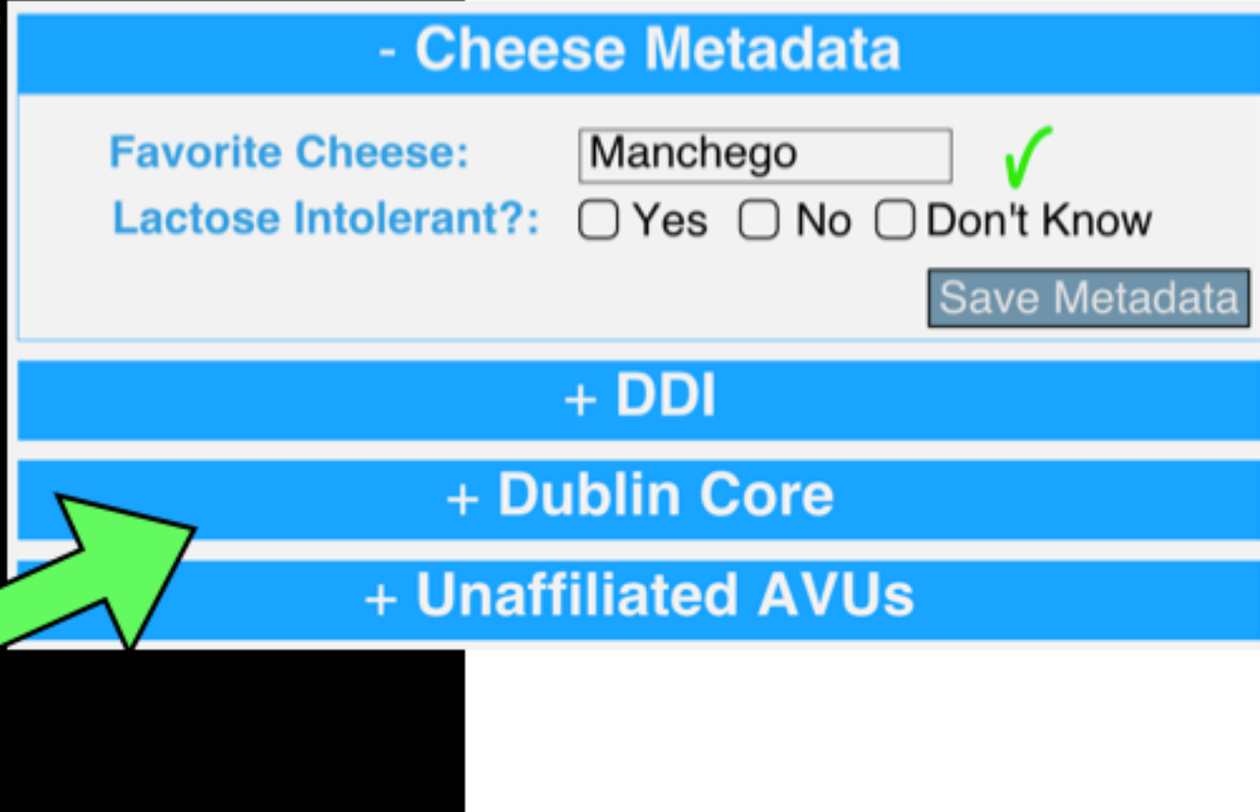
- ☐ Add structure to AVUs
  - ☐ Describe groupings, type, validation, meaning, policies, bindings
  - ☐ Automate data entry for curation
  - ☐ Automate nicely formatted displays
  - ☐ Declarative automated metadata extraction
  - ☐ Simple, flexible

cheese.mdtemplate

```
{
    “name”:  “cheese”,
    “description” : “Cheese Metadata”,
    “author” : “Rick Skarbez”,
    “required” : “true”,
    “elements” : [
       {
          “name” : “favCheese”,
          “description” : “Favorite
          Cheese”,
          “required” : “true”,
          “type” : “STRING”
       },
       {
          “name” : “lactoseIntolerant”,
          “description” : “lactoseIntolerant?”,
          “validationStyle” :
             “IN_LIST”,
          “validationOptions” :
             [“Yes”, “No”,
             “Don’t Know”]
       }
    ]
}
```

