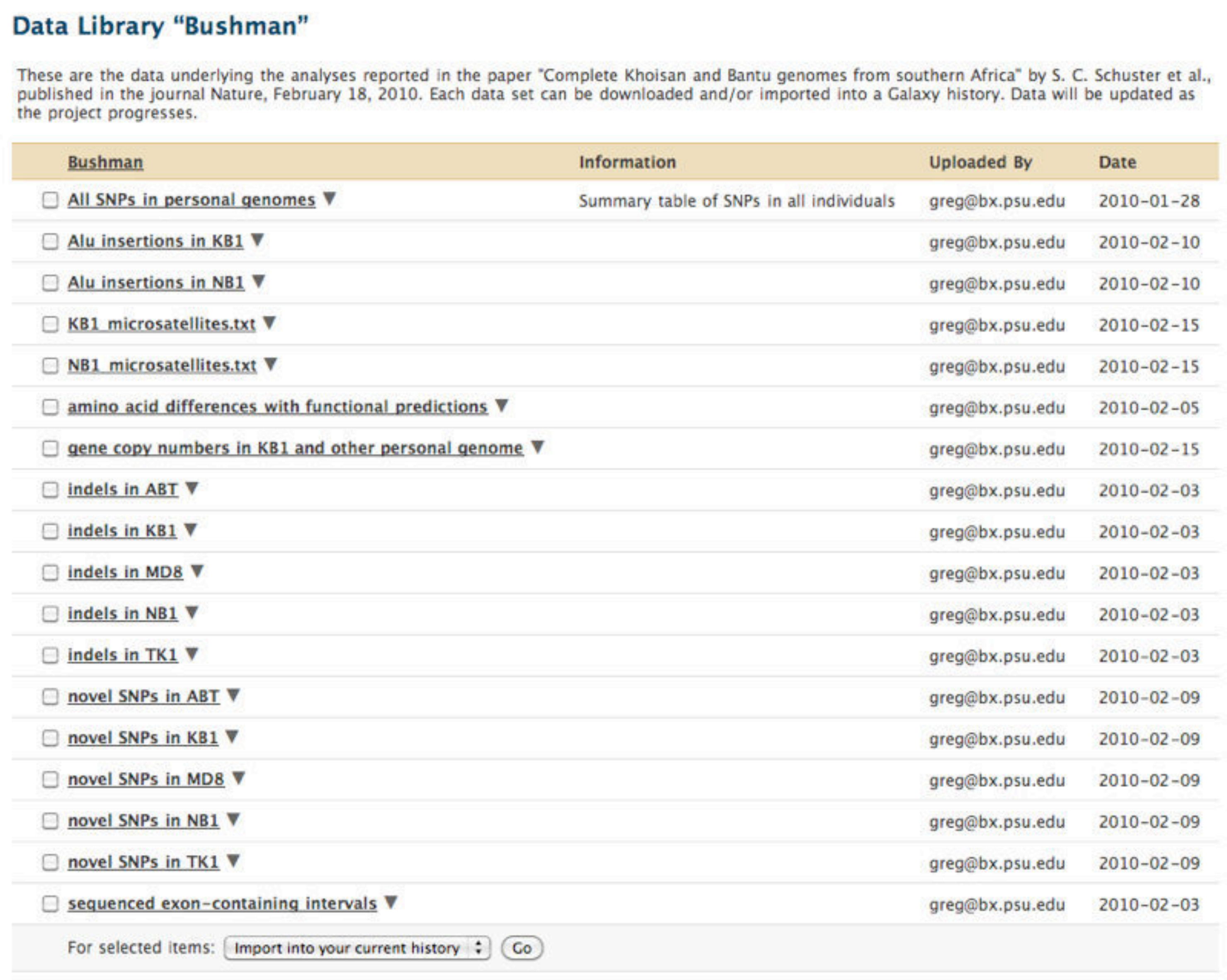
# Galaxy data libraries and sample tracking at NGS facilities

Greg Von Kuster
Galaxy Team

# What are data libraries?

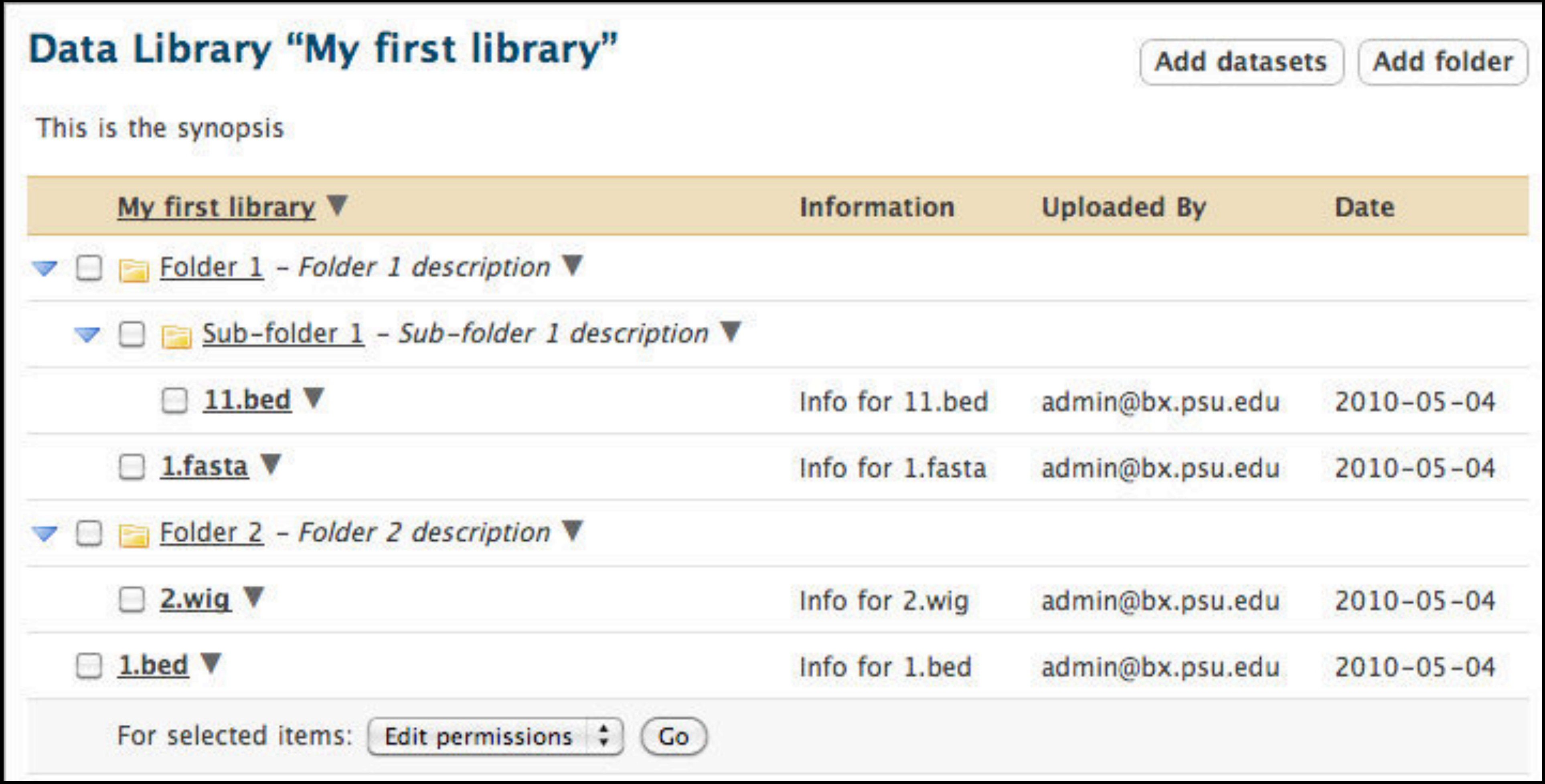


## Data Library “Bushman”

These are the data underlying the analyses reported in the paper "Complete Khoisan and Bantu genomes from southern Africa" by S. C. Schuster et al., published in the journal Nature, February 18, 2010. Each data set can be downloaded and/or imported into a Galaxy history. Data will be updated as the project progresses.

