Kipper: A software package for sequence database versioning for Galaxy bioinformatics servers

Damion Dooley[1], Aaron Petkau[2], Gary VanDomselaar[2], William Hsiao[1,3]

[1] University of British Columbia, Vancouver, BC, Canada. Email: damion.dooley@bccdc.ca

[2] National Microbiology Laboratory, Public Health Agency of Canada, Winnipeg, MB, Canada.

[3] BC Public Health Microbiology and Reference Laboratory, Vancouver, BC, Canada.

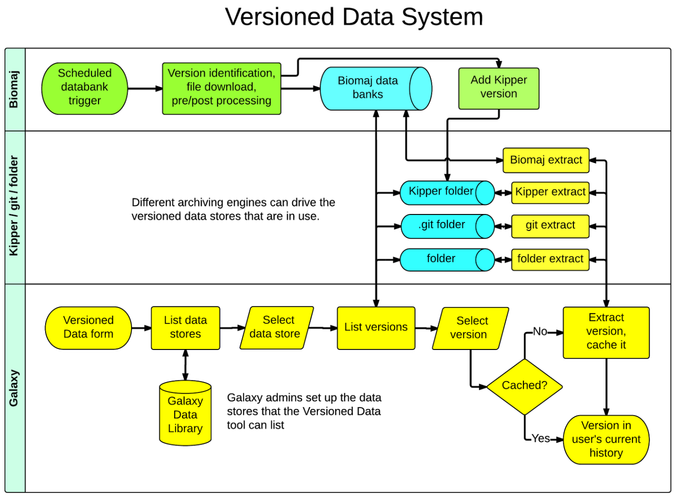
Project Website: https://github.com/Public-Health-Bioinformatics
Source Code: https://github.com/Public-Health-Bioinformatics/versioned_data
License: Licensed under the Academic Free License version 3.0

Abstract

There are various reasons to rerun bioinformatics tools and pipelines on sequencing data, including re-creating a past result, re-validation of a tool or workflow using a known dataset, or tracking the impact of database changes. For identical results to be achieved, updated reference sequence databases must be versioned. Server administrators have tried to fill the requirements by supplying users with one-off versions of databases, but these are time consuming to set up and to maintain. Disk storage and data backup performance has also discouraged maintaining multiple versions of databases since databases such as NCBI nr can consume 50Gb or more disk space per version, with growth rates that parallel Moore's law.

Our end-to-end open source versioning system combines our own Kipper software package - a simple key-value large-file versioning system - with Biomaj(a software system for downloading sequence databases), and Galaxy (a web-based bioinformatics workflow scheduling platform). Available versions of databases can be recalled and used via command-line or within Galaxy. The Kipper data store format makes publishing curated fasta databases especially convenient since in most cases it can store a range of versions into a file only slightly larger than the size of the latest version.

Kipper is under active development and we encourage feedback from the user community to improve its utility.


Versioned Data System
Biomaj
Kipper / git / folder
Galaxy
Scheduled databank trigger
Version identification, file download, pre/post processing
Biomaj data banks
Add Kipper version
Biomaj extract
Kipper folder
Kipper extract
.git folder
git extract
folder
folder extract
Different archiving engines can drive the versioned data stores that are in use.
Versioned Data form
List data stores
Select data store
List versions
Select version
Extract version, cache it
No
Cached?
Yes
Version in user's current history
Galaxy Data Library
Galaxy admins set up the data stores that the Versioned Data tool can list