Bibliospec (SQLite)

EBDM 2016

C.Panse & W.Wolski

FGCZ

Functional Genomics Center Zurich

- Core facility for genomics, metabolomics, proteomics
- Approx **40** scientist and technical staff
  - **10** in proteomics group, including **5** useR! with **2 R** package developers
  - **Applications:** Label Free Quantification (MS1, DIA, SRM, PRM), Protein Identification, PTMs, APMS , …

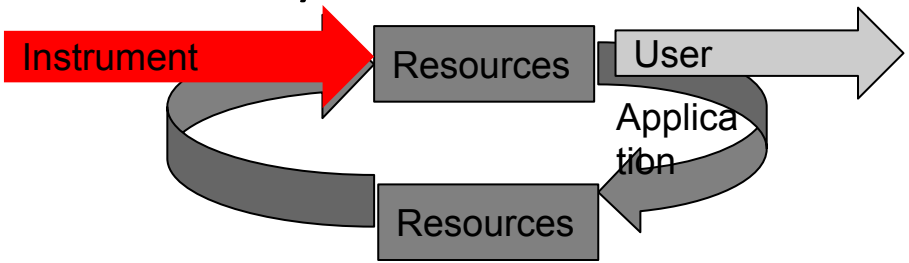
Proteomics Laboratory IT infrastructure

- Raw and processed data is stored since start of the FGCZ **2002**.
- The amount of data stored **doubles** every 18 months
- 2015: 14TB, 2016:12TB
- **NO XML** formats for MS data storage!

b-fabric information management platform

- Instrument vendor files: .raw, .bat ...
- applications: **.dat**, **tsv, csv**, **SQLite, pdf (with Rmarkdown)**

100% NO-PSI certified



ETH zürich | University of Zurich[UZH] | functional genomics center zurich | f·g·c·z

Project | Order | Agenda

Tasks | Projects | Samples | Extracts | Workunits | Datasets | More | ···b-fabric | wolski

Project 2185 running Identification of novel players in liver regeneration

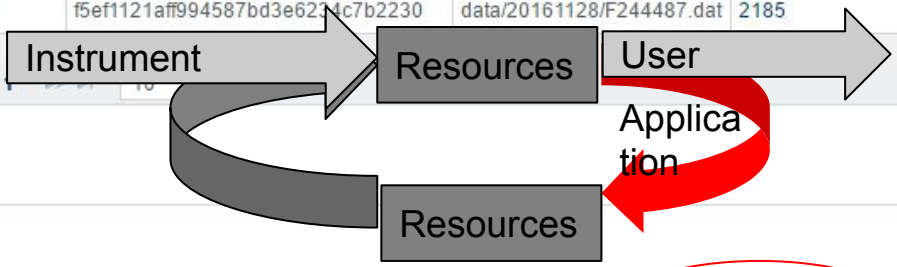
S | Search Site

Resource Basket

Select	Remove	Resource Id	Name	Status	Size	File Checksum	Relative Path	Project
☑	Remove	261990	data/20161128/F244483.dat	attached Access	120.907 MB	804dc253dc7662681b45d43f98f06303	data/20161128/F244483.dat	2185
☑	Remove	261979	data/20161128/F244482.dat	attached Access	124.898 MB	3e138436d5913cdf8ae2e168a8880921	data/20161128/F244482.dat	2185
☑	Remove	260635	data/20161110/F243851.dat	attached Access	25.286 MB	54a99a01f498c21f9b9cfbc00110d6ab	data/20161110/F243851.dat	2185
☑	Remove	261996	data/20161128/F244486.dat	attached Access	118.983 MB	b86ea489162873261f5f005897e54da7	data/20161128/F244486.dat	2185
☑	Remove	261991	data/20161128/F244489.dat	attached Access	113.849 MB	44c18c6888d539e4345607d1416bf977	data/20161128/F244489.dat	2185
☑	Remove	261992	data/20161128/F244487.dat	attached Access	99.452 MB	f5ef1121aff994587bd3e62[illegible]4c7b2230	data/20161128/F244487.dat	2185