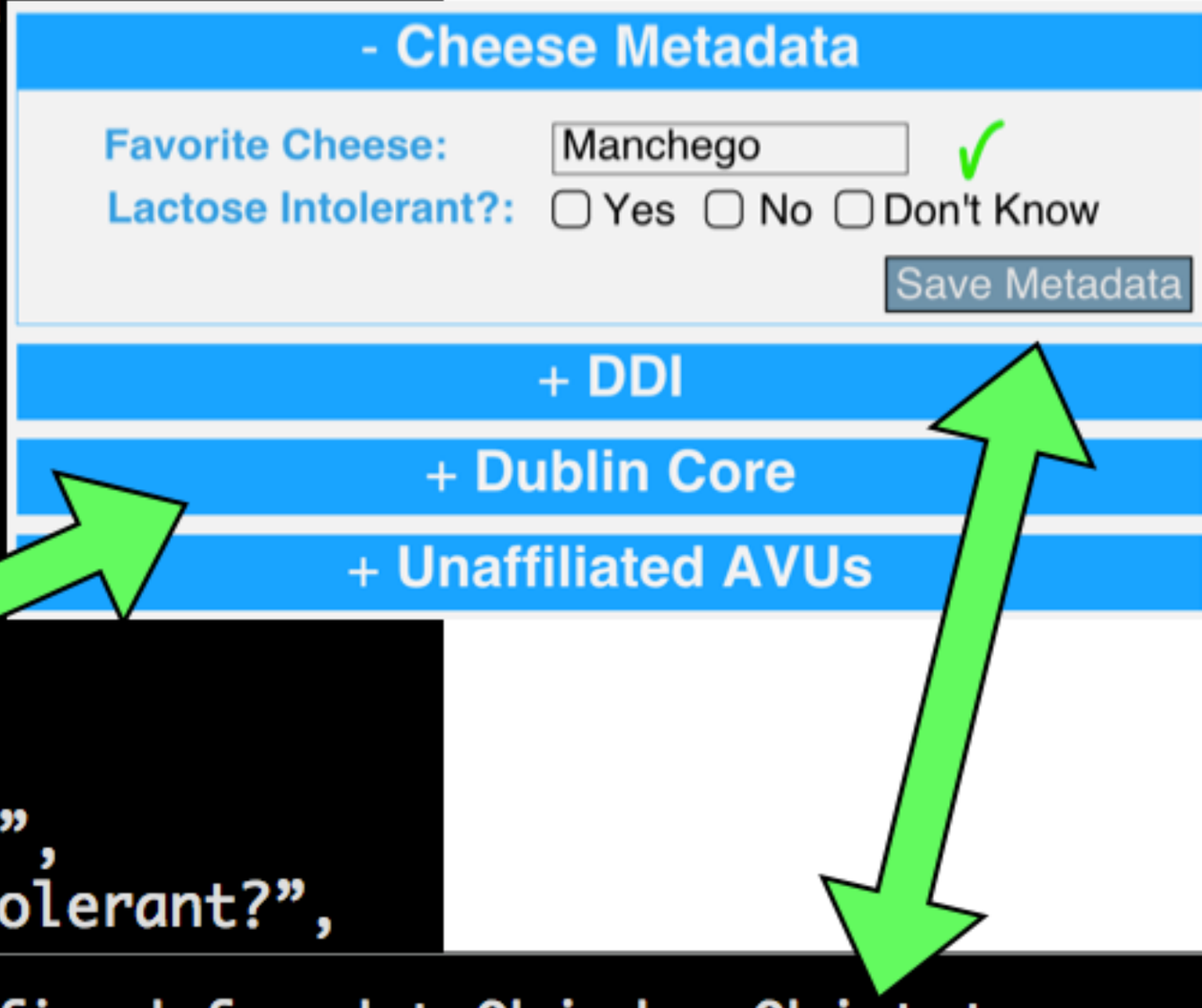
cheese.mdtemplate

```
{
    “name”:  “cheese”,
    “description” : “Cheese Metadata”
    “author” : “Rick Skarbez”,
    “required” : “true”,
    “elements” : [
       {
          “name” : “favCheese”,
          “description” : “Favorite
          Cheese”,
          “required” : “true”,
          “type” : “STRING”
       },
       {
          “name” : “lactoseIntolerant”,
          “description” : “lactoseIntolerant?”,
          “validationStyle” :
             “IN_LIST”,
          “validationOptions” :
             [“Yes”, “No”,
             “Don’t Know”]
       }
    ]
}
```



