



iRODS and NIEHS Environmental Health Science

Mike Conway – mike.conway@nih.gov

Deep Patel – deep.patel@nih.gov

National Institute of Environmental Health Sciences

What we do now

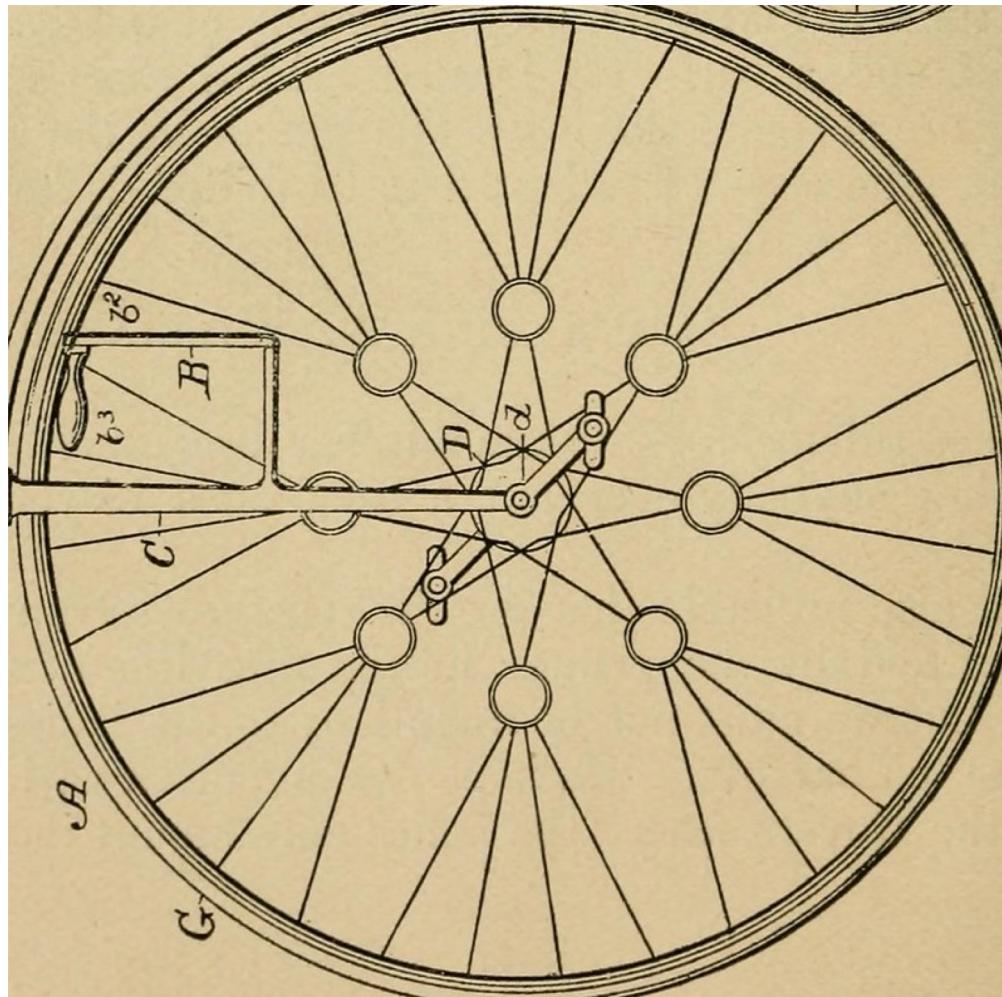
- Focus on environmental health, managing intramural research data
- iRODS is heavily involved in managing sequencing data
- We are tracking the [NIH Strategic Plan for Data Science](#)
- We are getting more and more involved with the Health Sciences community via [GA4GH](#)



Building Flexibility is the theme

- Real world problems
- Cheap and simple solutions that don't pollute community code
- In the spirit of open source we want to develop broadly usable tools/patterns that can provide foldback of effort from the community
- How to add microservices around core iRODS capabilities that simplify common tasks
- Look at how some of these raw materials lead towards some 'bigger fish' that the iRODS community can tackle together

iRODS Role



iRODS strength is as a hub that can serve as the canonical data and metadata store

