FlowRepositoryR: the FlowRepository R Interface

Josef Spidlen

April 16, 2015

Abstract

FlowRepository is a free public flow cytometry data repository intended for authors of peer-reviewed manuscripts to deposit their underlying flow cytometry data, provide annotations, and share annotated datasets upon publication. Primarily, FlowRepository is accessed via a web-based user interface (https://flowrepository.org/), however, it also includes an application programming interface (API), which allows for programmatic access from other software tools. FlowRepositoryR is an R library that utilizes this API allowing users to locate available datasets, review provided annotations and download releated data files to their local file system, all this conveniently from within R without the need of opening a web browser. Downloaded datasets can then be easily analyzed using flowCore and other libraries developed for this purpose.

Keywords: FlowRepository, flow cytometry, data repository, API, application programming interface, dataset download

1 Introduction

1.1 Background

Data associated with publications should be available and accessible. Transparency and public availability of protocols, data, analyses, and results are crucial to make sense of the complex biology of human diseases. Funding agencies, regulatory agencies, publishers, and the scientific community have all recognized the importance of protecting cumulative data outputs to accelerate subsequent exploitation through the community-based development of public data repositories (Field et al., 2009).

1.2 FlowRepository

Until recently, no public repository existed for flow cytometry data. In order to address this issue, we developed FlowRepository (Spidlen et al., 2012b,a) - a public resource for authors to deposit their flow cytometry data, provide MIFlowCyt (Lee et al., 2008) compliant annotation, and share annotated datasets upon publication. Development and maintenance of FlowRepository is generously supported by the Wallace H. Coulter Foundation, the International Society for Advancement of Cytometry (ISAC), the International Clinical Cytometry Society, various research grants and the flow cytometry community in general. Technically, FlowRepository has been developed by extending and adapting Cytobank (Kotecha et al., 2010), an online tool for

storage and collaborative analysis of cytometric data. Primarily, FlowRepository is accessed via a web-based user interface (`https://flowrepository.org/`), however, it also includes an XML-based application programming interface (API). This API allows for programmatic access from other software tools.

1.3 FlowRepositoryR

FlowRepositoryR is an R library that utilizes this API (Spidlen, 2015) to allow users locate available datasets, review annotations and download releated data files to their local file system. Below, we will demonstrate how this library can be used.

2 Typical Use

2.1 Requirements

You will need R libraries XML, RCurl, and tools in order to install FlowRepositoryR. These are being used to parse the XML used to communicate with the FlowRepository server and to establish the HTTP(s) connection. If you are installing from BioConductor, biocLite should resolve those dependencies for you. In addition, it is recommended to also have the RUnit library to be able to run unit tests. Finally, you will likely want to have flowCore (Hahne et al., 2009) and other related libraries in order to analyze data from FlowRepository datasets obtained using the FlowRepositoryR package. Assuming you have installed FlowRepositoryR already, we will start by loading the library.

```
> library(FlowRepositoryR)
```

2.2 List available datasets

FlowRepository has been live since 2012 and it continues to see a steady increase in users, data submissions and downloads. As of March 2015, there are 440 datasets, 215 of those are public. The majority of the private datasets are presumed to be related to manuscripts that are currently under peer-review and will be made public once these manuscripts are published. You can use the `flowRep.ls` function in order to list the identifiers of currently available datasets.

```
> dataSets <- flowRep.ls()
> ## We will only show a maximum of 10 identifiers so that we don't
> ## clutter the vignette
> dataSets[1:min(10, length(dataSets))]
```

```
 [1] "FR-FCM-ZZZ3" "FR-FCM-ZZZ4" "FR-FCM-ZZZA" "FR-FCM-ZZZE" "FR-FCM-ZZZF"
 [6] "FR-FCM-ZZZG" "FR-FCM-ZZZH" "FR-FCM-ZZZK" "FR-FCM-ZZZU" "FR-FCM-ZZZV"
```

2.3 Review information about a datasets

While an extended search functionality is being developed, for now we will assume that you know which dataset you are interested in. You can use the `flowRep.get` function in order to

obtain a dataset from FlowRepository. This will retrieve information about the dataset but it will not download the data.

```
> ## FR-FCM-ZZJ7 is a purposely picked dataset that is public and very
> ## small for the unit tests and the vignette and man pages to compile
> ## quickly. Also, FlowRepository is not tracking the downloads of this
> ## particular dataset since the stats would be based mainly on these
> ## automated downloads.
> ds <- flowRep.get("FR-FCM-ZZJ7")
> summary(ds)

A flowRepData object (FlowRepository dataset) GvHD data subset
2 FCS files, 2 attachments, NOT downloaded
```

This will return a FlowRepository dataset represented by an object of the *flowRepData* class. See section 3.1 for more details about the dataset, or you can also use the `str` command to inspect the returned object.