cheese.mdtemplate

```
{
    “name”:  “cheese”,
    “description” : “Cheese Metadata”
    “author” : “Rick Skarbez”,
    “required” : “true”,
    “elements” : [
       {
          “name” : “favCheese”,
          “description” : “Favorite
          Cheese”,
          “required” : “true”,
          “type” : “STRING”
       },
       {
          “name” : “lactoseIntolerant”,
          “description” : “lactoseIntolerant?”,
          “validationStyle”
             “IN_LIST”,
          “validationOptior
             [“Yes”, “No”,
             “Don’t Know”]
       }
    ]
}
```

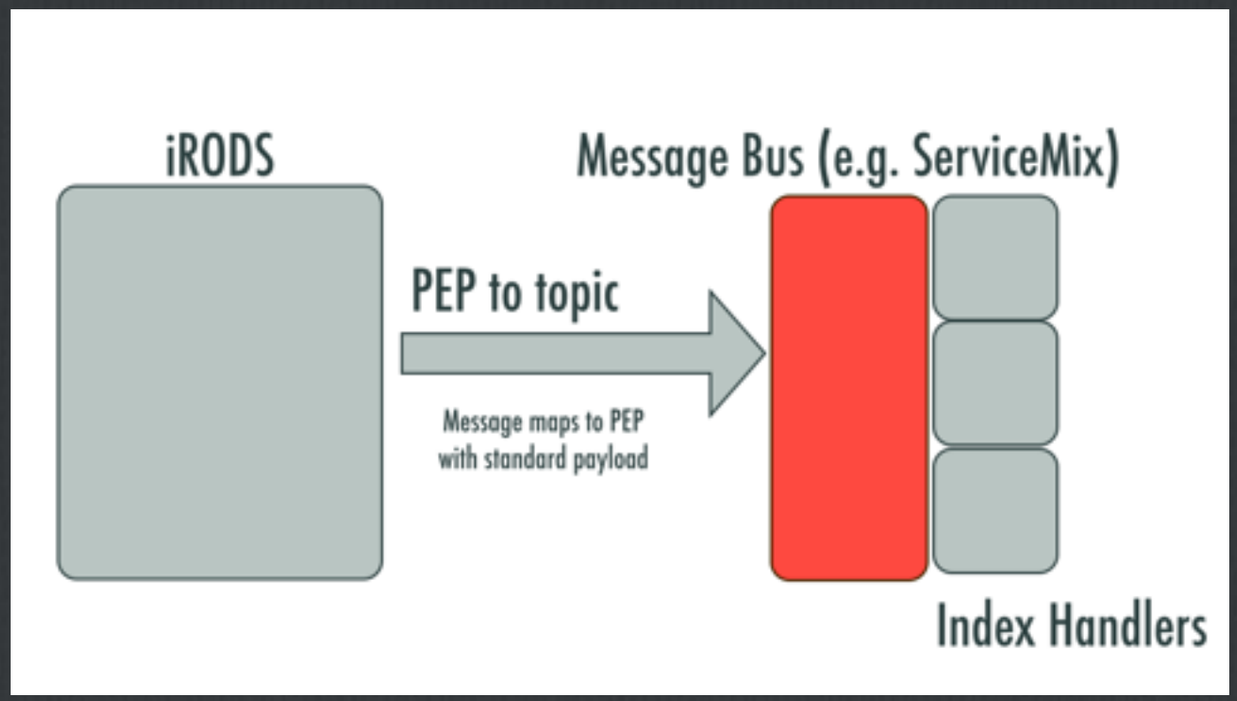



```
AVUs defined for dataObj demoObj.txt:
attribute: favCheese
value: Manchego
unit: <UUID assigned to cheese.mdtemplate>
---
attribute: lactoseIntolerant
value: No
unit: <UUID assigned to cheese.mdtemplate>
```

Indexing Framework

- ☐ A shift in iRODS emphasis, back to our roots?
- ☐ iRODS' sweet spot is policy management of data, so focus on policy management of metadata. iRODS as a discovery tool…not so much. GenQuery versus SPARQL…which would you pick?
- ☐ iRODS as a canonical, trusted store of metadata, and a framework to attach computation to data mediated by policies.