- Monitor and manage the data
- Index and leverage existing search technology, offloading the search problem to existing solutions
- Serve as a data curation hub, a curation ‘bus’ in a cloudy, multi-repository world



<https://flic.kr/p/S8qcUo>

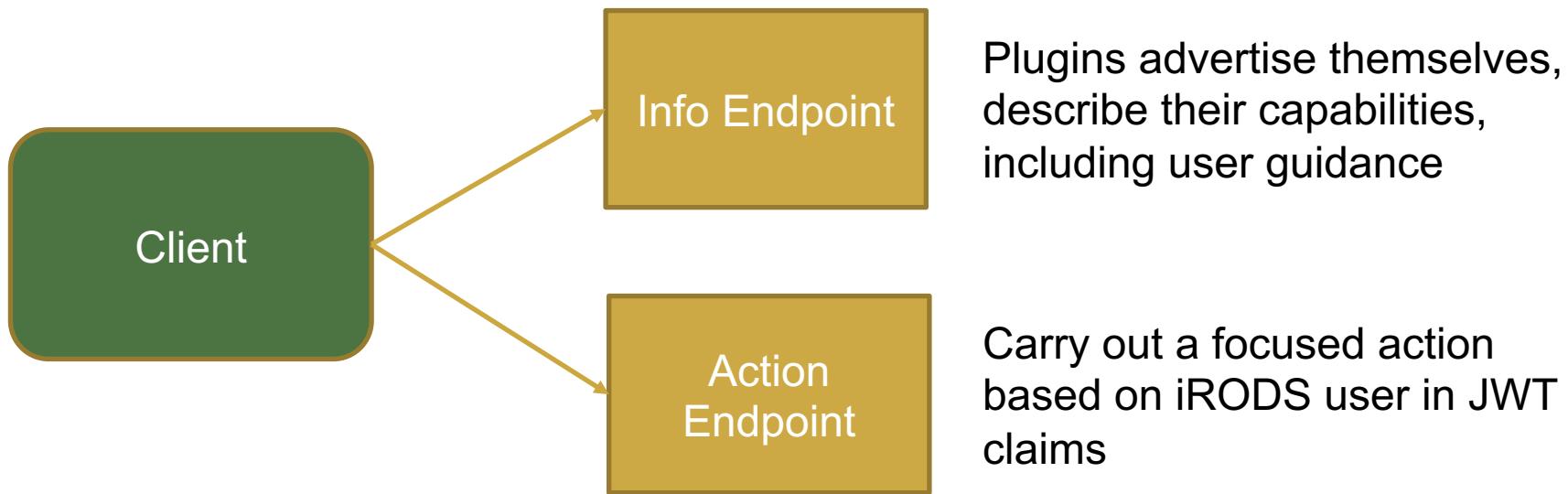


<https://flic.kr/p/2gZPyqN>

Search and Indexing

File Carts and Publishing

Using REST-ful API with a general ‘plugin’ model for Metalnx or any other microservice architecture



Client:

- Obtains a JWT by iRODS auth
- Has a configured list of endpoint addresses, interrogates info endpoints to establish the available plugins

Search!

- Simple way to plug in light-weight search capabilities (focused on a particular task/persona)
- Agnostic about underlying search technology, simpler than the QueryArrow approach?
- User friendly
 - Default Google-like search using baked in assumptions for unstructured strings
 - Attribute:Search Term advanced functions with wildcards/query operators for building more sophisticated searches
- UI and API-friendly output
 - Google like search results with sublinks
 - ils –LA listings

Search API

The screenshot shows the Elasticsearch API documentation interface. At the top, there are navigation icons and a user profile. Below is a section titled "Info" with endpoints for describing search capabilities and options. It includes two "GET" requests: one for index types and another for search attributes. Under the "Search" section, there is a "POST" request for generic search. On the right side of the interface, three orange arrows point from the text descriptions to the corresponding API endpoints.