Total: 6 / 6 Rows

Create Dataset



Run Application on Selected Resources

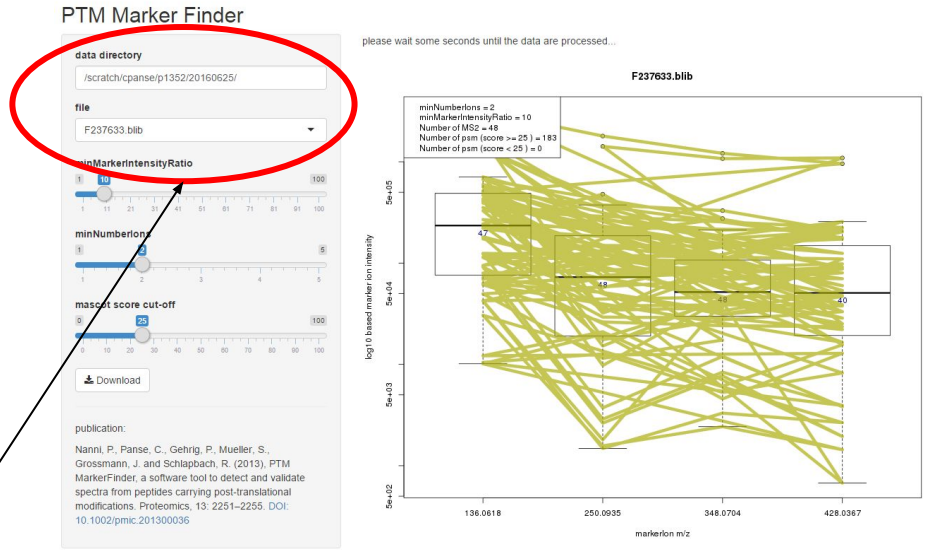
Mascot site localization export | Scaffold, mudpit for iTRAQ 8-plex | fgcz_mascot2specL_psmSet | ssrc | nonbatchdemo | scaffold_generic | gerneric_yaml | DIA Assay Library Generator | BiblioSpec

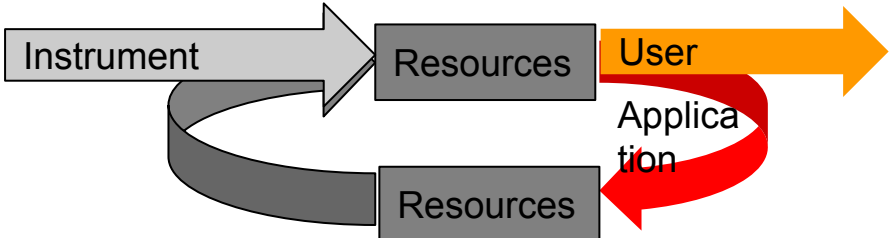
Shiny - (external) applications and interactive result viewers

Good container for data in interactive applications is **SQLite.**

SQLite is an in-process library that implements a self-contained, serverless, zero-configuration, transactional SQL database engine.

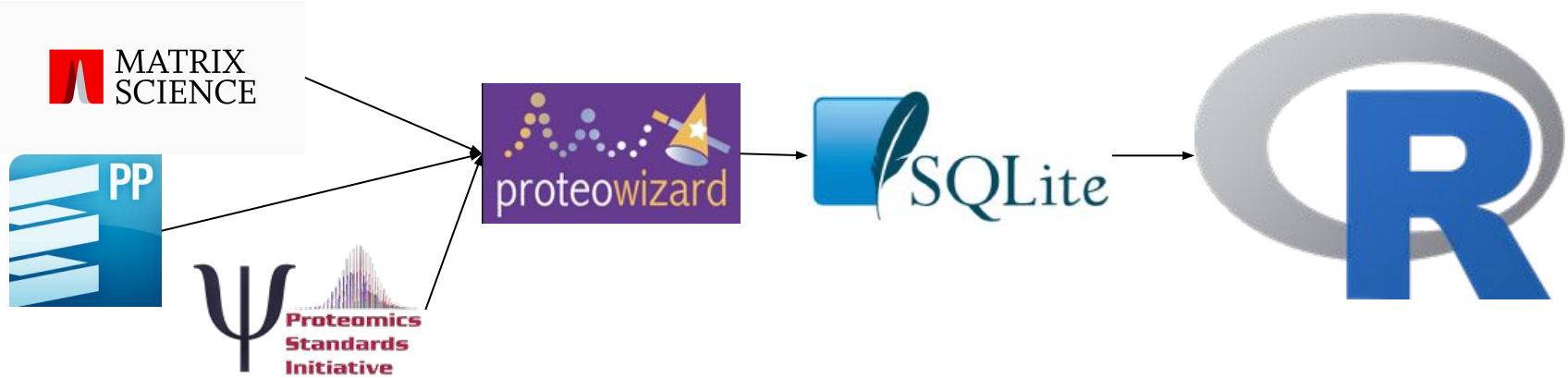
Specify resource ID. Storage is mounted to the shiny server





