Genomic data management at Canada's National Microbiology Laboratory with IRIDA and Galaxy

Aaron Petkau

Galaxy Admins Web Meetup

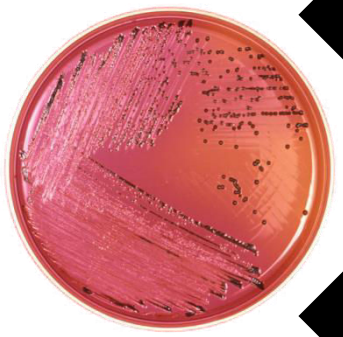
August 20, 2015

Background

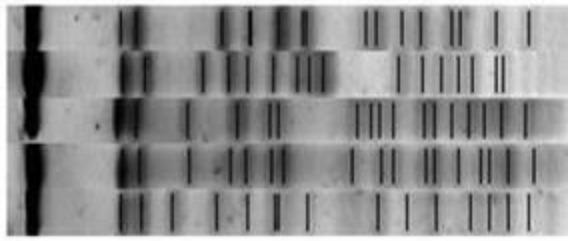
- Bioinformatics Core Facility at the National Microbiology Laboratory in Canada
  - Part of the Public Health Agency of Canada
  - Bacterial and viral genomics
  - Maintain a Galaxy instance + large compute cluster and shared storage
- Over the last few years moving to sequencing and comparing many (100s, 1000s) of genomes
  - Both for research and for outbreak surveillance and response

Outbreak Surveillance

Sample Source



Culture



PFGE Fingerprint

Whole Genome Sequencing Fingerprint

Current

Future

IRIDA

- Transition towards whole genome sequencing for outbreak surveillance and response
- Provides
  - Management of WGS data
  - Standardized pipelines for data analysis
  - Visualization of epidemiological data
  - REST API for integration with other software

http://irida.ca

Challenges

- Data Management
  - Keep sequenced genomes organized
  - Reduce copies of data
  - Accessible by users in Galaxy and other software
- Data Analysis
  - Standardized workflows
  - Workflows on 100s or 1000s of genomes

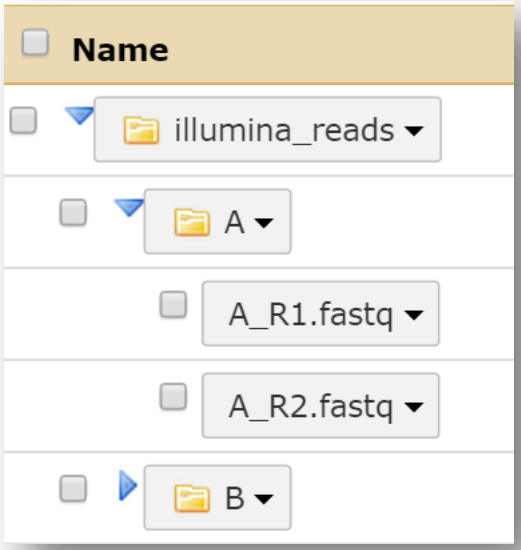
Initial Solutions

- Custom Galaxy API scripts
  - *sequences_imports.py, execute_workflow.py, extract_datasets.py*

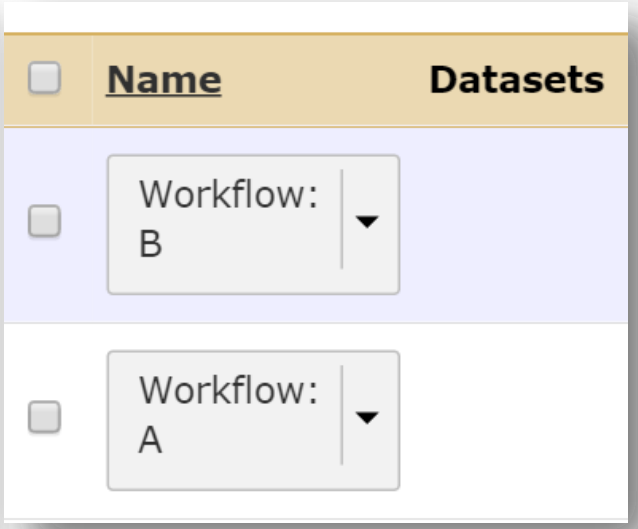
File System

```
Project_Test
├── Sample_A
│   ├── A_R1.fastq
│   └── A_R2.fastq
└── Sample_B
    ├── B_R1.fastq
    └── B_R2.fastq
```

Data Library



Run Multiple Workflows



Results: workflow_results.zip

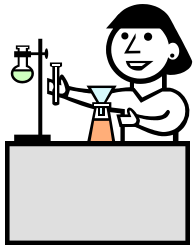
Problems

- Need Galaxy admin permissions
  - Users can't independently manage data
- Data duplication
  - No central management of sequencing data
  - Many copies of files for different software

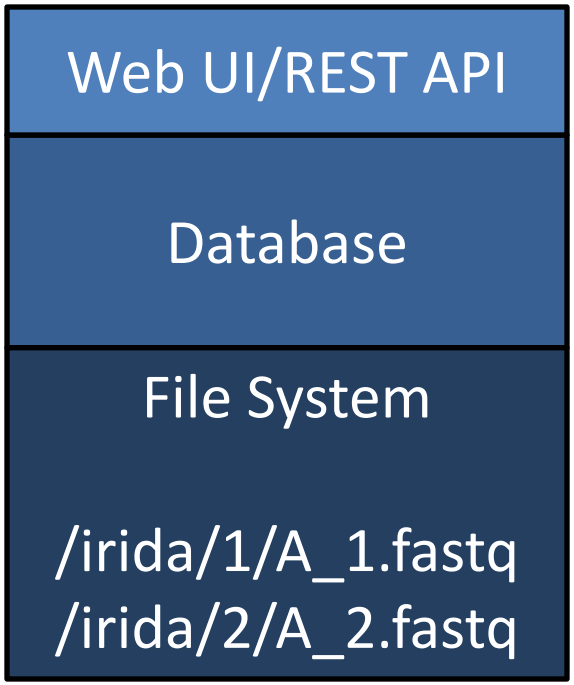
- Solutions became initial stages of IRIDA