2.4 Download the data

Data associated with a FlowRepository dataset can be downloaded using the `download` method of the *flowRepData* class.

```
> ds <- download(ds)

Downloading to E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/v
File GvHD2.fcs downloaded.
File GvHD8.fcs downloaded.
File about.txt downloaded.
File ExampleGate.png downloaded.
Download finished.

> summary(ds)

A flowRepData object (FlowRepository dataset) GvHD data subset
2 FCS files, 2 attachments, downloaded
```

Assuming the dataset exists and you have permissions to access it, this will download the whole dataset including all FCS files and attachment files associated with it. Unless specified otherwise (see section 2.6), the download method will create a new directory in your current working directory, name it based on the identifier of the dataset, and dowload the files there. A separate `attachments` subfolder will be created for the attachments. The location where these files were downloaded can be obtained from the local path slot of the file proxies. For example, the local path of the first downloaded dataset can be obtained as follows:

```
> localpath(fcs.files(ds)[[1]])
```

```
[1] "E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/vignettes/Fl
```

If we wanted the local path of all the downloaded FCS files, we could use the `lapply` function as follows:

```
> unlist(lapply(fcs.files(ds), function(x) paste(localpath(x))))
```

```
[1] "E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/vignettes/Fl
[2] "E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/vignettes/Fl
```

Analogously, we can locate all the attachments as follows:

```
> unlist(lapply(attachments(ds), function(x) paste(localpath(x))))
```

```
[1] "E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/vignettes/Fl
[2] "E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/vignettes/Fl
```

2.5 Downloading private datasets

In order to download a private dataset, you will need to register with FlowRepository. Open your web browser and navigate to `http://flowrepository.org/`. Then follow the *Login* link in the top right corner of the page. Next, either Sign-in or follow the registration link if you haven't signed up yet. FlowRepository uses OpenID or Google+ authentication. Those are used for web-based authentication. The FlowRepositoryR package (and FlowRepository API in general) use a email/password based authentication. This needs to be set in your profile independently. Once you have logged in in your web browser, click on the *Welcome Your Name* link in the top right corner next to the *Logout* link. This will enter your profile. Next, follow the *Edit* link from the actions panel on your left. Scroll down and set your API password as shown in Figure 1. The API password shall use 8 or more characters and include at least one number, one upper-case character and one lower-case character. Set your password and confirm it by clicking on the *Update* button.

Once you have set your password online, you can use the `setFlowRepositoryCredentials` to set your FlowRepository API credentials, which will give you access to non-public datasets created by you or shared with you in FlowRepository.

```
> setFlowRepositoryCredentials(email="boo@gmail.com", password="foo123456")
```

Alternativelly, you can provide the `filename` argument instead of the email and passwords arguments, which will read your credentials from a text file. This file shall include 2 lines, email address in the first line, password in the second line. Finally, the function will prompt for credentials if called without arguments in an interactive mode.

Once your credentials are set, you can use the `include.private=TRUE` option of the `flowRep.ls()` function in order to include non-public dataset in the list of available datasets. In the `download` method, if credentials are set then those will be used automaticaly. You can disable this by passing the `use.credentials=FALSE` argument to the `download` method of a *flowRepData* object.

To conclude this section, let's forget the set credentials as the `boo@gmail.com` email and `foo123456` password are not real credentials to access FlowRepository.

```
> forgetFlowRepositoryCredentials()
```