What is bibliospec?

- Converts **MS peptide identification** results into a relational database in SQLite
- ProteoWizard **BiblioSpec** executable (written in C++).
- Can read .idXML, .pep.XML, .group.XML, .dat etc.
- R-package **bibliospec** is an ORM using R-reference classes
  - methods to return most frequently used views as data.frame
  - methods for spectral count based protein quantification



Code example

```
> library(bibliospec)
> BB <- Blib()
> files <- NULL
> datdir <- 'd:/projects/p2069/dataSearchResults/ProteinPilot/xml/'
> files <- dir(datdir)
> files<-file.path(datdir,files)
> files
[1] "d:/projects/p2069/dataSearchResults/ProteinPilot/xml//GRAY_HD.group.xml"  "d:/projects/p2069/dataSearchResults/ProteinPilot/xml//GRAY_MS.group.xml"
[3] "d:/projects/p2069/dataSearchResults/ProteinPilot/xml//WHITE_HD.group.xml" "d:/projects/p2069/dataSearchResults/ProteinPilot/xml//WHITE_MS.group.xml"
>
```

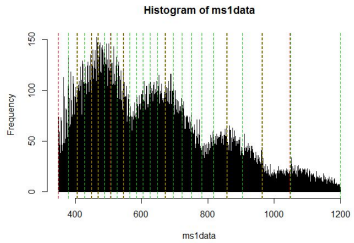
```
BB$build(files, outfile = "blibFiles/PP0.95.blib", cutoff=0.95)
BB$filter(infile="blibFiles/PP0.95.blib", outfile = "blibFiles/PP0.95.filtered.blib", minpeaks =10 )
```

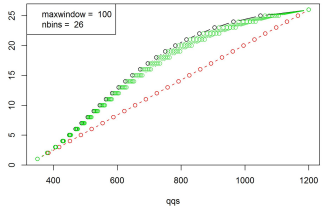
```
> library(bibliospec)
> BS <- Bibliospec(dbfile="../blibFiles/PP0.95.blib")
database connected
> dim(BS$summary())
[1] 44  8
> BS$getNrPSM()
  count(*)
1   528401
> allPeaks <- BS$getPeaks()
getspectra
> dim(allPeaks)
[1] 149345045         3
> round(dim(allPeaks)[1],digits = -6)
[1] 1.49e+08
>
```

Calls ProtViz **BlibBuild** executable.
Since you can't ship exe's with R packages
this function uses `utils::download.files` to download exe
from github.

Packages to analyse MS peptide identification results

cdsw - computing dynamic swath windows (github)