IRIDA

- Initial development focused on management of WGS data



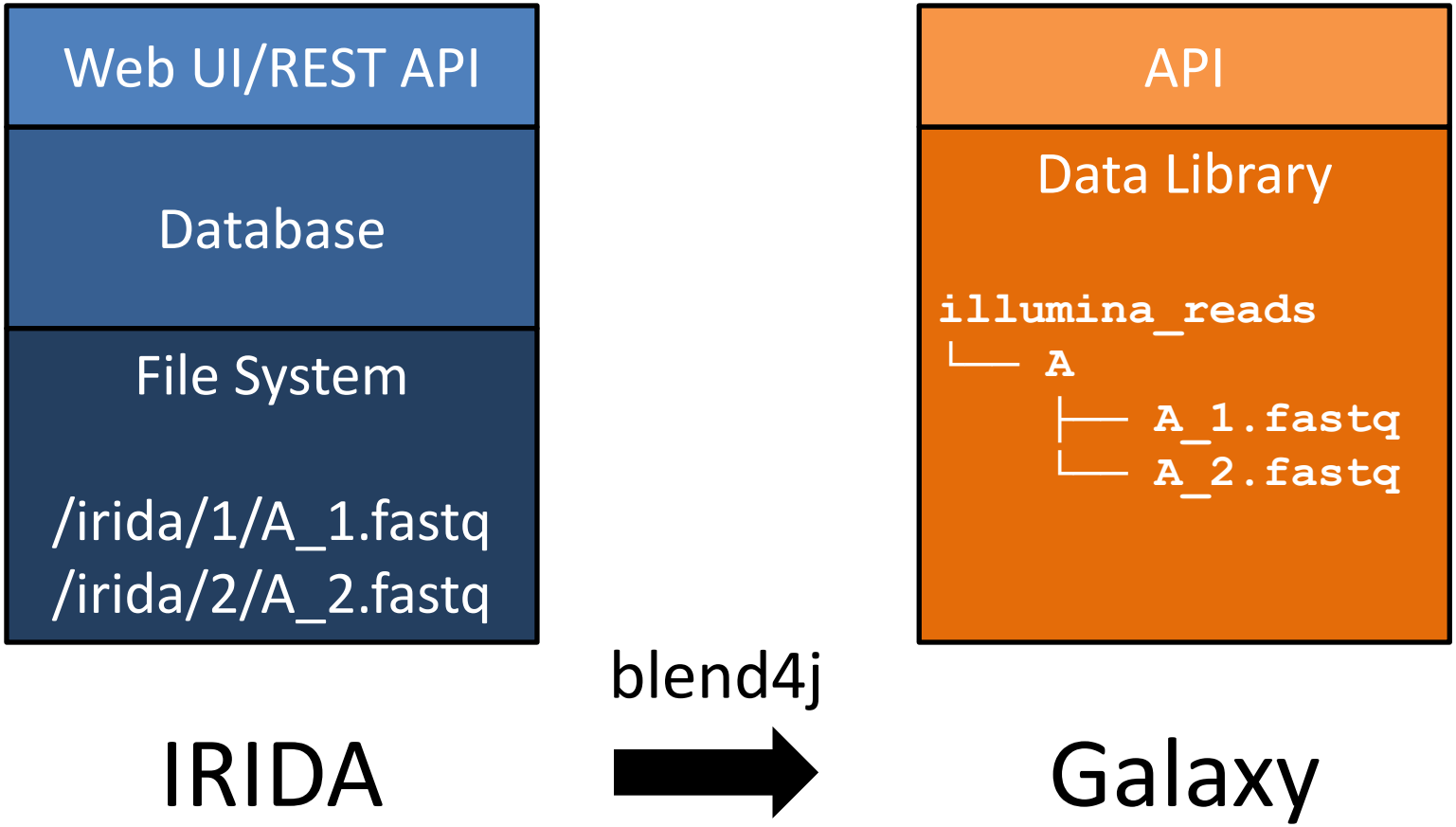
Researcher manages WGS data from Web UI



WGS uploaded via REST API

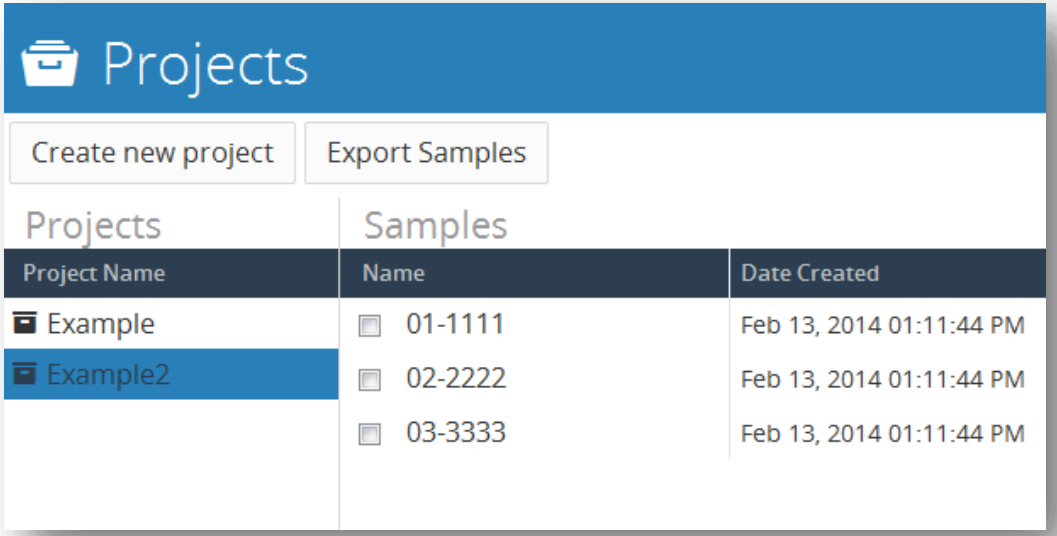
IRIDA/Galaxy: Solution 1

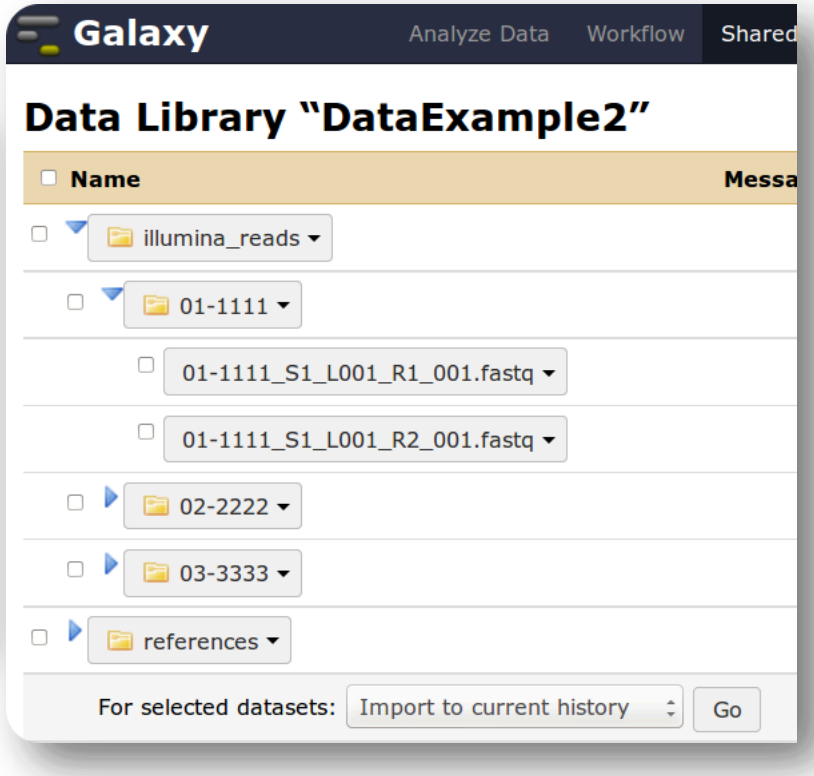
- Data pushed from IRIDA to Galaxy



IRIDA/Galaxy: Solution 1

- Data pushed from IRIDA to Galaxy





IRIDA → Galaxy

Issues

- IRIDA to Galaxy export code within IRIDA codebase
  - More difficult to test
  - Upgrades or bug fixes requires release of new IRIDA version
- Only send data to one Galaxy instance

Current IRIDA/Galaxy Import

- Pull data from IRIDA into Galaxy
- Modeled after other Galaxy **Data Source** tools
- Installable through a Toolshed
  - Will be released publically once IRIDA is released
  - Some configuration for authentication information

config.ini:

```
[Galaxy]
admin_key:  key
galaxy_url: http://galaxy

[IRIDA]
client_id:  galaxy-irida-importer
client_secret: secret
irida_url: http://irida
```