Basic Indexing Topology



- Asynch messaging driven by iRODS PEPs track data and metadata events and deliver to the indexers
- Indexers are a simple OSGi component model that plug in and receive these messages, and can do near-real-time synchronization of external indexers
- Different parts of grid can be projected into different indexers
- Metadata is preserved and managed in grid, indexes are ephemeral, and can be created and re-created as needed

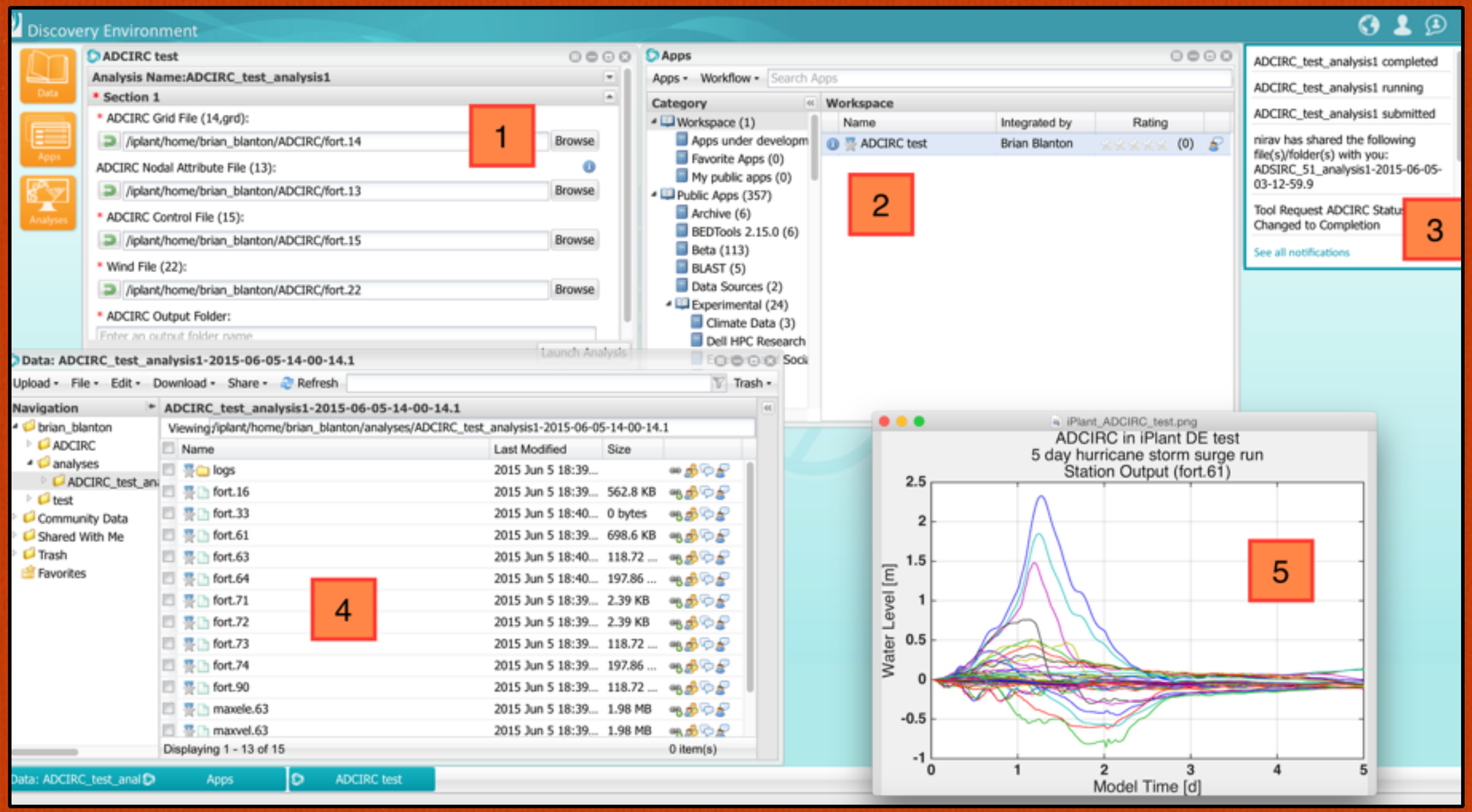
Two Indexer Types

- ☐ Metadata indexers (AVUs and Catalog), try to push as much info into JSON to reduce re-query of catalog. Example…RDF statements stored in a standard AVU format -> triple store

- ☐ Data indexers (file content), try to localize indexing to physical storage location and handle multiple indexers interested in same data…Example…ElasticSearch and file format recognition

Virtual Collections

- Any index that one may query and generate an ILS -LA type listing can be a virtual collection (SQL, SPARQL, boolean search)
- Virtual collections serialized as JSON, which are tied to drivers to produce the listing
- New collections may be created orthogonal to hierarchical arrangement, on an equal footing to the iRODS catalog.