| Bushman | Information | Uploaded By | Date |
|---|---|---|---|
| □ All SNPs in personal genomes ▼ | Summary table of SNPs in all individuals | greg@bx.psu.edu | 2010-01-28 |
| □ Alu insertions in KB1 ▼ | | greg@bx.psu.edu | 2010-02-10 |
| □ Alu insertions in NB1 ▼ | | greg@bx.psu.edu | 2010-02-10 |
| □ KB1 microsatellites.txt ▼ | | greg@bx.psu.edu | 2010-02-15 |
| □ NB1 microsatellites.txt ▼ | | greg@bx.psu.edu | 2010-02-15 |
| □ amino acid differences with functional predictions ▼ | | greg@bx.psu.edu | 2010-02-05 |
| □ gene copy numbers in KB1 and other personal genome ▼ | | greg@bx.psu.edu | 2010-02-15 |
| □ indels in ABT ▼ | | greg@bx.psu.edu | 2010-02-03 |
| □ indels in KB1 ▼ | | greg@bx.psu.edu | 2010-02-03 |
| □ indels in MD8 ▼ | | greg@bx.psu.edu | 2010-02-03 |
| □ indels in NB1 ▼ | | greg@bx.psu.edu | 2010-02-03 |
| □ indels in TK1 ▼ | | greg@bx.psu.edu | 2010-02-03 |
| □ novel SNPs in ABT ▼ | | greg@bx.psu.edu | 2010-02-09 |
| □ novel SNPs in KB1 ▼ | | greg@bx.psu.edu | 2010-02-09 |
| □ novel SNPs in MD8 ▼ | | greg@bx.psu.edu | 2010-02-09 |
| □ novel SNPs in NB1 ▼ | | greg@bx.psu.edu | 2010-02-09 |
| □ novel SNPs in TK1 ▼ | | greg@bx.psu.edu | 2010-02-09 |
| □ sequenced exon-containing intervals ▼ | | greg@bx.psu.edu | 2010-02-03 |

# a hierarchical container for datasets



## Data Library “My first library”

This is the synopsis

| My first library ▼ | Information | Uploaded By | Date |
|---|---|---|---|
| ▽ □ Folder 1 – *Folder 1 description* ▼ | | | |
| ▽ □ Sub-folder 1 – *Sub-folder 1 description* ▼ | | | |
| □ 11.bed ▼ | Info for 11.bed | admin@bx.psu.edu | 2010-05-04 |
| □ 1.fasta ▼ | Info for 1.fasta | admin@bx.psu.edu | 2010-05-04 |
| ▽ □ Folder 2 – *Folder 2 description* ▼ | | | |
| □ 2.wig ▼ | Info for 2.wig | admin@bx.psu.edu | 2010-05-04 |
| □ 1.bed ▼ | Info for 1.bed | admin@bx.psu.edu | 2010-05-04 |

For selected items: Edit permissions Go

# How do I put data in?

# How do I put data in?

- Upload a single file
- Import datasets from a Galaxy history
- Upload a directory of files
- Upload files from file system paths

# uploading a directory

- The directory location is configurable
- Includes an option to not copy files into Galaxy's normal files directory, leaving them in their original location

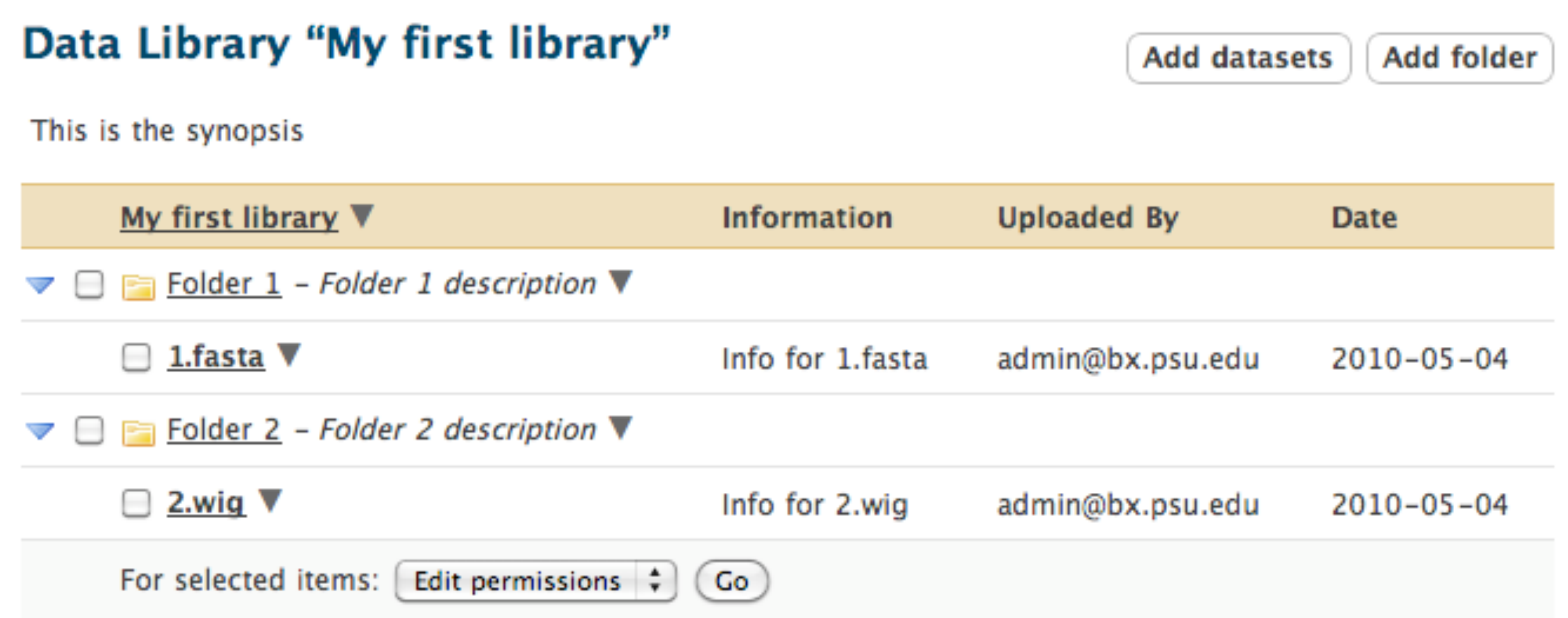
# uploading from file system paths

- Allows for any number of file system paths (files or directories), saving the directory structure if desired
- Includes an option to not copy files into Galaxy's normal files directory, leaving them in their original location

Oops, I uploaded the wrong data, what can I do?