Web Browser

IRIDA Server

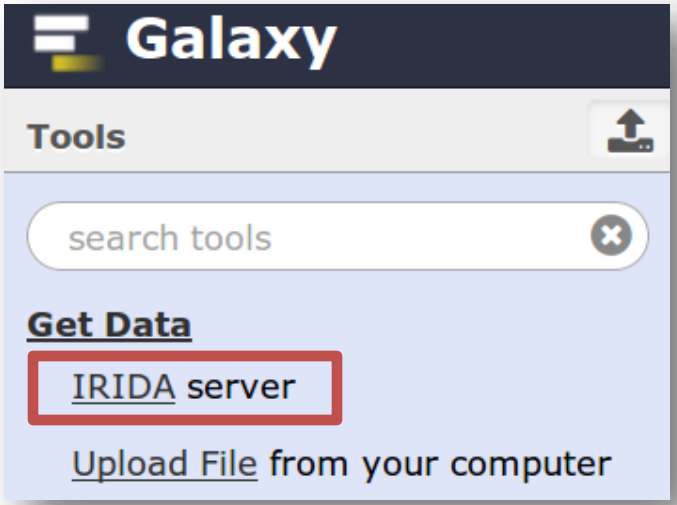
Galaxy Server

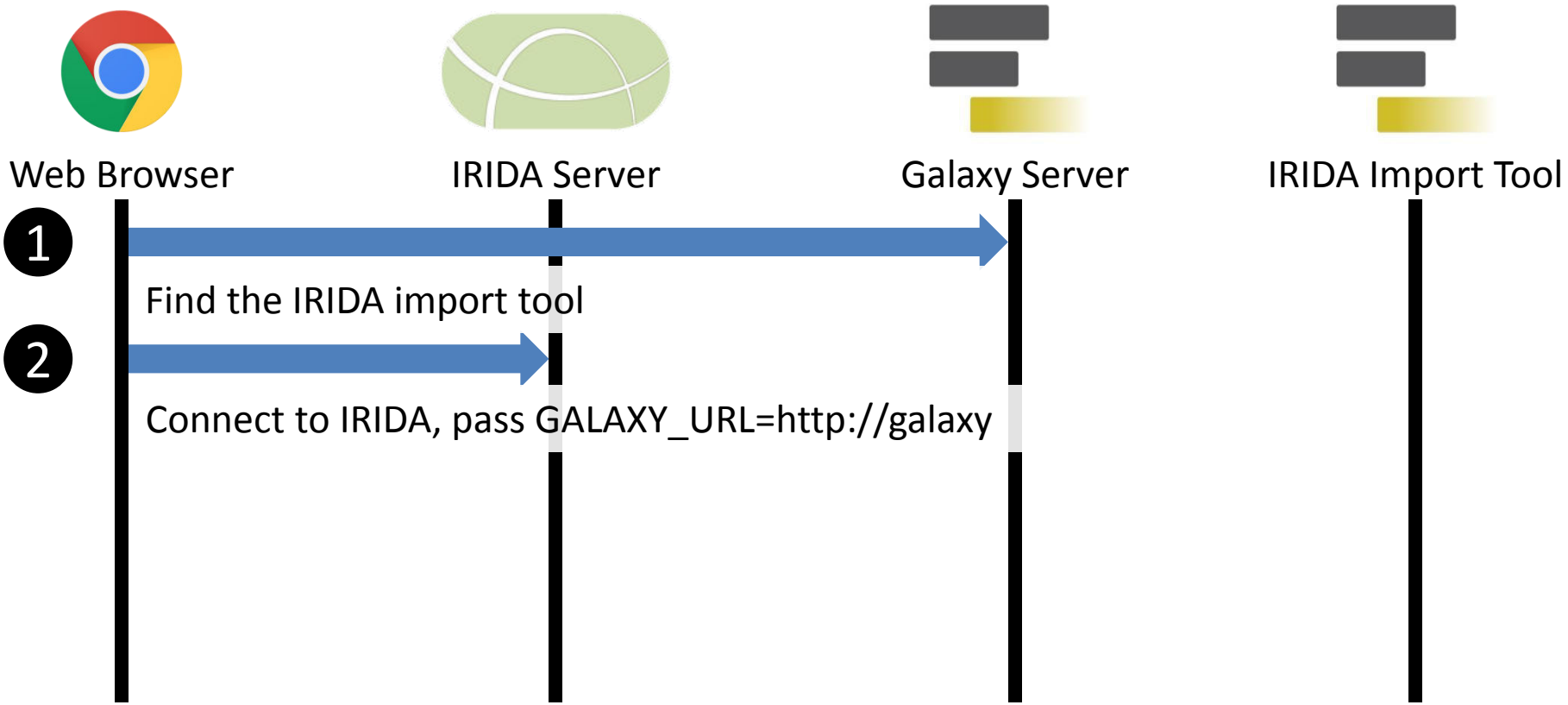
IRIDA Import Tool

1

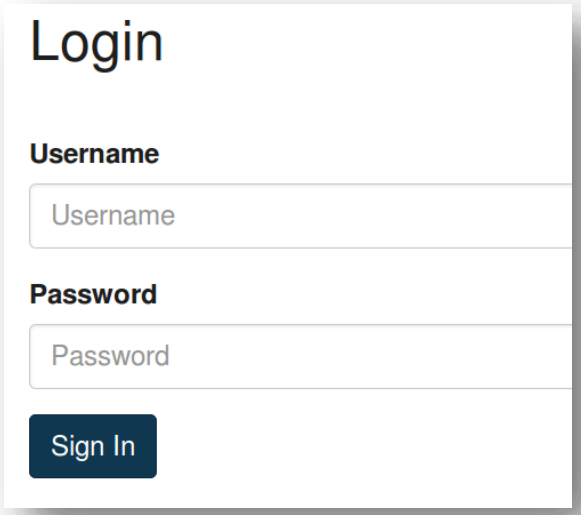
Find the IRIDA import tool

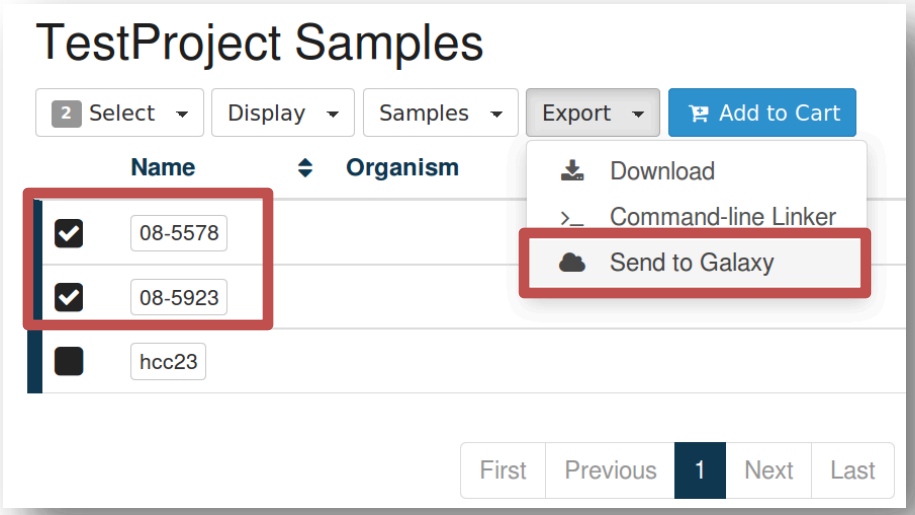
1



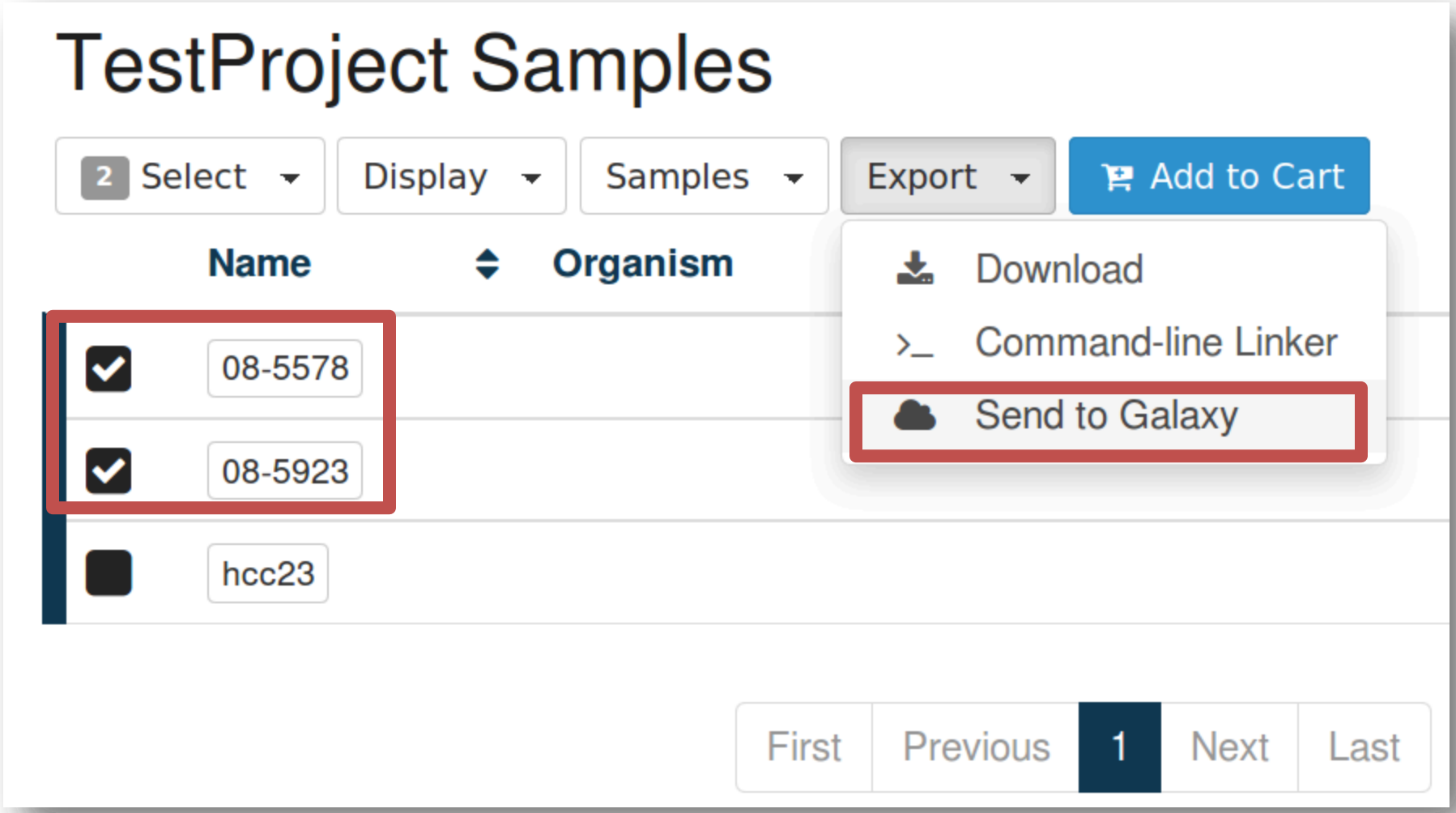

Web Browser
IRIDA Server
Galaxy Server
IRIDA Import Tool
1
Find the IRIDA import tool
2
Connect to IRIDA, pass GALAXY_URL=http://galaxy