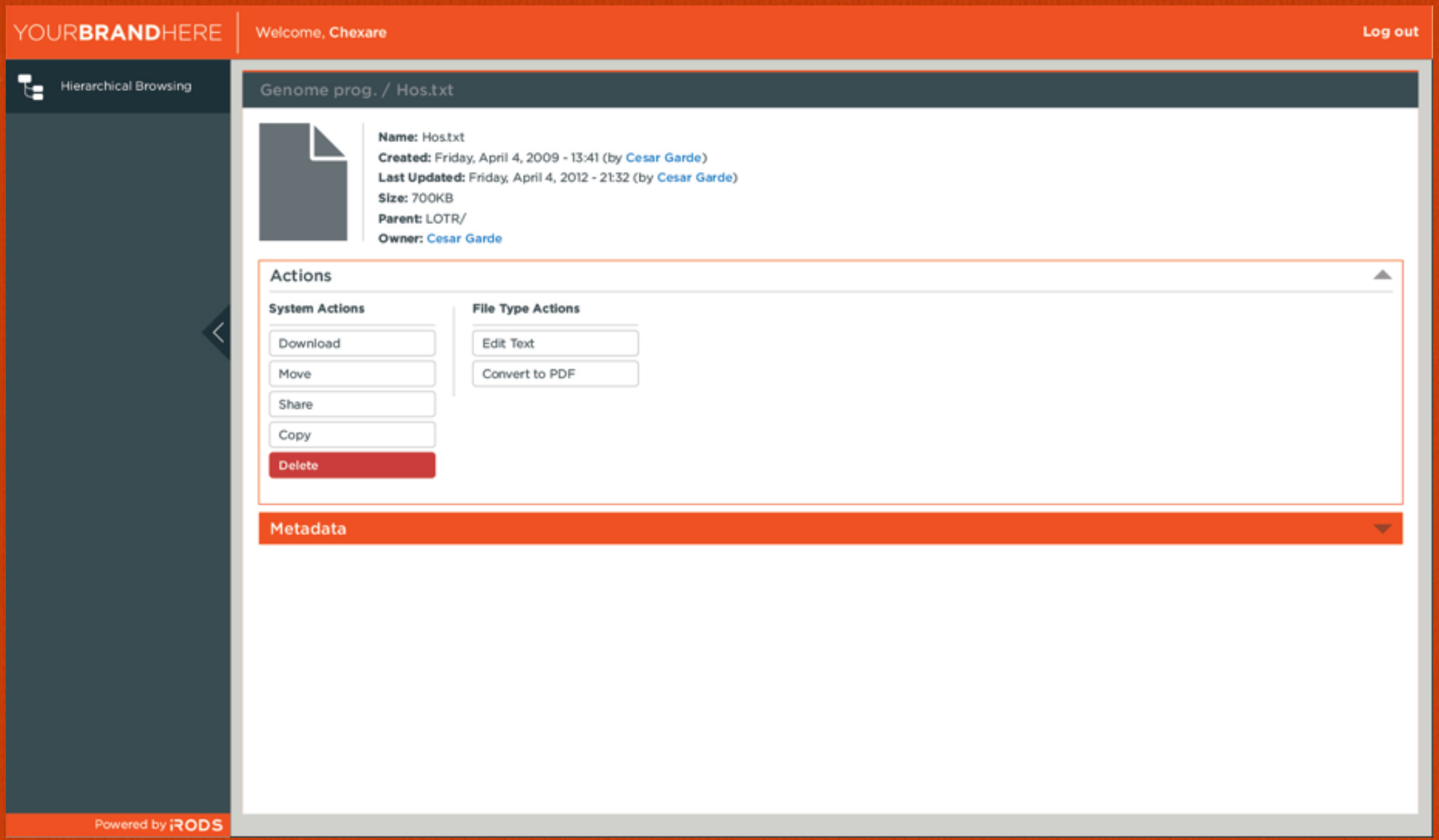
Computation + Data

working to integrate and extend iPlant's Discovery Environment

From Brain Blanton @ RENCI running on iPlant Discovery Environment

Pluggable Computation

- ☐ Abstraction of ‘App’ with driver/back end API
- ☐ Shared notification framwork within iRODS
- ☐ Use of iRODS indexing framework for audit logs
- ☐ Provenance tracking and reproducibilty
- ☐ Extend clients through configuration of applications within iRODS