Info Endpoints for describing search capabilities and options

GET /indexes Find index types supported by this api

GET /attributes/{index_name} Find search attribute terms for a specific index

Search Search on index

POST /search Generic search on one or all available indexes

← Endpoint for registration
← Describe search schema
← Execute a search

A search plugin has a schema that describes its purpose and targeted persona. The available search attributes are described by the /attributes endpoint.

Many small plugins versus one complicated, multifaceted plugin.

Metalnx configuration

```
#####
# Pluggable search configuration. Turn on and off pluggable search globally, and
configure search endpoints.
# N.B. pluggable search also requires provisioning of the jwt.* information above
#####
# configured endpoints, comma delimited in form https://host.com/v1
pluggablesearch.endpointRegistryList=http://proj_sample_search:8082/v1,http://
metadata_search:8082/v1
# enable pluggable search globally and show the search GUI components
pluggablesearch.enabled=true
classicsearch.enabled=false
# JWT subject claim used to access search endpoint for data gathering. User
searches will utilize the name of the individual
pluggablesearch.endpointAccessSubject=pluggablesearch
# timeout for info/attribute gathering, set to 0 for no timeout
pluggablesearch.info.timeout=0
# timeout for actual search, set to 0 for no timeout
pluggablesearch.search.timeout=0
```

Post-discovery in a web UI

The screenshot shows the metalnx web interface. At the top, there's a dark header bar with the text "National Institutes of Health - U.S. Department of Health and Human Services". Below it, a secondary header bar has the text "The Junction for NIEHS Staff | Go to NIEHS Public Site". To the right of this bar is a search input field with placeholder text "Search here ...". A "Search" button is located to the right of the search input.

The main content area has a teal background and features the "metalnx" logo. On the left side of this area, there's a sidebar with several navigation items: "Resources" (with a database icon), "Rules" (with a document icon), "Users" (with a person icon), "Groups" (with a group icon), and "Collections" (with a folder icon). Below these is a "Global Search" input field with a magnifying glass icon.

The central part of the page contains a search interface. It includes a "Select Schema" dropdown menu with options like "Epigenomics Projects", "Epigenomics Samples and Runs", and "Metadata Search". Next to it is an "Enter a search" input field and a green "Search" button.

At the bottom of the page, there's a horizontal navigation bar with various links: "For Lab Staff", "For Office Staff", "For Managers", "Working Here", "Computer Support", "Library", "News", "Policies & Ethics", "Purchasing", "Safety & Security", "Training", and "Division".

Available schema have been discovered by polling the /info endpoints

Describing the schema attributes

The screenshot shows the metalnx web application interface. On the left is a sidebar with navigation links: Resources (databases), Rules (document icon), Users (person icon), Groups (globe icon), Collections (file folder icon), Global Search (magnifying glass icon), Templates (cloud icon), Shared Links (link icon), Favorites (star icon), Public (globe icon), and Trash (trash bin icon). The main area has a header with "Epigenomics Samples and Runs" and a search bar. A modal window titled "Hide Hint" provides search examples and instructions. Below is a table titled "Available Fields" with columns for field name, description, and example.

Available Fields	Description	Example
AnalystEmail	Email ID of the analyst	brian.papas@nih.gov
AnalystName	Name of the analyst	Brian Papas
ASPNumber	Animal Study Protocol approval number	2011-0016
Branch	Branch name, affiliated with project PI	DIR/STL
Date	Date of sample submission form	
GenomeRef		mm10
Index1		AGTTCC
Index2		AGTTCC
IRBNumber	Institutional Review Board protocol number	12-N-0095
LibrariesPreparedBy	Was the library prepared by Investigator or Epigenomics Core	Investigator