2

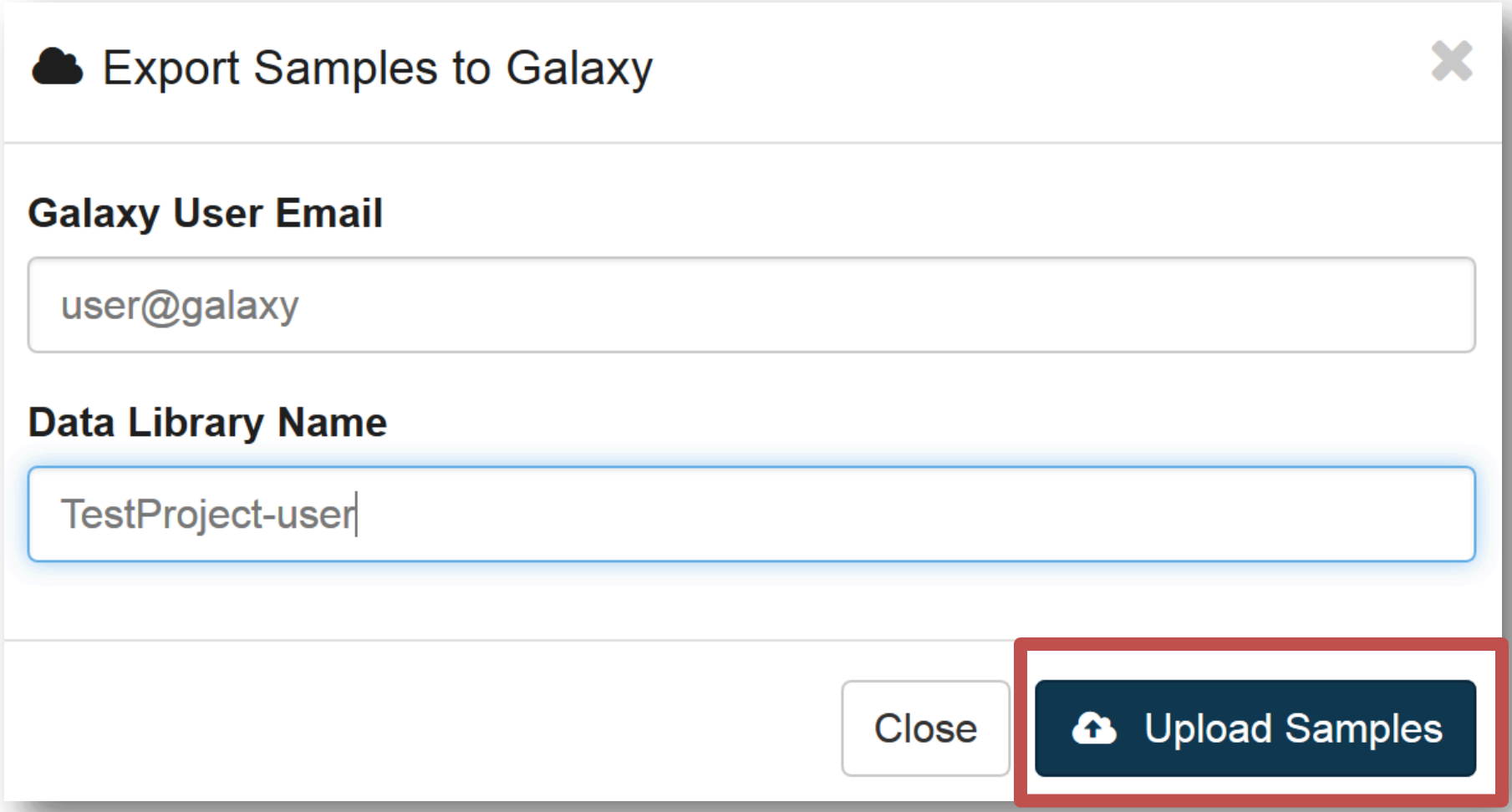

Login
Username
Username
Password
Password
Sign In

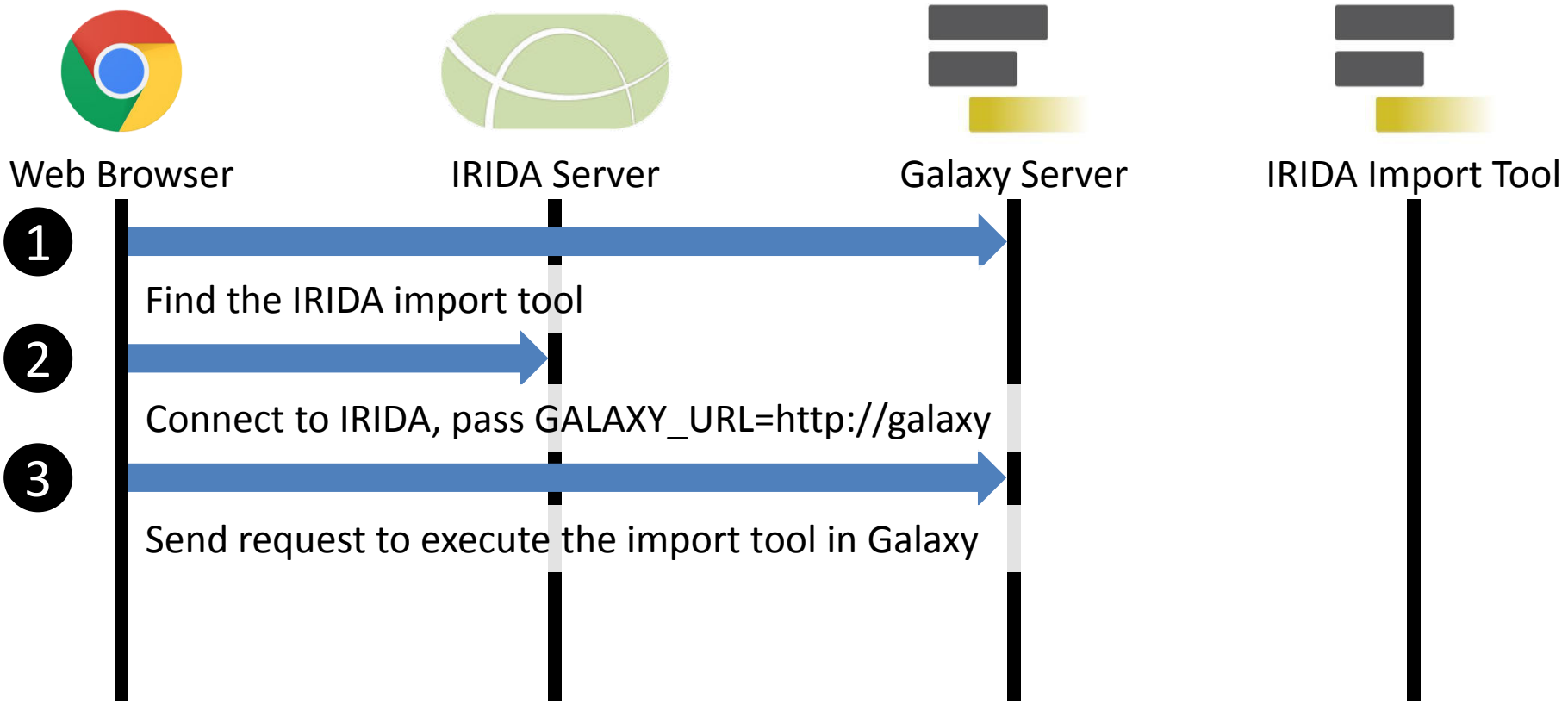

TestProject Samples
2 Select
Display
Samples
Export
Add to Cart
Name
Organism
Download
Command-line Linker
Send to Galaxy
08-5578
08-5923
hcc23
First
Previous
1
Next