Computation as a pluggable abstraction

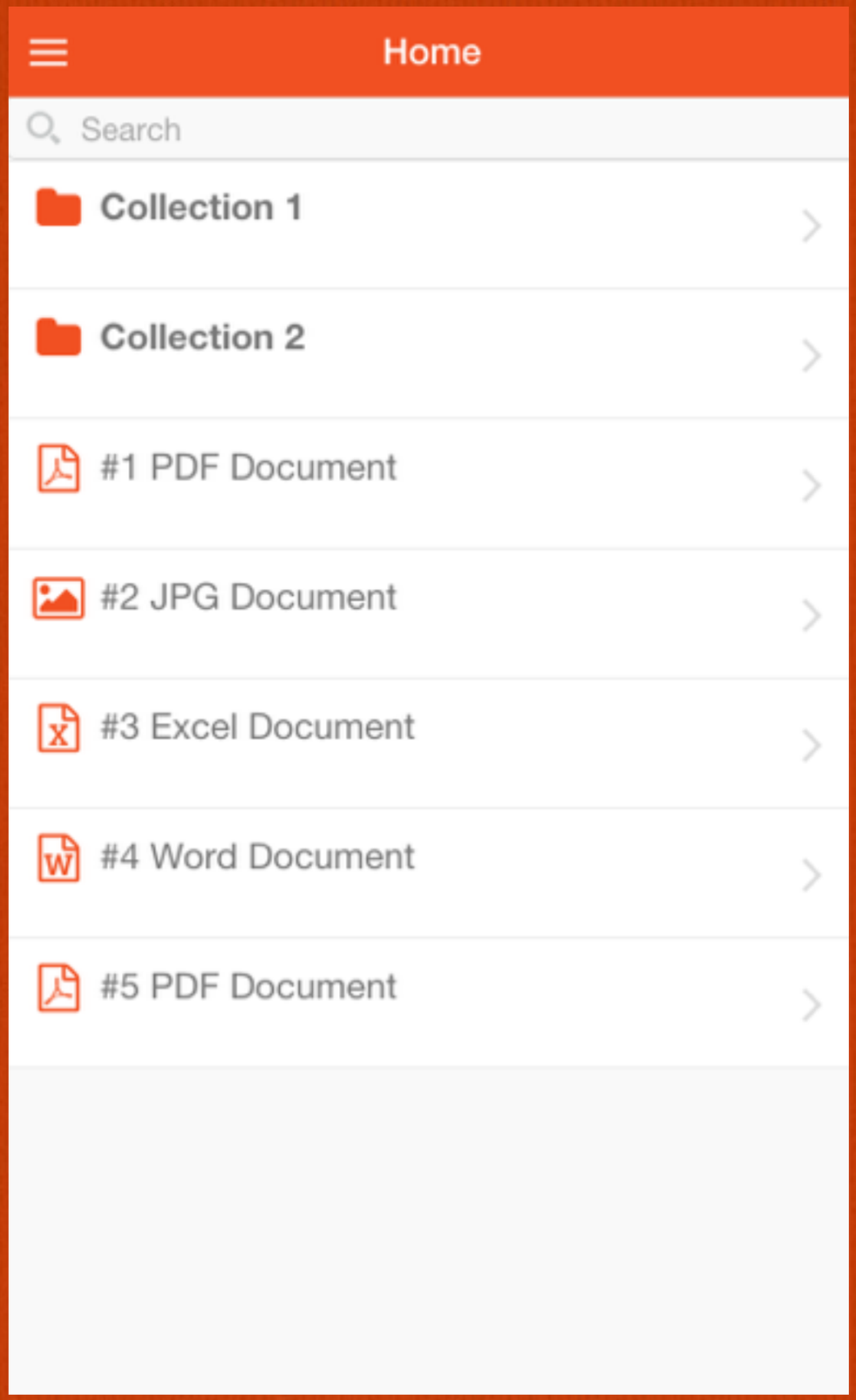
- ☐ Abstraction of ‘application’
  - ☐ iRODS Rule
  - ☐ WSO
  - ☐ Condor
  - ☐ Docker
  - ☐ HPC Grid
  - ☐ Indexing Framework Component