# delete it!

- Folders and datasets can be deleted at any level (state is saved for contents, if any)
- Deleted items are not displayed



Data Library "My first library"

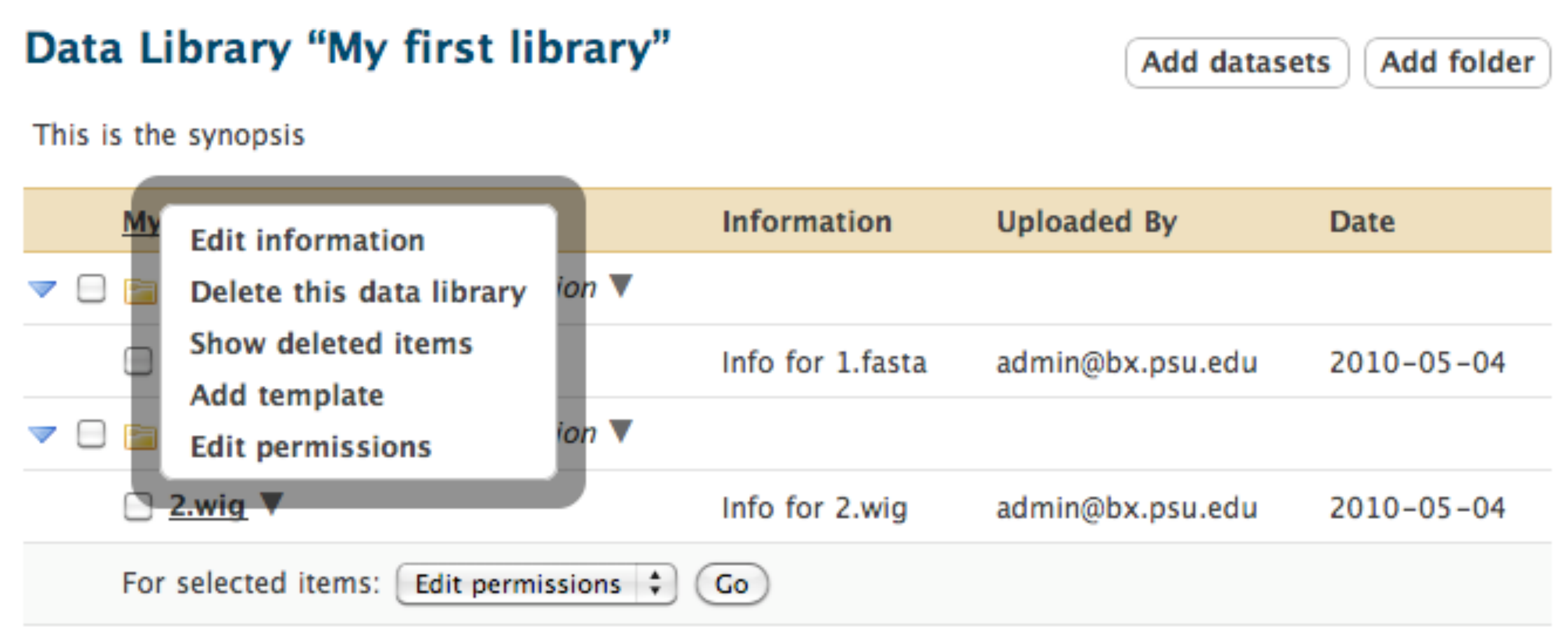
This is the synopsis

| | | My first library ▼ | Information | Uploaded By | Date |
|---|---|---|---|---|---|
| ▼ | □ | Folder 1 – *Folder 1 description* ▼ | | | |
| | | □ 1.fasta ▼ | Info for 1.fasta | admin@bx.psu.edu | 2010-05-04 |
| ▼ | □ | Folder 2 – *Folder 2 description* ▼ | | | |
| | | □ 2.wig ▼ | Info for 2.wig | admin@bx.psu.edu | 2010-05-04 |

For selected items: Edit permissions Go

# Can I undelete?

# yes!  Show deleted items



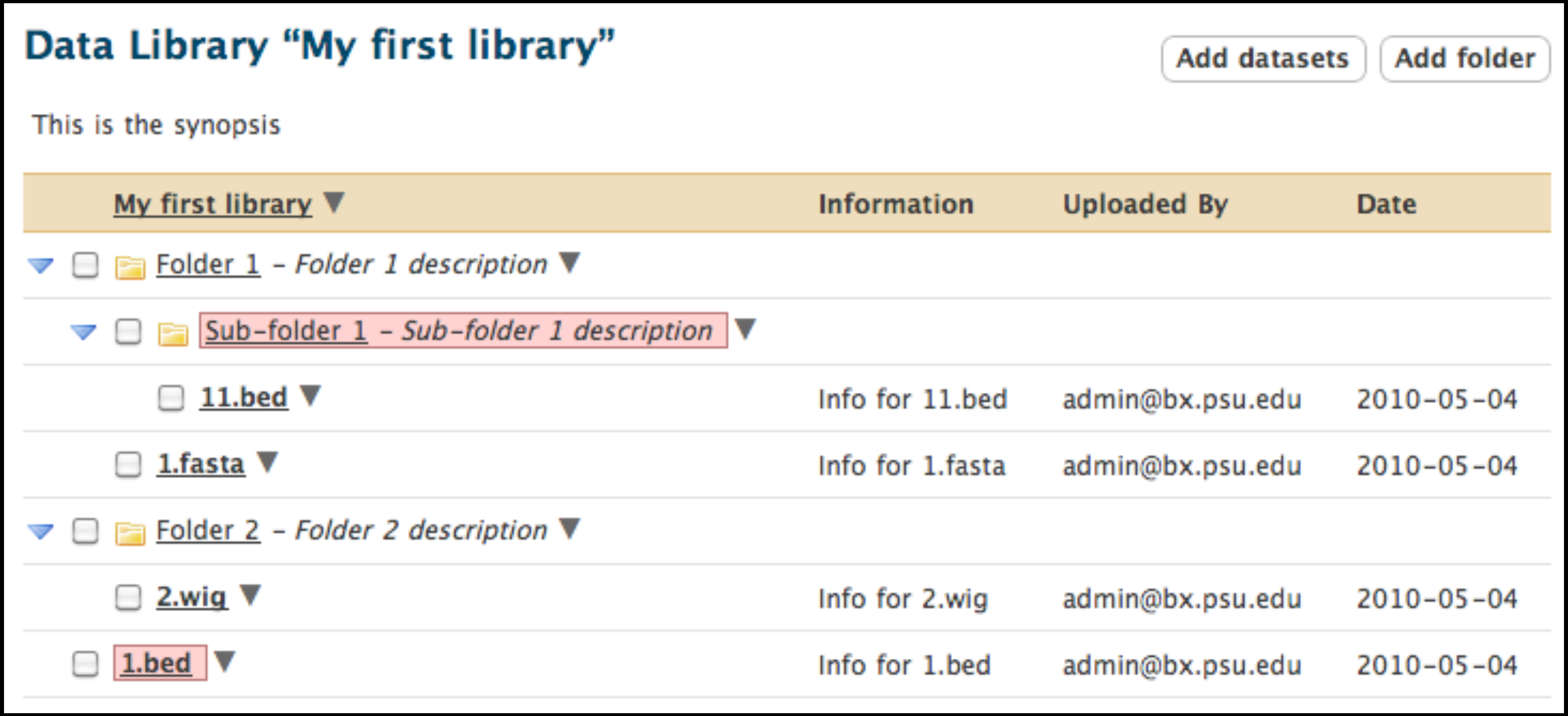
Data Library "My first library"

This is the synopsis

| My | Information | Uploaded By | Date |
|---|---|---|---|
| ☐ ion ▼ | | | |
| ☐ | Info for 1.fasta | admin@bx.psu.edu | 2010-05-04 |
| ☐ ion ▼ | | | |
| ☐ 2.wig ▼ | Info for 2.wig | admin@bx.psu.edu | 2010-05-04 |