Search Options

- Enter a plain string...
 - The plugin, appropriate to the persona, will translate into a default query
 - e.g. sample name, sample name unique, run id like NOVA01*
- Enter Attribute:Query
 - The plugin will fashion a query appropriate to the targeted attributes
 - e.g. GenomeRef:mm10 PreparationKit:Nextera
- Attribute:Query style can also support ‘builder’ type query interfaces in later iterations

Search result options

Search can support familiar web search results with customizable properties and sublinks

```
search_data ▼ {  
    index-schema-  
    description  
  
    search_result  
        ▼ [ ▼ {  
            title           string  
                          Descriptive title for search result  
  
            url_link        string  
                          Resolvable https link to result data location  
  
            subtitle        string  
                          Optional subtitle that can be presented as a highlight, publication info, etc  
  
            content_text    string  
                          Bag of attribute-value paired metadata attached to search hit  
  
            properties      result_properties > {...} ↩  
            links          search_data_linkset > {...} ↩  
        }]  
}
```

Results in familiar web search format with built-in support for headers and sub-links

The screenshot shows the metaINX search results interface. The top navigation bar includes the metaINX logo, a user dropdown (conwaymc), and a search bar with the query "Epigenomics Samples and Runs" and filter "ns*". The left sidebar contains links for Resources, Rules, Users, Groups, Collections, Global Search, Templates, Shared Links, Favorites, Public, and Trash. The main content area displays a "Search Results" section for "Epigenomics Samples and Runs". It shows a title "Library prepared by: Epigenomics Core" with a direct link, followed by "Principal investigator: [REDACTED]" and "Sequencing system: NextSeq-High Output". Below this, a "Related sample fastq files" section lists several files, each with a small icon and a direct link. Arrows from the text descriptions on the right point to the corresponding elements in the interface.

Title with direct link to a run folder

Properties automatically formatted

Sub-links or 'see also' links

Grid style results now available

List Grid

Name	Modified	Size (kB)
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-
[REDACTED]	Wed Dec 31 1969 19:00:00 GMT-0500 (Eastern Standard Time)	-

Virtual Collections and FAIR data

- Originally implemented in DFC but is fairly powerful when combined with this plugin pattern
- Any query that returns an ils –LA listing can serve as a virtual collection. Saved queries can become a new iRODS collection
- Virtual collections can span federations, these query endpoints could be added to ils, icd
- MetaInx already supports proxy pages (off by default) for query results that show available metadata and allows requests for access for ‘metadata is public, data is private’ scenarios

The iRODS indexing capability and the search plugin architecture are being aligned

- Current NIEHS file props and AVU search plugin –
<https://github.com/angrygoat/pluggable-search-props-and-metadata>
- Current file props and AVU indexer implementation
 - Based on a utility visitor framework
 - <https://github.com/DICE-UNC/jargon/tree/master/jargon-datautils/src/main/java/org/irods/jargon/datautils/visitor>
 - Filtering
 - Restart
 - Automatic metadata stack management
 - <https://github.com/trel/metadata-and-file-props-indexer>

Publication and the file shopping cart!

A file shopping cart is an iRODS ‘hidden’ file in the user home area that contains a manifest of selected paths in iRODS

The screenshot shows the metaInx web interface. On the left, there is a sidebar with various navigation links: Resources, Rules, Users, Groups, Collections, Global Search, Templates, Shared Links, Favorites, Public, and Trash. Below this is a 'Navigation Tools' dropdown menu with options like Move, Copy, Apply Template, Download, Add To Cart, and Delete. A yellow arrow points from the text 'Add files to cart' to the 'Add To Cart' option in this menu. In the center, there is a 'Cart List' table with two rows. The first row has columns for 'Name' and 'Absolute Path'. The second row has columns for 'Name' and 'Absolute Path'. The 'Name' column for the first row contains 'DataCommonsDesigns_01312018.pdf'. The 'Absolute Path' column for the first row contains '/commonsProdZone/home/conwaymc/DataCommonsDesigns_01312018.pdf'. The 'Name' column for the second row contains 'testupload'. The 'Absolute Path' column for the second row contains '/commonsProdZone/home/conwaymc/testupload'. A yellow arrow points from the text 'Apply publishing plugins to cart contents' to the 'testupload' entry in the cart list.