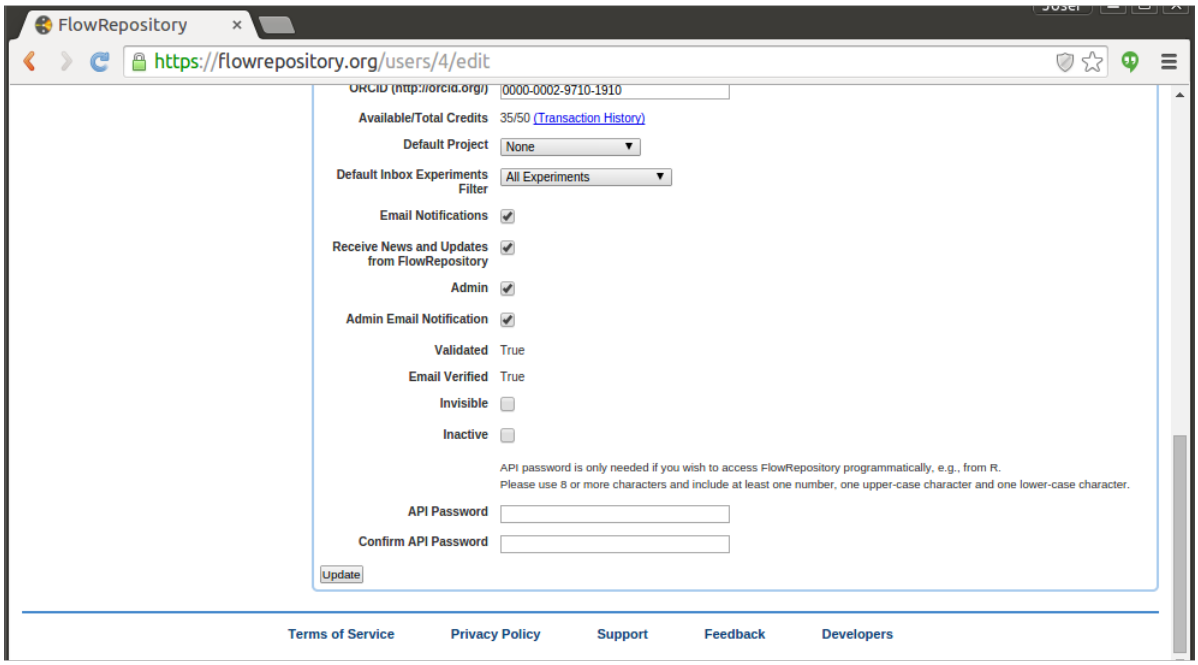

Figure 1: **Setting FlowRepository API access password.** FlowRepository uses OpenID or Google+ authentication for web-based access, but those are separate from the application programming access, which needs to be set in your profile by providing an API password.

2.6 Additional download options

The `dirpath` argument may be passed to the to the `download` method of a *flowRepData* object. This can be used to specify the directory on the local file system where the dataset shall be downloaded. By default, the files will be downloaded to a folder named based on the dataset identifier (FR-FCM-xxxx), which will be created in your current working directory.

If you don't want to see the progress about files as they are being downloaded, you can turn this off by passing the `show.progress=FALSE` argument to the `download` method of a *flowRepData* object.

2.7 Downloading only certain files from a dataset

Should you wish to download only some files of a FlowRepository dataset, you can do so by using the `download` method of the *fileProxy* objects (*i.e*, *fcsProxy* or *attachmentProxy*). For example

```
> myDataset <- flowRep.get("FR-FCM-ZZJ7")
> summary(myDataset)

A flowRepData object (FlowRepository dataset) GvHD data subset
2 FCS files, 2 attachments, NOT downloaded
```

```
> ## And download a single attachment file
> at1 <- download(attachments(myDataset)[[1]])

File about.txt downloaded.

> localpath(at1)

[1] "E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/vignettes/a

> summary(at1)

A fileProxy object (proxy for a file) about.txt (downloaded)

> ## A single FCS file proxy can be downloaded
> fcs1 <- download(fcs.files(myDataset)[[1]])

File GvHD2.fcs downloaded.

> localpath(fcs1)

[1] "E:/biocbld/bbs-3.1-bioc/tmpdir/Rtmp6VFOAo/Rbuild207849a932b8/FlowRepositoryR/vignettes/G

> summary(fcs1)

A fileProxy object (proxy for a file) GvHD2.fcs (downloaded)
```

3 Representing FlowRepository Datasets

3.1 The *flowRepData* Class

FlowRepository datasets are represented by *flowRepData* objects. Slots of this class capture the metadata (information about) the dataset as follows:

`id`: Object of class `character` containing the FlowRepository identified of the dataset. These identifiers are typically in the form of `FR-FCM-`xxxx where xxxx represents 4 alphanumeric characters.

`public.url`: Object of class `character` or NULL containing the public URL of this dataset. This will commonly be in the form of `https://flowrepository.org/id/`identifier, where identifier is the FlowRepository identified of the dataset.

`name`: Object of class `character` or NULL containing the name of this dataset.

`public`: Object of class `logical` or NULL containing the information whether this dataset is public.

`primary.researcher`: Object of class `character` or NULL containing the name of the primary researcher associated with this dataset.

`primary.investigator`: Object of class `character` or NULL containing the name of the primary investigator associated with this dataset.

`uploader`: Object of class `character` or NULL containing the name of the uploader of this dataset.

`experiment.dates`: Object of class `character` or NULL containing the dates associated with this dataset. Typically, there will be two dates associated with the dataset, the first one for the start of the experiment, the second one for the end of the experiment. A single date indicates the start of an experiment that may still be ongoing. The dates shall be encoded as "YYYY-MM-DD".