For selected items: Edit permissions

# deleted items are red

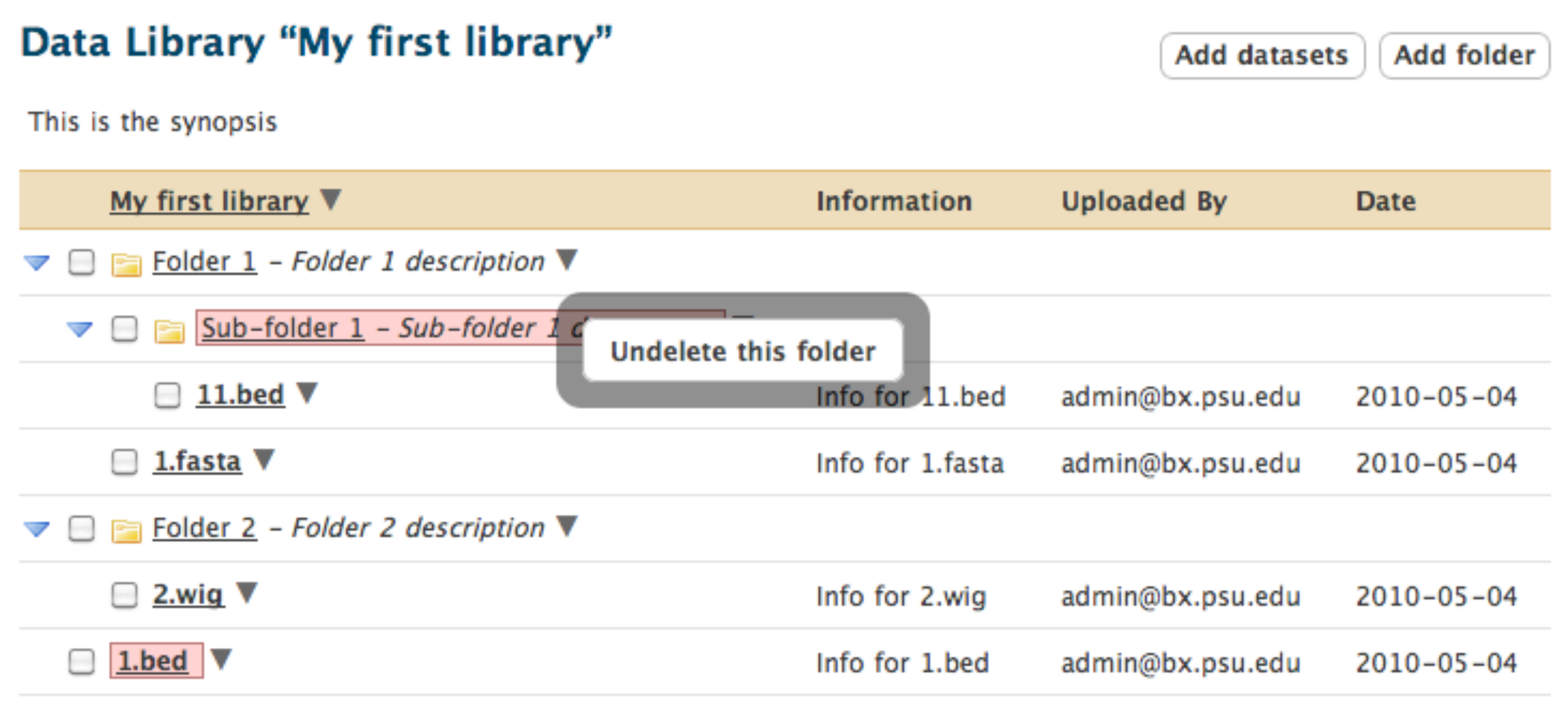


## Data Library “My first library”

This is the synopsis

| My first library ▼ | Information | Uploaded By | Date |
|---|---|---|---|
| □ Folder 1 – *Folder 1 description* ▼ | | | |
| □ Sub-folder 1 – *Sub-folder 1 description* ▼ | | | |
| □ 11.bed ▼ | Info for 11.bed | admin@bx.psu.edu | 2010-05-04 |
| □ 1.fasta ▼ | Info for 1.fasta | admin@bx.psu.edu | 2010-05-04 |
| □ Folder 2 – *Folder 2 description* ▼ | | | |
| □ 2.wig ▼ | Info for 2.wig | admin@bx.psu.edu | 2010-05-04 |
| □ 1.bed ▼ | Info for 1.bed | admin@bx.psu.edu | 2010-05-04 |

# ...and you can undelete



## Data Library "My first library"

This is the synopsis

| | My first library ▼ | Information | Uploaded By | Date |
|---|---|---|---|---|
| □ | Folder 1 – *Folder 1 description* ▼ | | | |
| □ | Sub-folder 1 – *Sub-folder 1 d* | | | |
| □ | 11.bed ▼ | Info for 11.bed | admin@bx.psu.edu | 2010-05-04 |
| □ | 1.fasta ▼ | Info for 1.fasta | admin@bx.psu.edu | 2010-05-04 |
| □ | Folder 2 – *Folder 2 description* ▼ | | | |
| □ | 2.wig ▼ | Info for 2.wig | admin@bx.psu.edu | 2010-05-04 |
| □ | 1.bed ▼ | Info for 1.bed | admin@bx.psu.edu | 2010-05-04 |

How do I use the data in a library?

# using the data

- Users that can access a library dataset can import it into their Galaxy history for analysis

# using the data

- Users that can access a library dataset can import it into their Galaxy history for analysis
- Importing a library dataset into a history creates a pointer to the same single disk file, minimizing disk space

# using the data

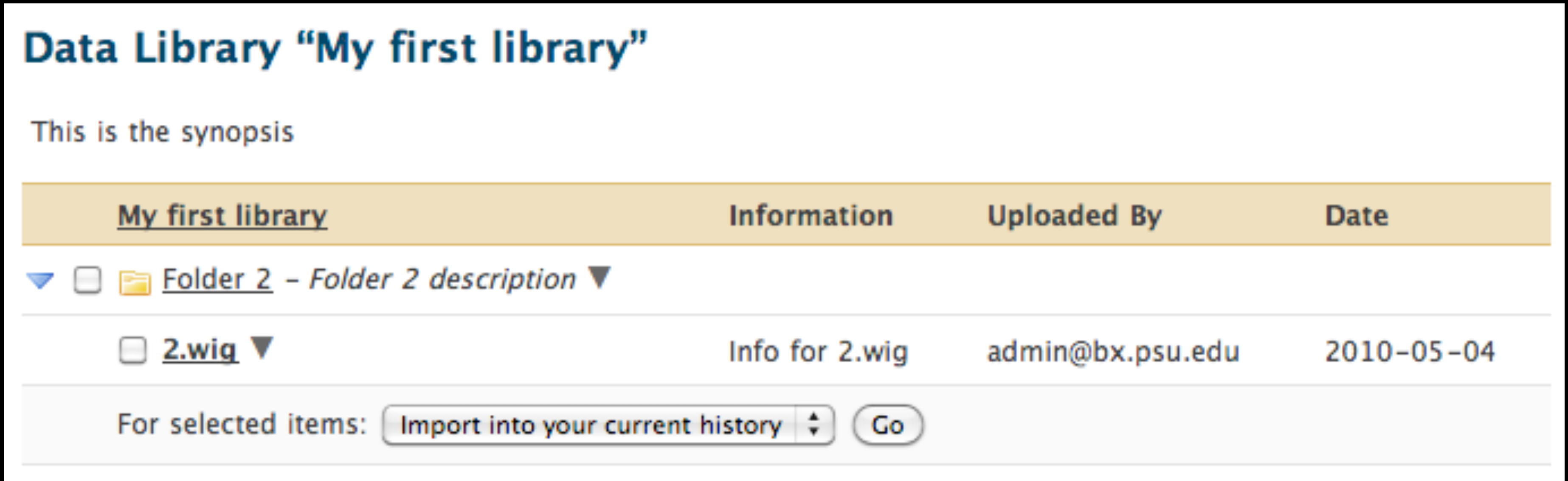
- Users that can access a library dataset can import it into their Galaxy history for analysis
- Importing a library dataset into a history creates a pointer to the same single disk file, minimizing disk space
- Versioning is supported for library datasets

# Can I protect the data?

# data library security

- Restricts access to the entire library or specific datasets contained within it
- The default is no restriction, making items "public"