Last

Upload Samples from IRIDA



Upload Samples from IRIDA

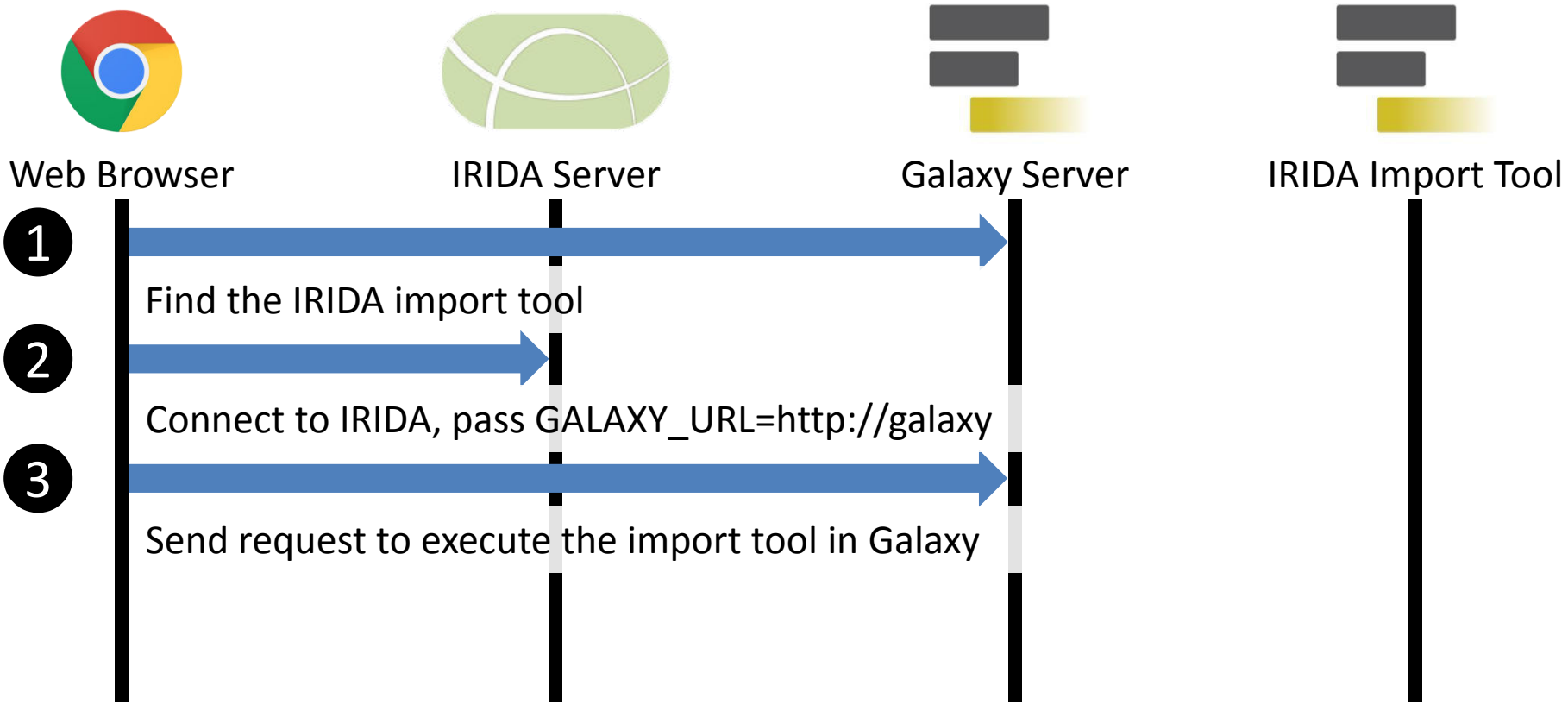




3

```
{
  “oauth2” : “7nXsHN”,
  “user”: “user@galaxy”,
  “library” : “TestProject-user”,
  “samples” : [
    {“name” : “08-5578”,
     “files”: [{“href” : http://irida/file/1}]
    }
  ]
}
```