Other Topics

- ☐ Too little time!
- ☐ We'll see web interfaces
- ☐ We'll talk about API and framework/protocol integration

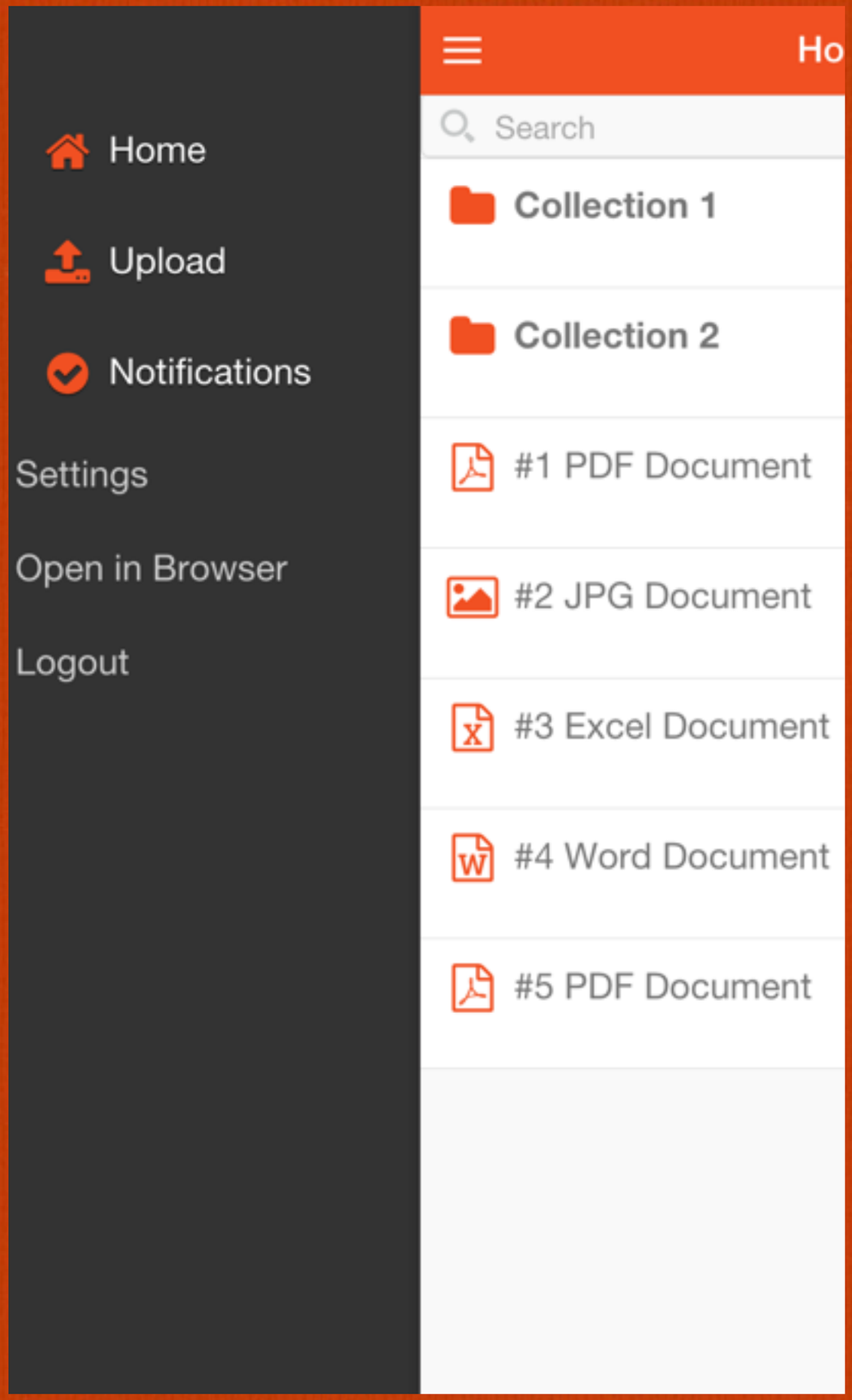
Mobile Preview

A companion to web interface development for ubiquitous mobile access to selected data and notifications