`purpose`: Object of class `character` or NULL stating the purpose of this dataset (experiment).

`conclusion`: Object of class `character` or NULL stating the conclusion associated with this dataset (typically conclusions reached by analyzing the data).

`comments`: Object of class `character` or NULL stating additional comments associated with this dataset.

`funding`: Object of class `character` or NULL stating the funding used to collect the data in this dataset.

`qc.measures`: Object of class `character` or NULL stating the quality control measures taken in order to ensure the quality of data in this dataset.

`miflowcyt.score`: Object of class `numeric` or NULL stating the MIFlowCyt compliance score of this experiment. MIFlowCyt is the Minimum Information about a Flow Cytometry Experiment - an ISAC Recommendation listing the minimum information that shall be provided as annotation of flow cytometry datasets. The MIFlowCyt compliance score is a value between 0 and 100 percent indicating the level of compliance with MIFlowCyt. Details about how FlowRepository calculates this score ara available here: `http://flowrepository.org/quick_start_guide#MIFlowCytScoreReport`

`keywords`: Object of class `list` (of objects of class `character`) enumerating keywords associated with this dataset.

`publications`: Object of class `list` (of objects of class `character`) enumerating publications associated with this dataset. Publications are typically listed as "PMID:12345678" or "PMCID:PMC1234567".

`organizations`: Object of class `list` of objects of class `flowRepOrganization` (see section 3.2) enumerating organizations associated with this dataset.

`fcs.files`: Object of class `list` of objects of class `fcsProxy` enumerating FCS files associated with this dataset.

`attachments`: Object of class `list` of objects of class `attachmentProxy` enumerating attachments associated with this dataset.

3.2 The *flowRepOrganization* Class

The *flowRepOrganization* class represents the name and address of an organization associated with a dataset stored in FlowRepository. Slots of this class capture the information as follows:

`name`: Object of class `character` containing the name of the organization.

`street`: Object of class `character` or NULL containing the street of the address of the organization.

`city`: Object of class `character` or NULL containing the city of the address of the organization.

`zip`: Object of class `character` or NULL containing the zip (or postal code) of the address of the organization.

`state`: Object of class `character` or NULL containing the state (or province) of the address of the organization.

`country`: Object of class `character` or NULL containing the country of the address of the organization.

References

D. Field, S. A. Sansone, A. Collis, T. Booth, P. Dukes, S. K. Gregurick, K. Kennedy, P. Kolar, E. Kolker, M. Maxon, S. Millard, A. M. Mugabushaka, N. Perrin, J. E. Remacle, K. Remington, P. Rocca-Serra, C. F. Taylor, M. Thorley, B. Tiwari, and J. Wilbanks. Megascience. 'Omics data sharing. *Science*, 326(5950):234–236, Oct 2009.

F. Hahne, N. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. flowCore: a Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 10:106, 2009.

N. Kotecha, P. O. Krutzik, and J. M. Irish. Web-based analysis and publication of flow cytometry experiments. *Curr Protoc Cytom*, Chapter 10:Unit10.17, Jul 2010.

J. A. Lee, J. Spidlen, K. Boyce, J. Cai, N. Crosbie, M. Dalphin, J. Furlong, M. Gasparetto, M. Goldberg, E. M. Goralczyk, B. Hyun, K. Jansen, T. Kollmann, M. Kong, R. Leif, S. McWeeney, T. D. Moloshok, W. Moore, G. Nolan, J. Nolan, J. Nikolich-Zugich, D. Parrish, B. Purcell, Y. Qian, B. Selvaraj, C. Smith, O. Tchuvatkina, A. Wertheimer, P. Wilkinson, C. Wilson, J. Wood, R. Zigon, R. H. Scheuermann, and R. R. Brinkman. MIFlowCyt: the minimum information about a Flow Cytometry Experiment. *Cytometry A*, 73(10):926–930, Oct 2008.

J. Spidlen. Flowrepository Application Programming Interface, 2015. URL `http://flowrepository.org/images/pdf/FlowRepositoryAPI.pdf`. Accessed: 2015-03-18.

J. Spidlen, K. Breuer, and R. Brinkman. Preparing a Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) compliant manuscript using the International Society for Advancement of Cytometry (ISAC) FCS file repository (FlowRepository.org). *Curr Protoc Cytom*, Chapter 10:Unit 10.18, Jul 2012a.

J. Spidlen, K. Breuer, C. Rosenberg, N. Kotecha, and R. R. Brinkman. FlowRepository: a resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry A*, 81(9):727–731, Sep 2012b.