Import details in JSON



3

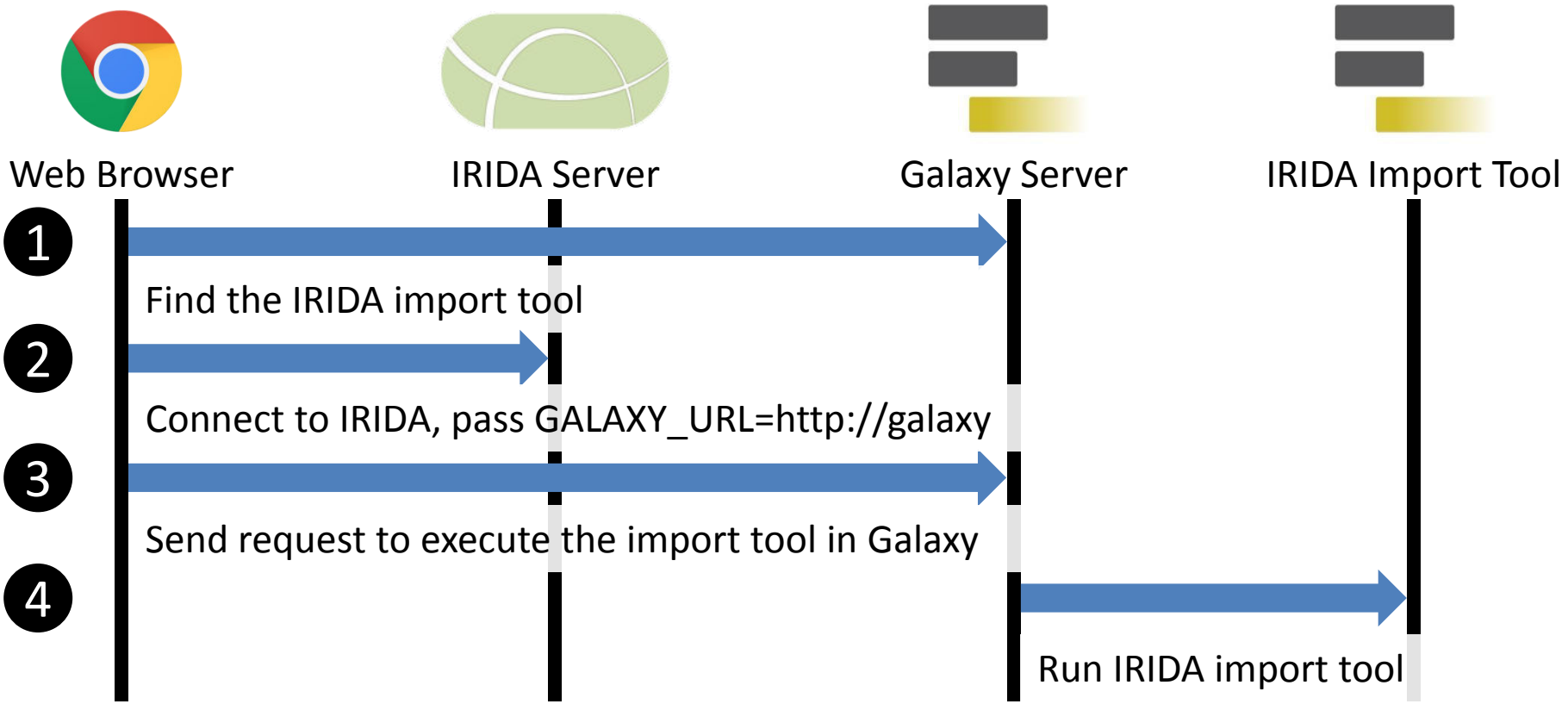
```
{
“oauth2” : “7nXsHN”,
“user”: “user@galaxy”,
“library” : “TestProject-user”,
“samples” : [
 {“name” : “08-5578”,
  “files”: [{“href” : http://irida/file/1}]
 }
]
}
```

IRIDA Authentication

Galaxy Library

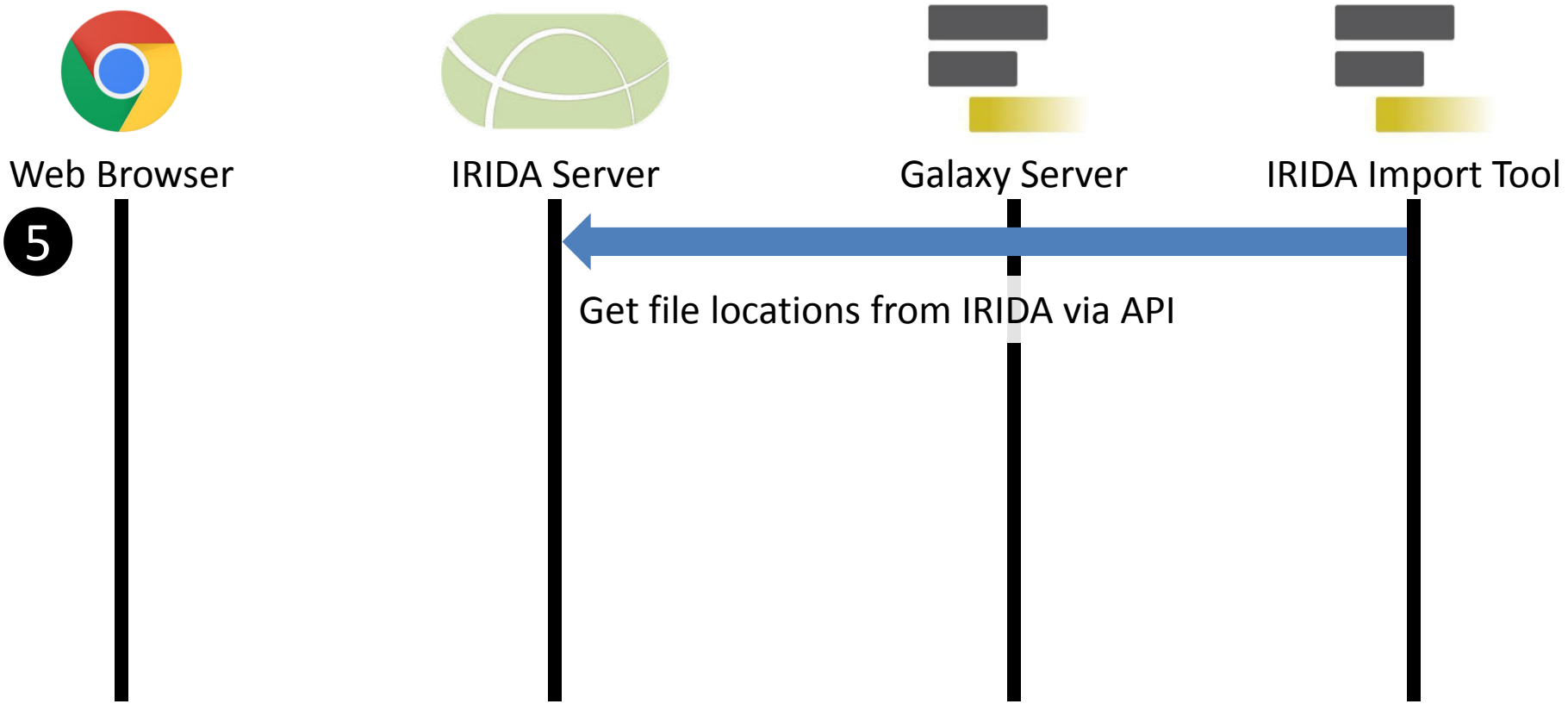
IRIDA Samples

Import details in JSON


Web Browser
IRIDA Server
Galaxy Server
IRIDA Import Tool
1
Find the IRIDA import tool
2
Connect to IRIDA, pass GALAXY_URL=http://galaxy
3
Send request to execute the import tool in Galaxy
4
Run IRIDA import tool