Name	Absolute Path
DataCommonsDesigns_01312018.pdf	/commonsProdZone/home/conwaymc/DataCommonsDesigns_01312018.pdf
testupload	/commonsProdZone/home/conwaymc/testupload

Add files to
cart

Apply publishing plugins to cart
contents

Publishing plugin API

- Like the search plugin, publishing plugins advertise themselves through an info endpoint.
- Publishers are simple but can accommodate complex operations through integrating with an external endpoint.
- Options:
 - Publish and **return a notification**
 - Publish and **download** the results to the client
 - Publish and **follow a link to an iRODS path** as a result
 - Publish and **follow a link to an external site** to complete the action

GEO publishing

- Our current production scenario is to publish sequencing data to GEO (Gene Expression Omnibus). Currently this gathers data from our iRODS based Commons and generates a spreadsheet
- The publishing plugin is set to a processing type of ‘download’ so a spreadsheet is generated by the plugin
- Next step is go generate a zip with the selected FastQ files assembled together with the spreadsheet
- Final step would be to add an intermediate curation service that would capture user data entry and tie back to original data

iRODS for managing research data via FAIR principles in a multi-repository world

WHERE DO I PUT MY RESEARCH DATA?

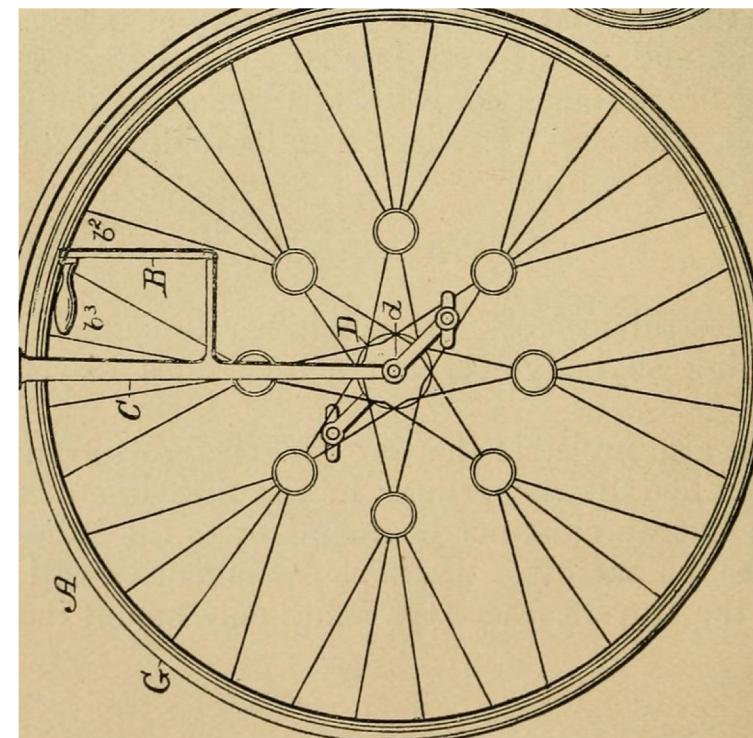


<https://flic.kr/p/8JLUEQ>

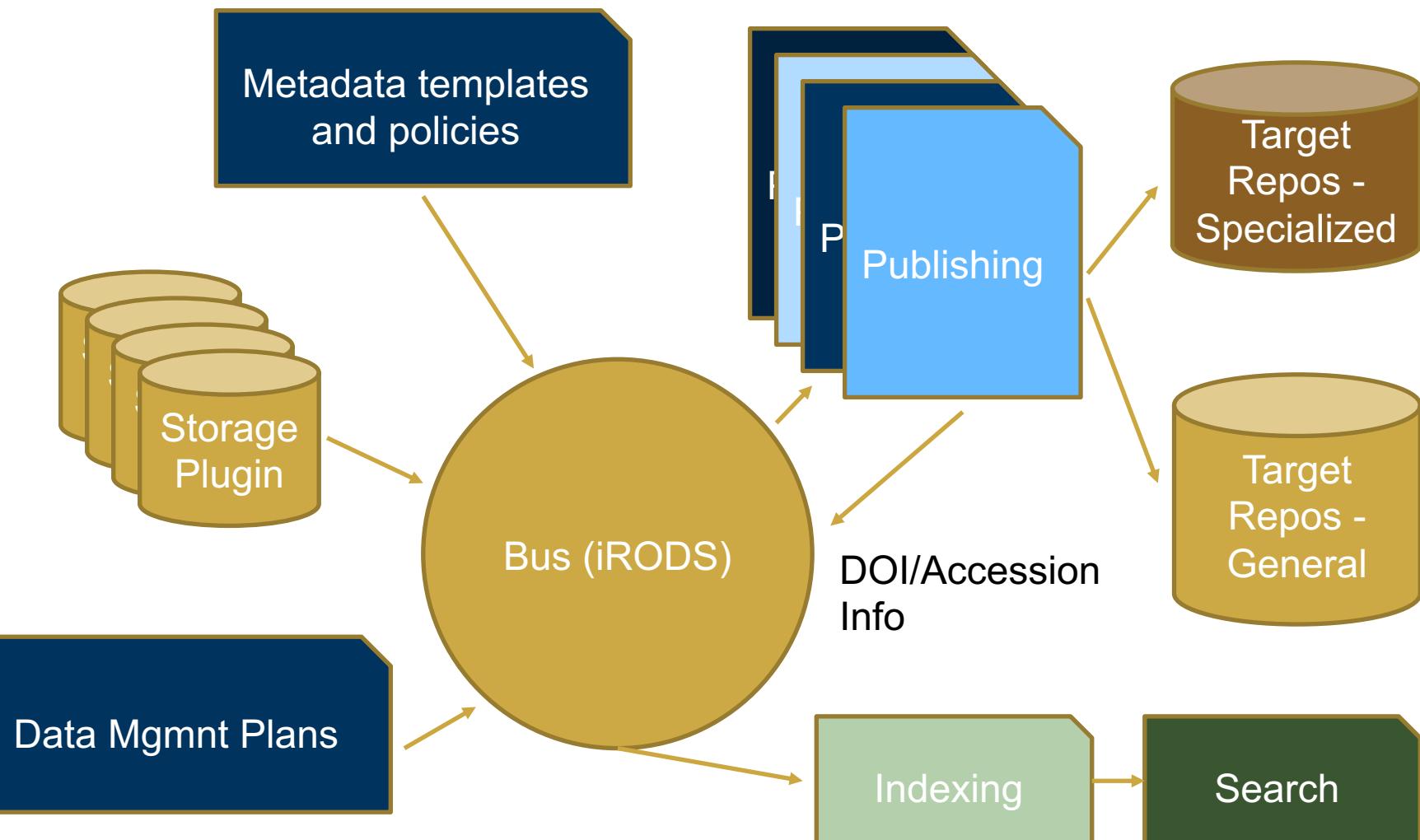
iRODS for preservation/archival storage with a twist

We need for a data submission ‘bus’ that can support:

- The assembly and curation of data sets and metadata models that are then pushed in segments to appropriate specialized repositories (e.g. dbGaP, GEO, SRA)
- The recording of DOIs and accession numbers via publishing
- Data model crosswalk to target repositories as part of publishing



Data flow through a data submission bus





Some useful references on this ‘bus’ concept

- COPO as a working example of a ‘bus’ (CyVerse UK) -
<https://f1000research.com/articles/9-495>
- Geoscience Digital Data Resource and Repository Service (GeoDaRRS) Workshop Report -
<https://opensky.ucar.edu/islandora/object/technotes:570>