# so Dick can see...



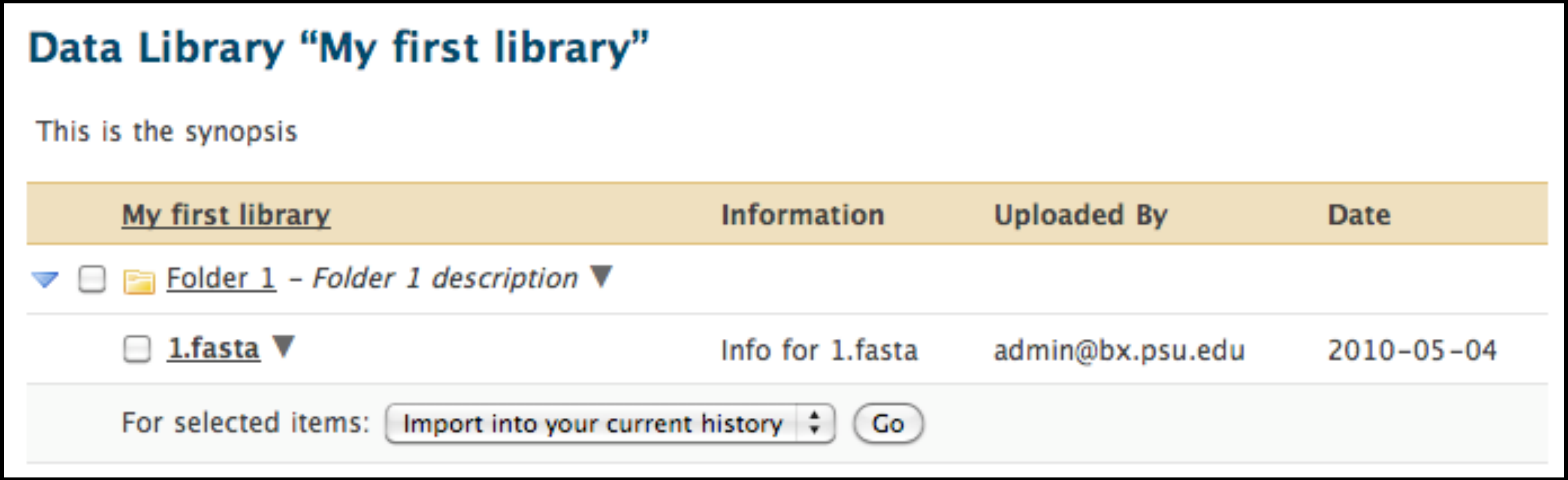
## Data Library "My first library"

This is the synopsis

| My first library | Information | Uploaded By | Date |
|---|---|---|---|
| □ Folder 2 – *Folder 2 description* ▼ | | | |
| □ 2.wig ▼ | Info for 2.wig | admin@bx.psu.edu | 2010-05-04 |

For selected items: Import into your current history  Go

# ...but Jane can see



## Data Library "My first library"

This is the synopsis

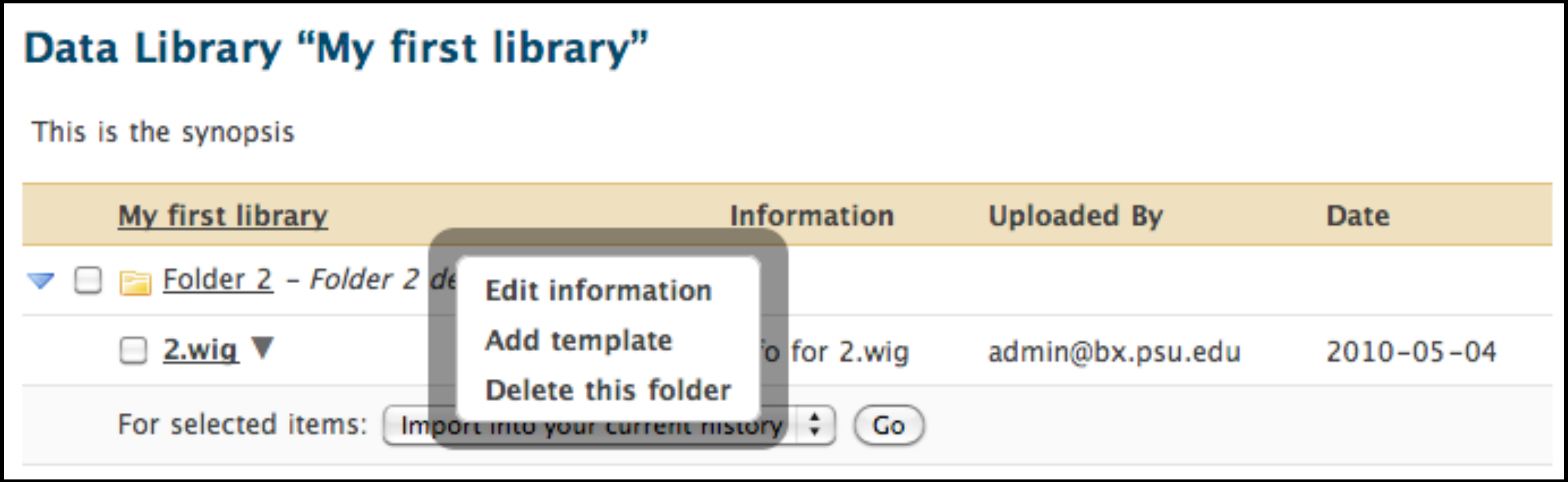
| My first library | Information | Uploaded By | Date |
|---|---|---|---|
| □ Folder 1 – *Folder 1 description* ▼ | | | |
| □ 1.fasta ▼ | Info for 1.fasta | admin@bx.psu.edu | 2010-05-04 |

For selected items: Import into your current history

# data library security also...

- Grants permission to specific users to perform actions on library items
- The default is no permission to do anything (except access)

# so Dick can...



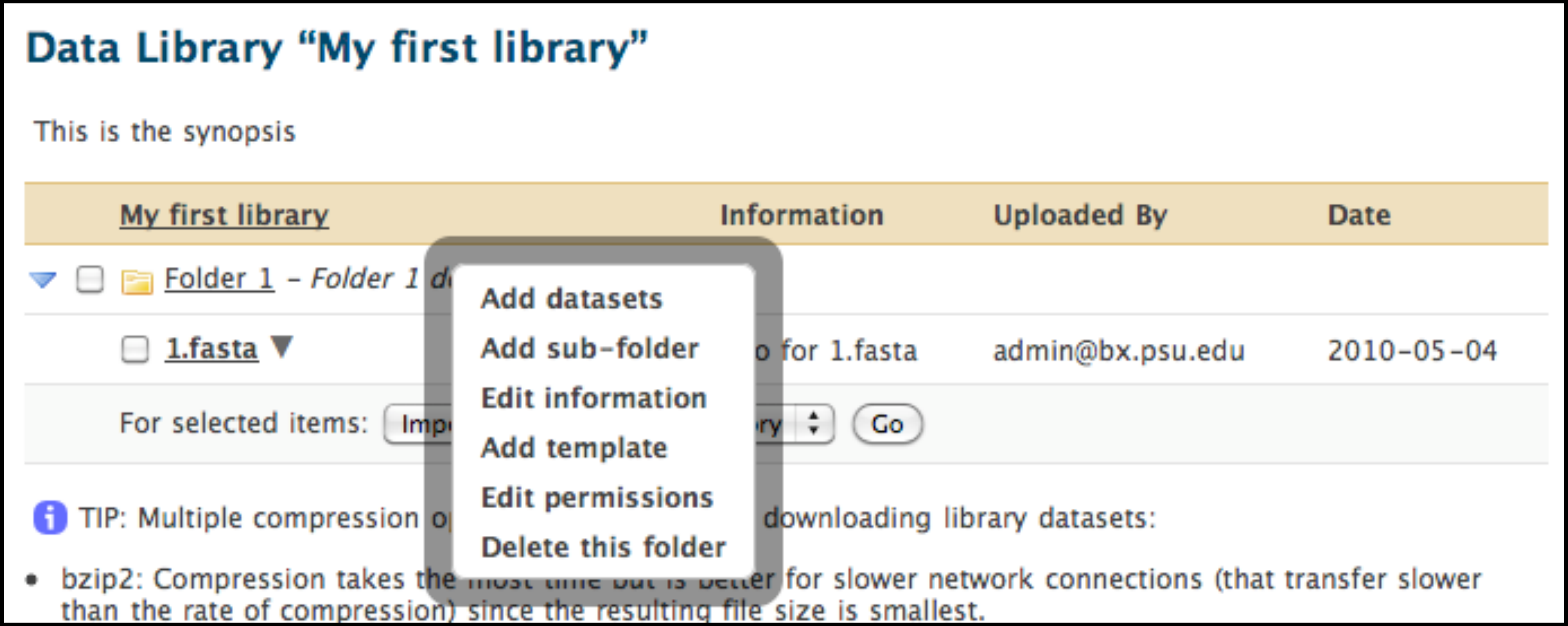
Data Library “My first library”