4

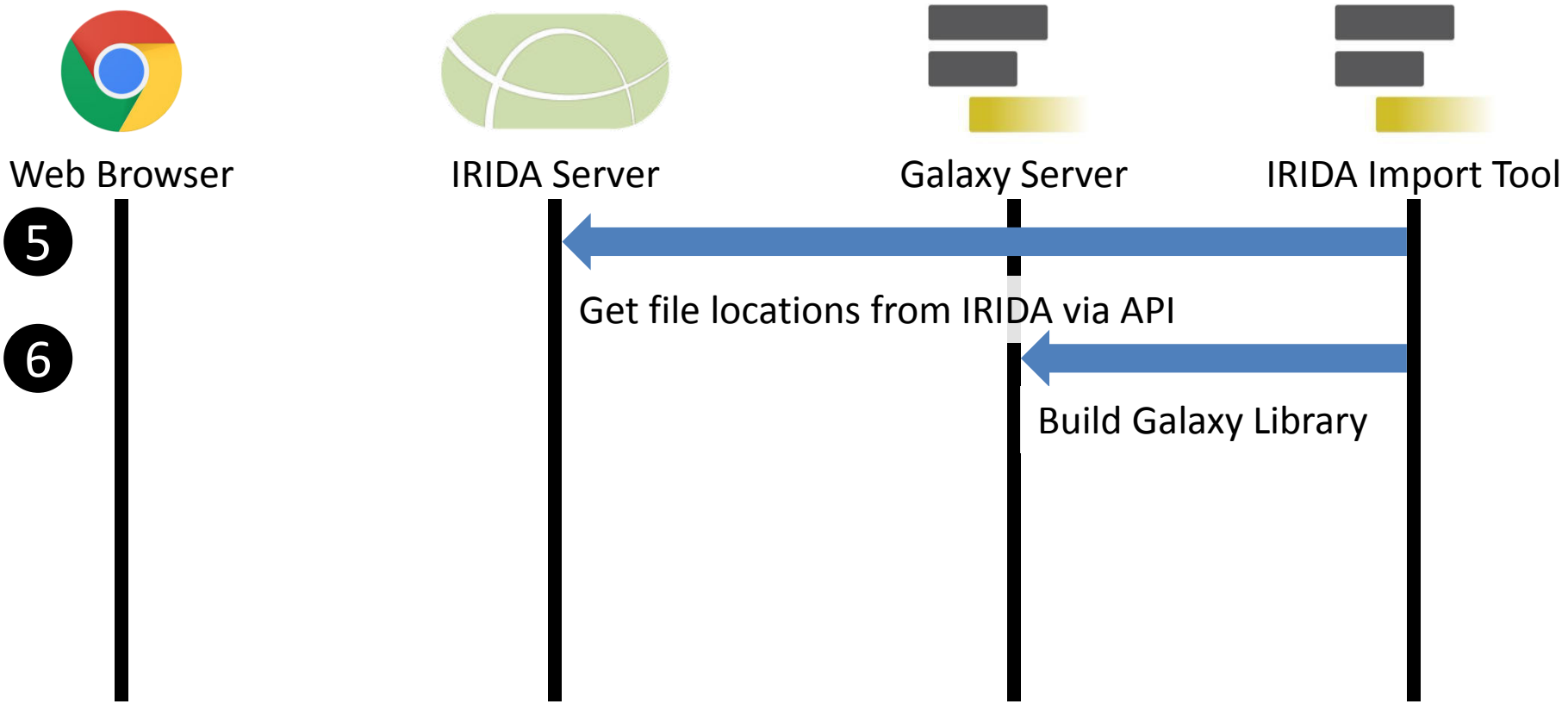

Unnamed history
1 shown, 1 hidden
0 bytes
1: IRIDA Export
This job is currently running