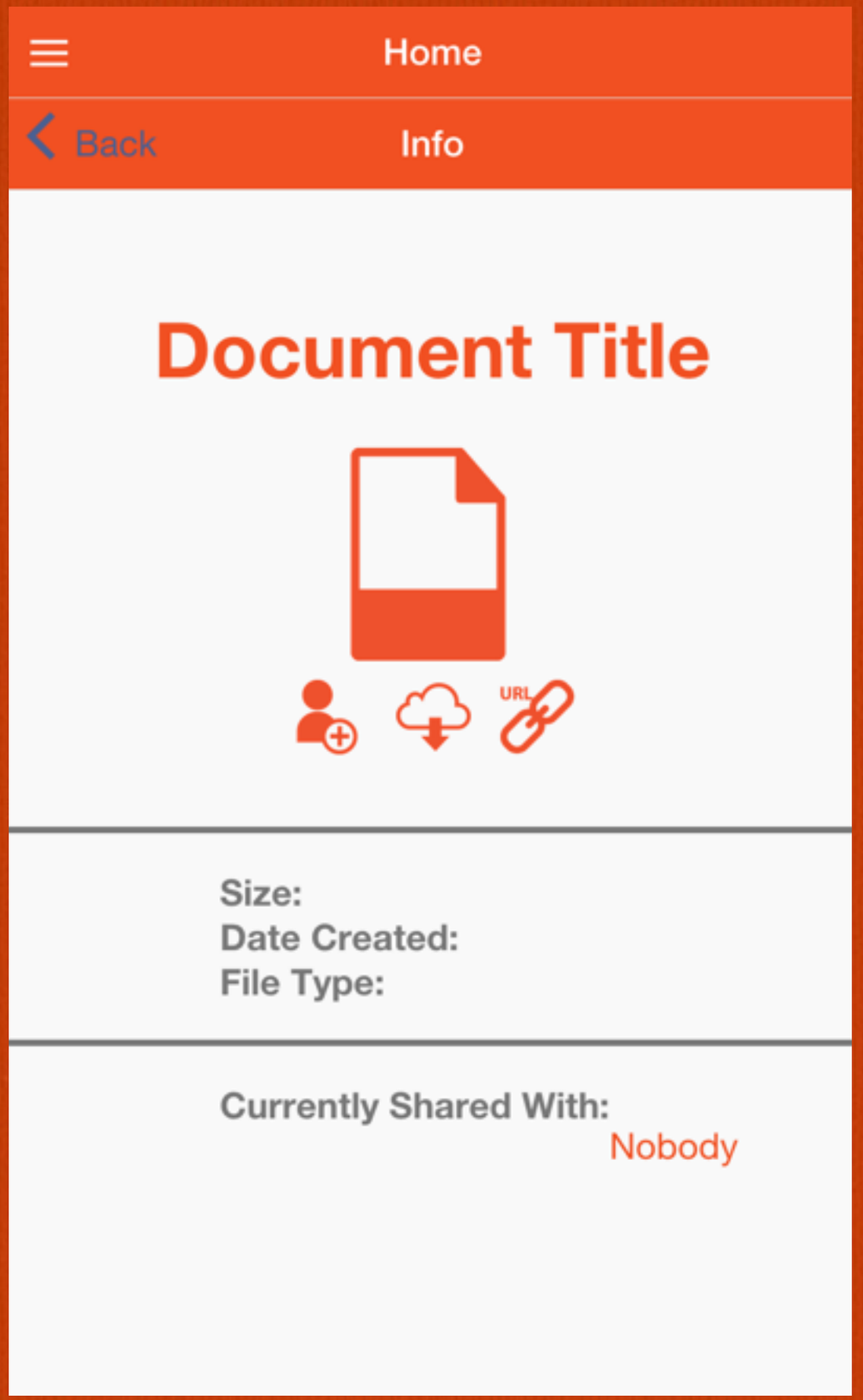
Mobile Preview

A companion to web interface development for ubiquitous mobile access to selected data and notifications



Home Page for Data

Simple actions to download and share



Notification-Centric

A lightweight, common notification mechanism for iRODS is in the offing