This is the synopsis

| | My first library | Information | Uploaded By | Date |
|---|---|---|---|---|
| ▽ □ | Folder 2 – *Folder 2 de* | | | |
| | □ 2.wig ▼ | o for 2.wig | admin@bx.psu.edu | 2010-05-04 |

For selected items: Import into your current history ⇕ Go

# but Jane can...



## Data Library “My first library”

This is the synopsis

| | My first library | Information | Uploaded By | Date |
|---|---|---|---|---|
| ▼ □ | Folder 1 - *Folder 1 d* | | | |
| □ | 1.fasta ▼ | o for 1.fasta | admin@bx.psu.edu | 2010-05-04 |

For selected items: Imp ry Go

TIP: Multiple compression o downloading library datasets:

- bzip2: Compression takes the most time but is better for slower network connections (that transfer slower than the rate of compression) since the resulting file size is smallest.

How does this work?

# flexible security policies

# flexible security policies

- Use built-in role-based authorization

# flexible security policies

- Use built-in role-based authorization
- Library level: access library, add, modify, manage permissions

# flexible security policies

- Use built-in role-based authorization
- Library level: access library, add, modify, manage permissions
- Folder level: add, modify, manage permissions

# flexible security policies

- Use built-in role-based authorization
- Library level: access library, add, modify, manage permissions
- Folder level: add, modify, manage permissions
- Datasets: access, modify, manage permissions

# and...

- Security settings are automatically inherited downward in the hierarchy, but can be overridden, providing distinct security policies at any level in the hierarchy

What’s the difference between the “library access” permission on a library and the “access” permission on a dataset?

# subtle, but important

- If a library is public but a contained dataset is not, anyone can see the library, but not everyone will see the dataset

# subtle, but important

- If a library is public but a contained dataset is not, anyone can see the library, but not everyone will see the dataset
- If a library is restricted, but a contained dataset is public, only those that can see the library will have access to the dataset

# subtle, but important

- If a library is public but a contained dataset is not, anyone can see the library, but not everyone will see the dataset
- If a library is restricted, but a contained dataset is public, only those that can see the library will have access to the dataset
- Users having any role associated with “library access” can see the library, but users must have all roles associated with “access” on the dataset in order to see the dataset

# data library templates

# data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)

# data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
- Template layout can easily be defined or changed at any time

# data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
- Template layout can easily be defined or changed at any time
- Templates can be inherited downward, but can be overridden at any level

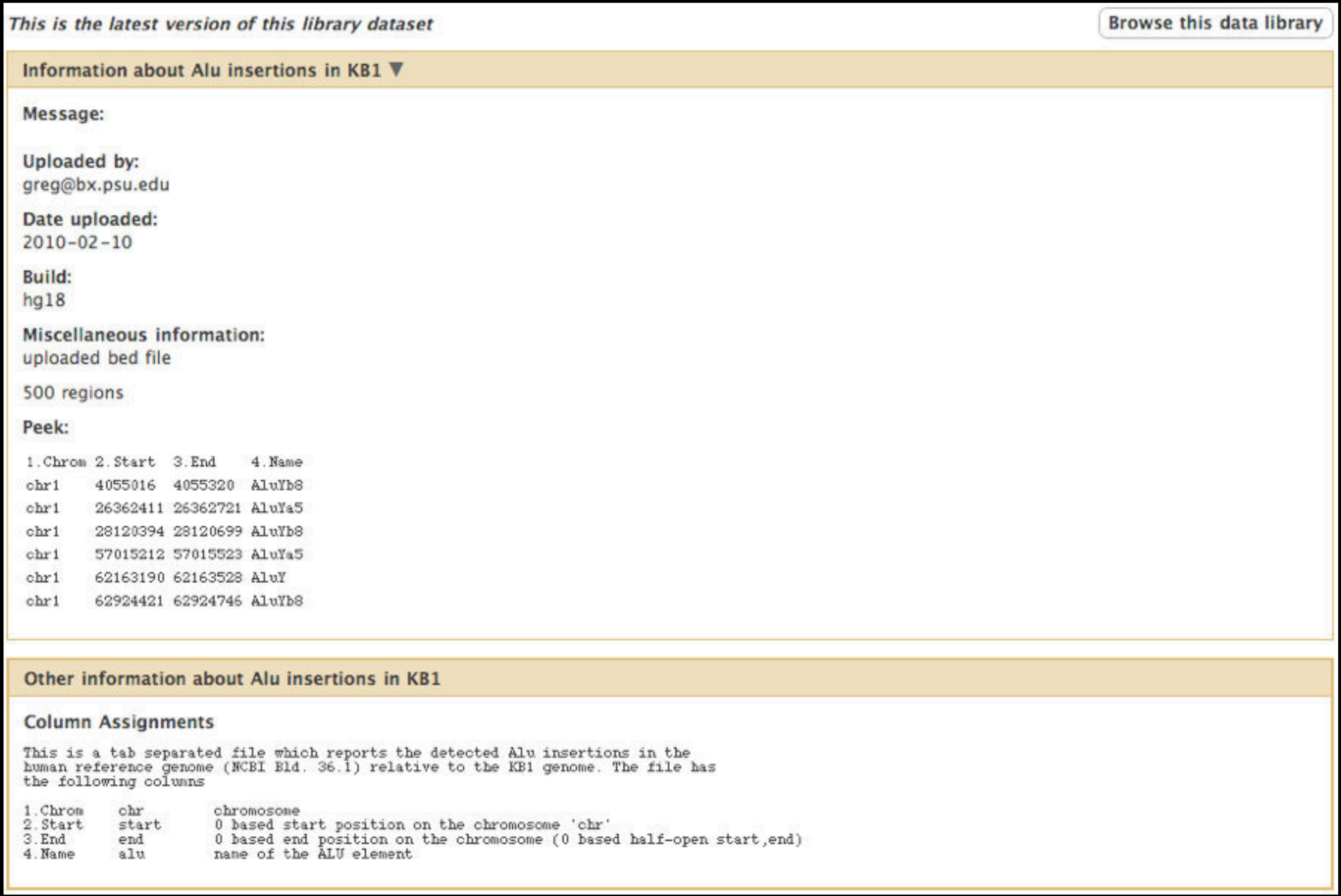
# data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
- Template layout can easily be defined or changed at any time
- Templates can be inherited downward, but can be overridden at any level
- Template inheritance can be turned on or off at any level

# data library templates

- Associate information with a library and its contents (e.g., associate sequencing run parameters with the resulting dataset)
- Template layout can easily be defined or changed at any time
- Templates can be inherited downward, but can be overridden at any level
- Template inheritance can be turned on or off at any level
- Templates can be deleted from any item

# a dataset template



*This is the latest version of this library dataset*

**Information about Alu insertions in KB1 ▼**

**Message:**

**Uploaded by:**
greg@bx.psu.edu

**Date uploaded:**
2010-02-10

**Build:**
hg18

**Miscellaneous information:**
uploaded bed file

500 regions

**Peek:**

| 1.Chrom | 2.Start | 3.End | 4.Name |
|---|---|---|---|
| chr1 | 4055016 | 4055320 | AluYb8 |
| chr1 | 26362411 | 26362721 | AluYa5 |
| chr1 | 28120394 | 28120699 | AluYb8 |
| chr1 | 57015212 | 57015523 | AluYa5 |
| chr1 | 62163190 | 62163528 | AluY |
| chr1 | 62924421 | 62924746 | AluYb8 |

**Other information about Alu insertions in KB1**

**Column Assignments**

```
This is a tab separated file which reports the detected Alu insertions in the
human reference genome (NCBI Bld. 36.1) relative to the KB1 genome. The file has
the following columns

1.Chrom    chr        chromosome
2.Start    start      0 based start position on the chromosome 'chr'
3.End      end        0 based end position on the chromosome (0 based half-open start,end)
4.Name     alu        name of the ALU element
```

Galaxy sample tracking streamlines the delivery of data from sequencing runs to customers how?