5 **File Locations**

/irida/1/08-5578_R1.fastq

/irida/2/08-5578_R2.fastq

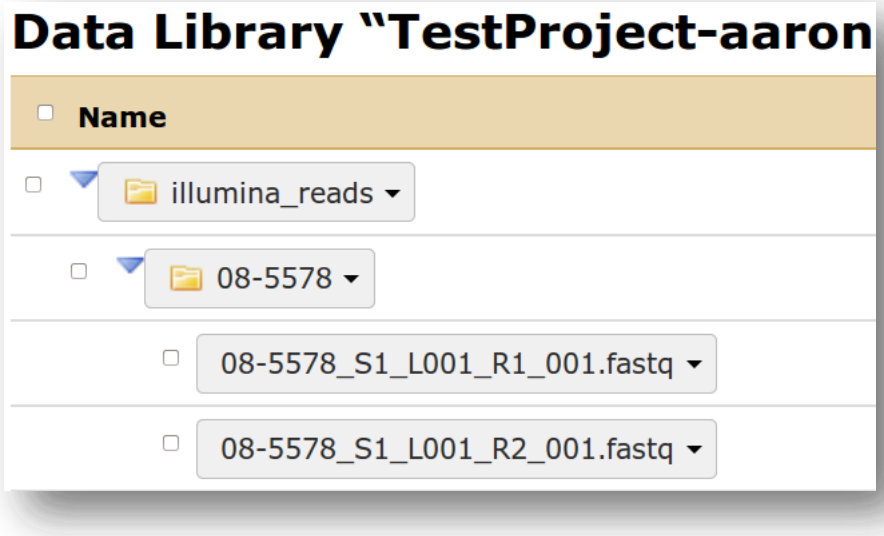


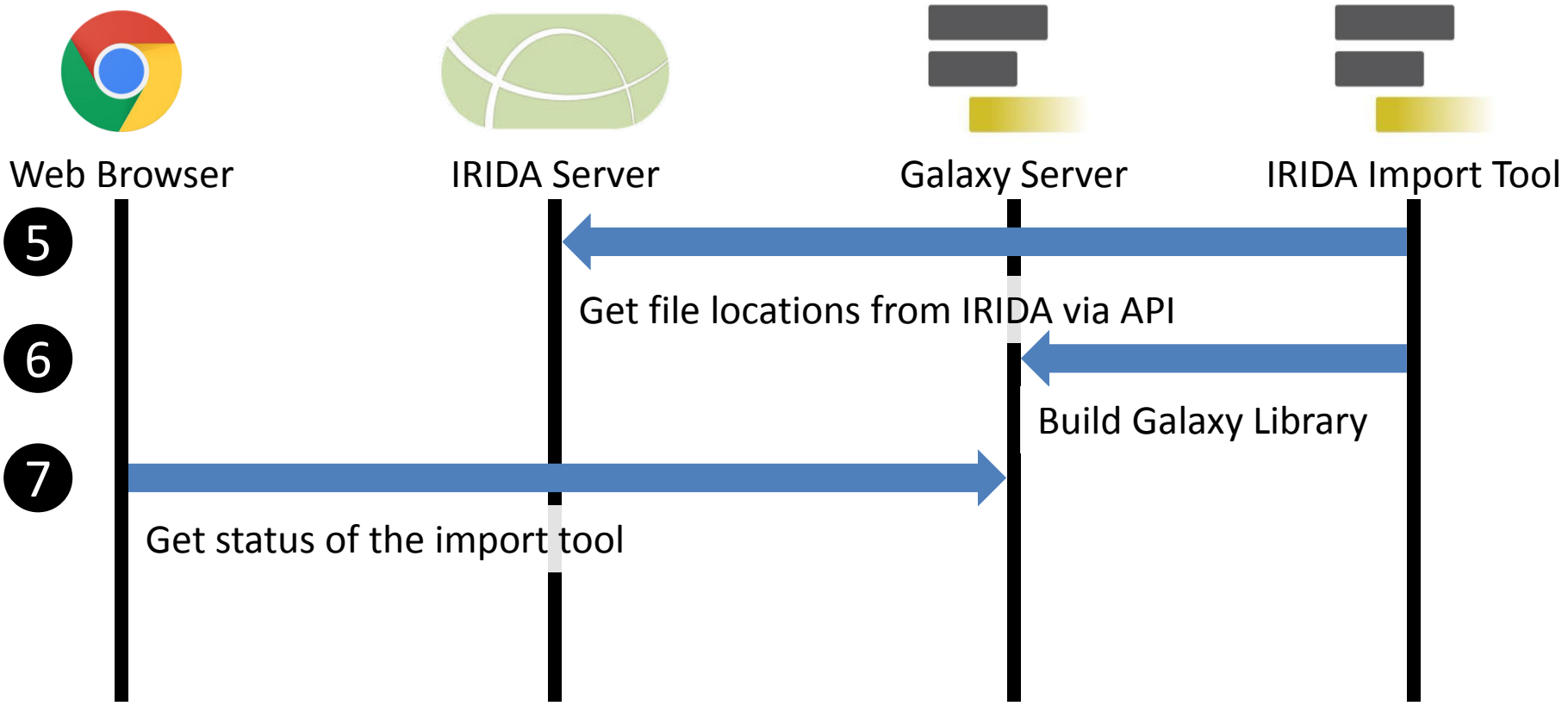
5 **File Locations**

/irida/1/08-5578_R1.fastq

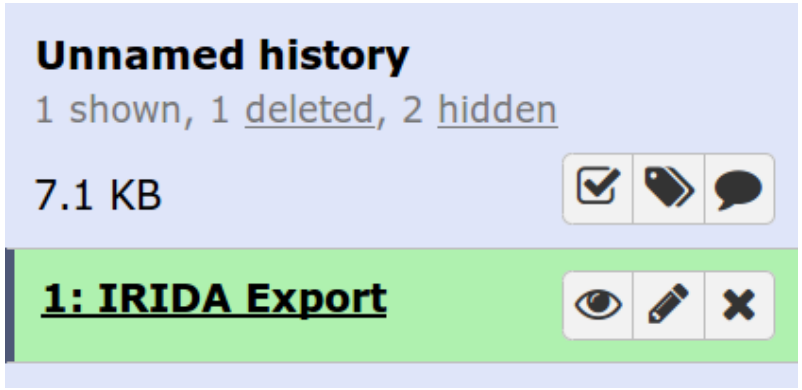
/irida/2/08-5578_R2.fastq

6





7



INFO: Exporting files from IRIDA to Galaxy...
…
INFO: Final summary:
2 file(s) exported and 0 file(s) skipped.

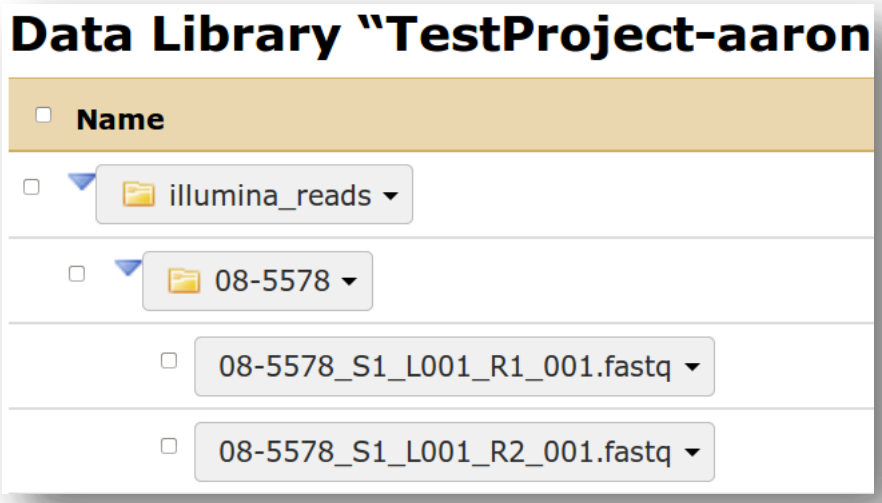
Status Report

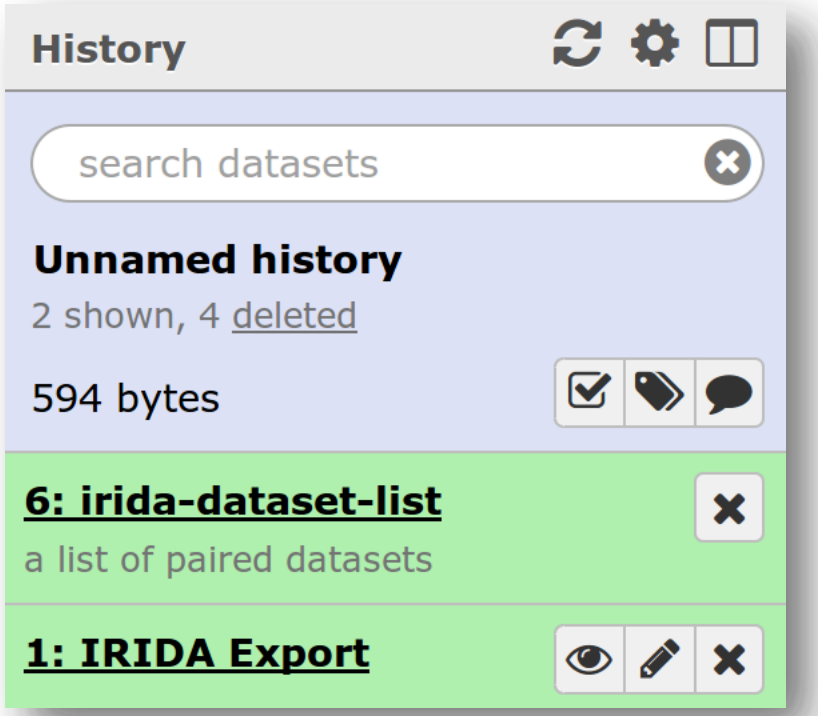
Summary

- Two challenges
- Data Management
  - Addressed by IRIDA and IRIDA/Galaxy import tool
- Data Analysis
  - Addressed by dataset collections enabling many genomes in the same workflow

Future Work

- Automated construction of dataset collections in a user's history





Acknowledgements

- Philip Mabon and Eric Enns
  - Galaxy gurus
- Joel Thiessen
  - Wrote the IRIDA import tool
- Daniel Bouchard
  - Updates to import tool
- Franklin Bristow, Josh Adam, Thomas Matthews
  - IRIDA and Galaxy tool development
- Gary Van Domselaar
  - Chief Bioinformatics
- IRIDA Team
- Galaxy Team