# example facility process

- A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results

# example facility process

- A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results
- The facility bar codes the samples, scanning the tubes at each station in the lab

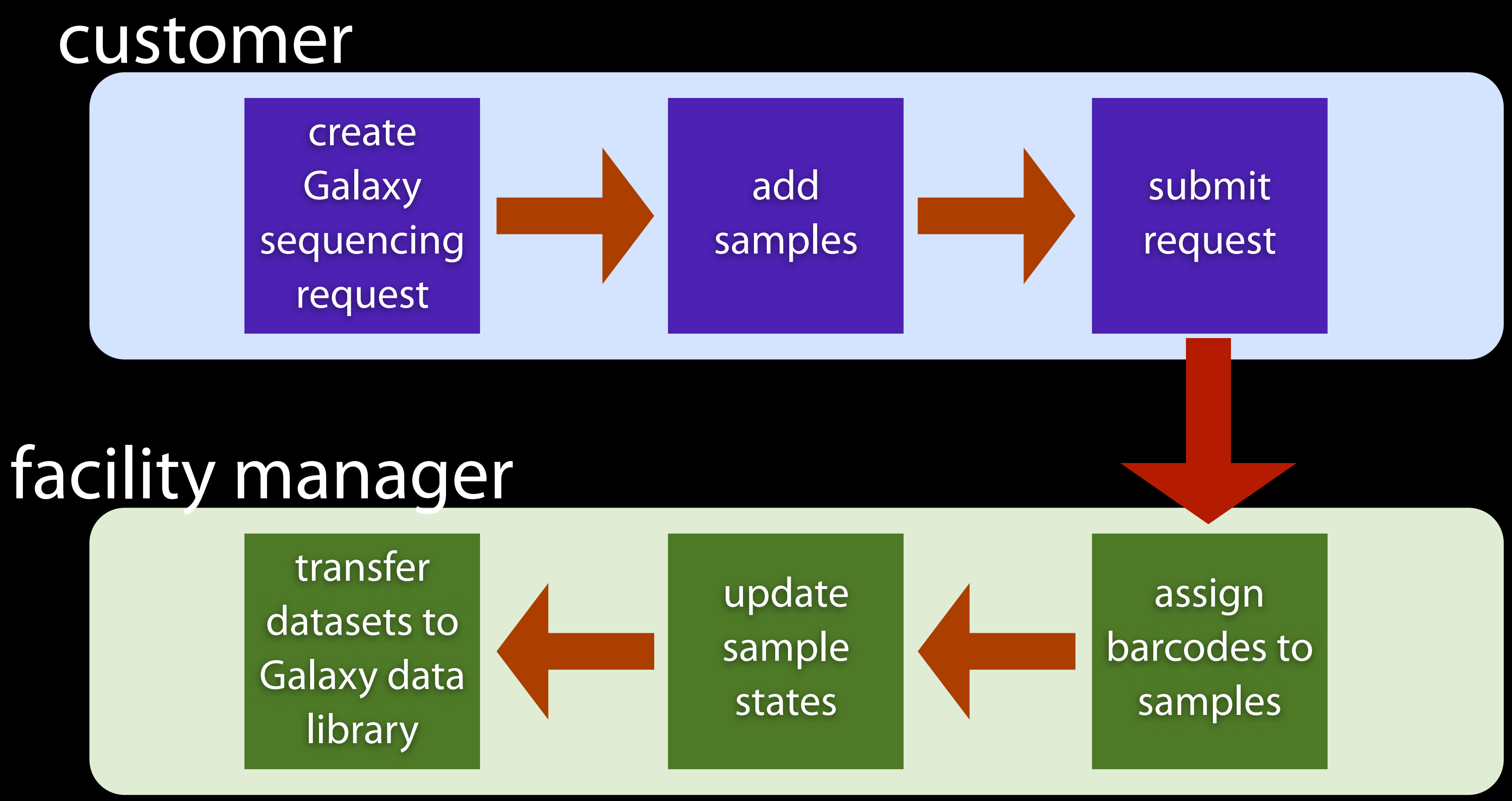
# example facility process

- A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results
- The facility bar codes the samples, scanning the tubes at each station in the lab
- The customer can watch the progress

# example facility process

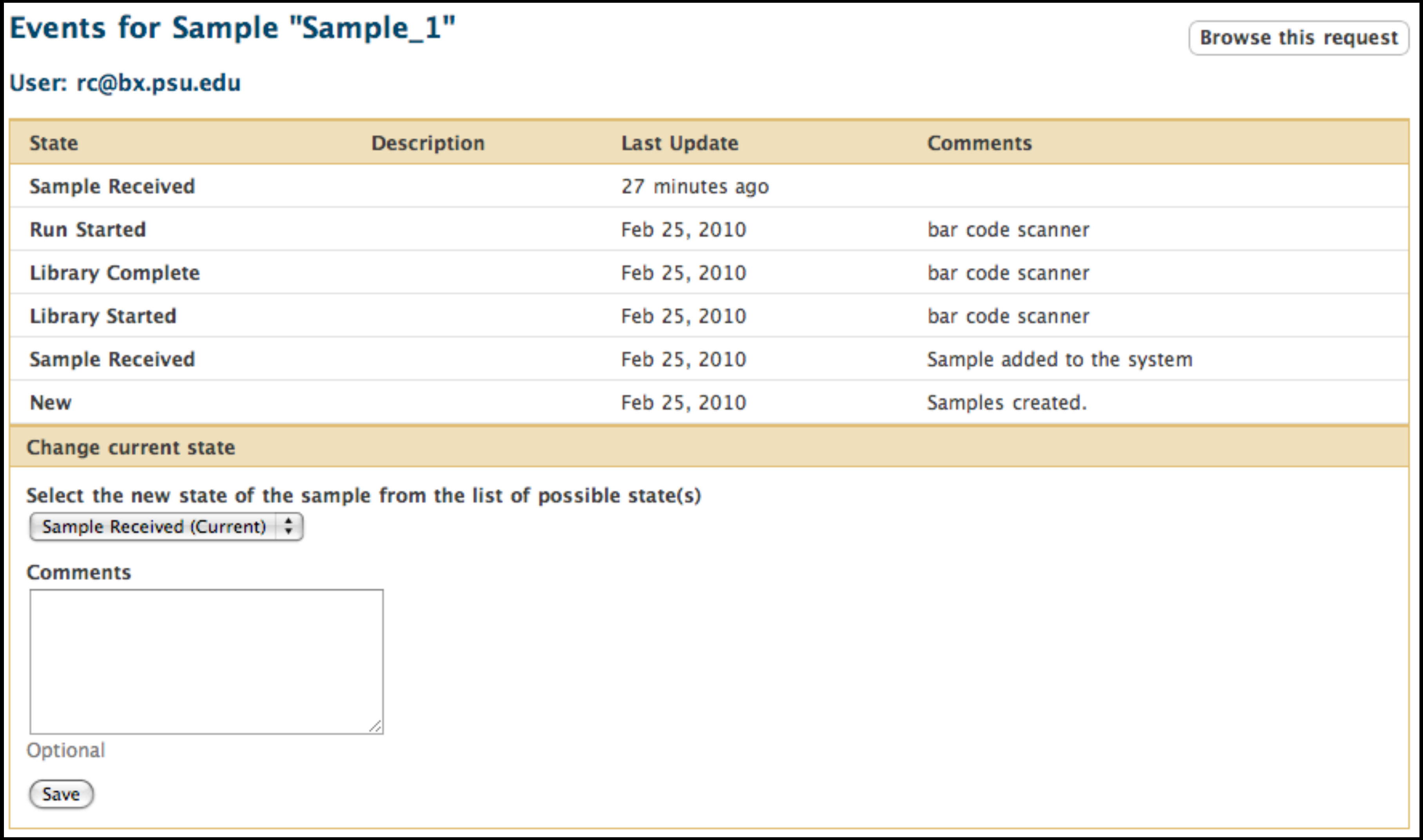
- A customer submits a sequencing request to a facility and delivers the samples, selecting a data library for the run results
- The facility bar codes the samples, scanning the tubes at each station in the lab
- The customer can watch the progress
- When the run is complete, the resulting data is transferred to the requested data library

# who does what?



customer
create Galaxy sequencing request
add samples
submit request
facility manager
assign barcodes to samples
update sample states
transfer datasets to Galaxy data library

# example sequencing request lifecycle



## Events for Sample "Sample_1"

User: rc@bx.psu.edu

| State | Description | Last Update | Comments |
|---|---|---|---|
| Sample Received | | 27 minutes ago | |
| Run Started | | Feb 25, 2010 | bar code scanner |
| Library Complete | | Feb 25, 2010 | bar code scanner |
| Library Started | | Feb 25, 2010 | bar code scanner |
| Sample Received | | Feb 25, 2010 | Sample added to the system |
| New | | Feb 25, 2010 | Samples created. |

Change current state

Select the new state of the sample from the list of possible state(s)

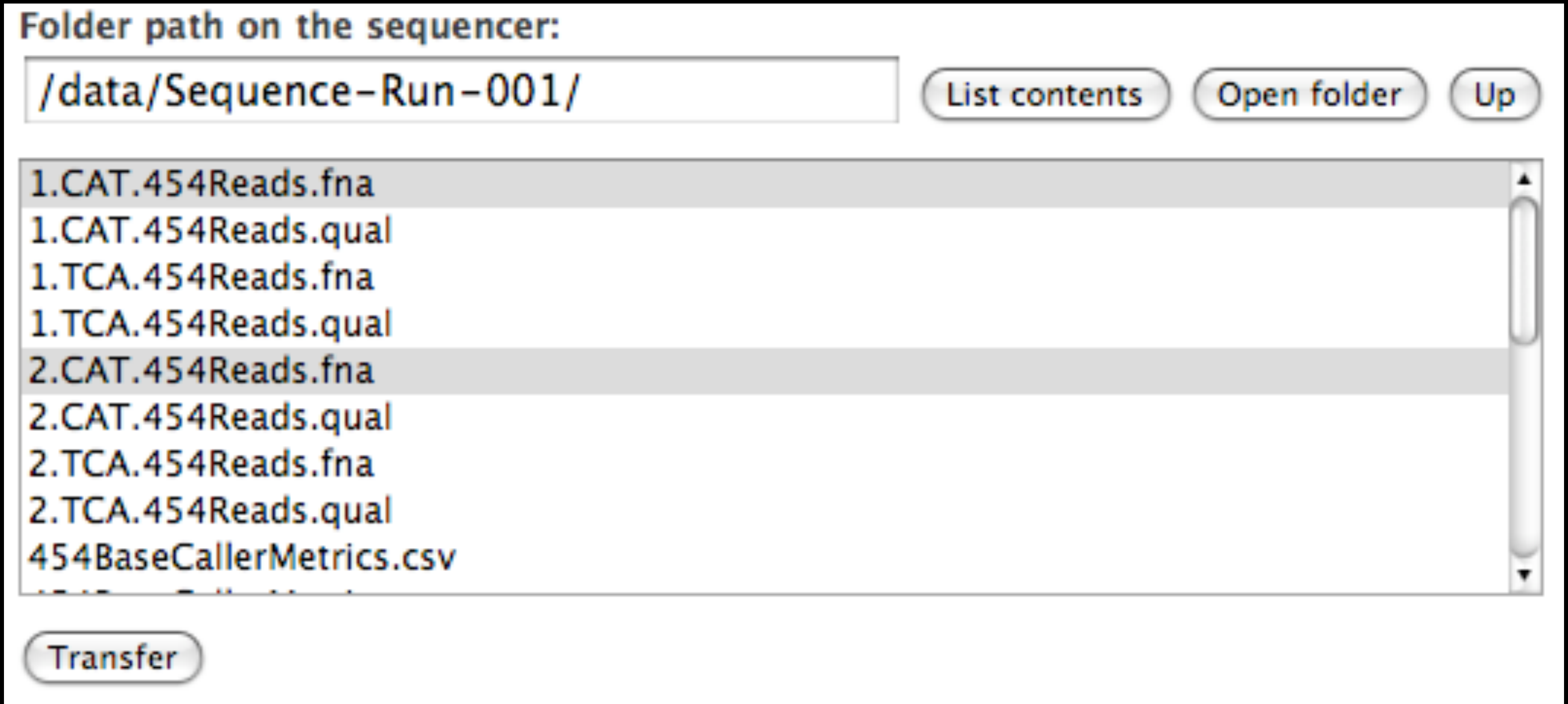
Sample Received (Current)

Comments

Optional

# transferring the data

- Remote file browser to select datasets on the sequencer and transfer them to the requested Galaxy data library



Folder path on the sequencer:

/data/Sequence-Run-001/

1.CAT.454Reads.fna
1.CAT.454Reads.qual
1.TCA.454Reads.fna
1.TCA.454Reads.qual
2.CAT.454Reads.fna
2.CAT.454Reads.qual
2.TCA.454Reads.fna
2.TCA.454Reads.qual
454BaseCallerMetrics.csv

What if my facility does things differently?

# You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time

# You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
- Lifecycle “states” of the request are defined by the “stations” in the lab

# You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
- Lifecycle “states” of the request are defined by the “stations” in the lab
- Bar code scanners can be used, but manual data entry is also supported

# You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
- Lifecycle “states” of the request are defined by the “stations” in the lab
- Bar code scanners can be used, but manual data entry is also supported
- Configure the sequencer information for communication with Galaxy

# You define the process

- The layout of the request and sample forms is defined by the lab and can be changed over time
- Lifecycle “states” of the request are defined by the “stations” in the lab
- Bar code scanners can be used, but manual data entry is also supported
- Configure the sequencer information for communication with Galaxy
- Permissions to submit requests is granted to specific users

# What if my facility uses a LIMS? Will it work with Galaxy?

# Yes!

- Galaxy sample tracking complements existing LIMS applications, it is not intended to replace them.

# Yes!

- Galaxy sample tracking complements existing LIMS applications, it is not intended to replace them.
- Galaxy uses a generic messaging engine with a very simple XML api for communication with the sequencer. This same messaging engine can communicate with a LIMS application.

# For more information about data libraries...

# For more information about data libraries...

- See me at the bar

# bonus topic

Wouldn’t it be cool if there was a place that allowed me to upload tools that I’ve developed, and share them with others?

How about if this place allowed me to easily find someone else's tool that I'm interested in using in my own Galaxy instance?

Maybe allow me to browse tools by category, or search for tools using names or descriptions?

Would’t it be nice if I could upload new and improved versions of tools that are already available there...

...you know, since I’m a better programmer than the one that wrote it the first time?

the time has come...

http://usegalaxy.org/community