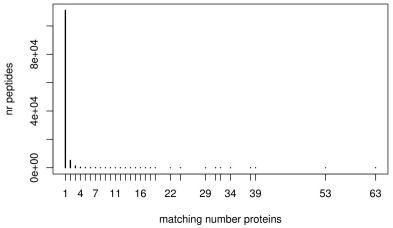


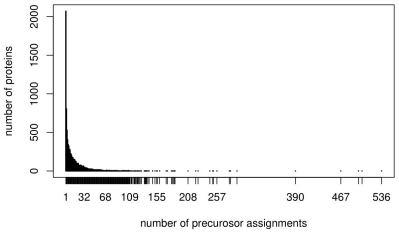


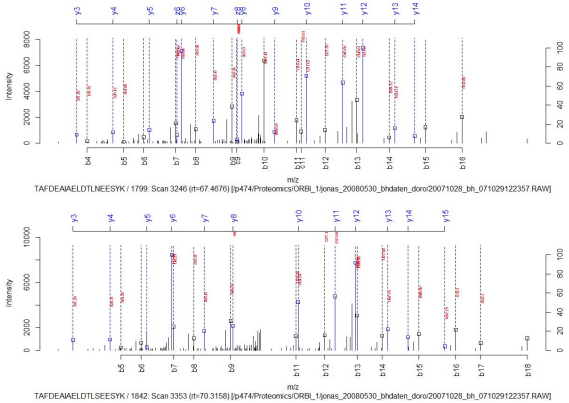
SpecL - generate Assay libraries for DIA (Bioconductor)

ProtViz - annotates and visualizes search results and performs PTM analysis (github, CRAN)

Prozor - protein inference - computes minimal set of proteins explaining peptides (github, CRAN)







Packages to analyse MS peptide identification results

https://github.com/protviz

Examples

```
library(bibliospec)

# use the sqlite file provided in the package
dbfile <- file.path(path.package("bibliospec"),
 "extdata/peptideStd.sqlite")

# call constructor
BS <- Bibliospec(dbfile=dbfile)
# test; should return TRUE
BS$getNrPSM() == 137

S <- BS$getPsmSet()

## Not run:
 library(specL)
 print(S)
 lapply(S[1:10], plot)

## End(Not run)

peaks <- BS$getPeaks()

print(BS$summary())
head(peaks)
```

Examples

```
BB <- Blib()
datdir <- file.path(path.package("bibliospec"),
 "extdata")
tmp <- file.path(datdir,dir(datdir))
zip <- grep("*.zip",tmp,value=T)
unzip(zip ,exdir = datdir)
dat <-grep("*\\.dat",file.path(datdir,dir(datdir)),value=T)
BB$build(dat , outfile = "test0_0.blib", cutoff=0.0 )
BS <- Bibliospec("test0_0.blib")
BS$summary()

spectrMet <- BS$getSpectraMeta()

peaks <- BS$getPeaks()
head(peaks)
length(unique(peaks$RefSpectraId))

BS <- NULL
```

- **Examples** to **document** and **TEST** code
- roxygen2, devtools
- data in the `inst/extdata` directory
  - .tsv, SQLite, .dat, .fasta etc.
  - realistic sizes
- `data/` - we are NOT using serialized R objects
  - makes refactoring almost impossible
- `R CMD check` may take minutes
- post and pre conditions checks in functions and in examples (`stopifnot`)

R / Shiny/ SQLite - the perfect team?

What **SQLite** can do for you:

- although data is on disk it feels like you had in memory.
- For large data (hundred millions of rows) **SQLite** outperforms almost any operation on a R data.frame
- Alleviates R memory problems

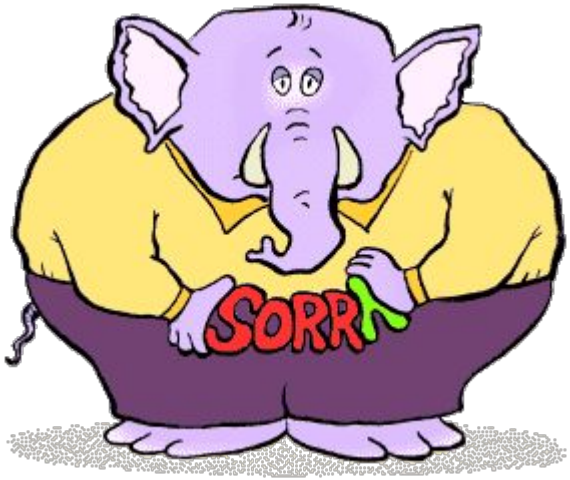
- SQLite : **CRAN - DBI, dplyr RSQLite**
  - Thermo Scientific Proteome Discoverer
  - timsTOF by Bruker Daltonics
  - **android**
  - Visual Studio
  - Scaffold
- HDF5 : **CRAN - h5**
  - TofWerk vendor format

Visual Studio

TOFWERK Time-of-Flight MS

Thermo SCIENTIFIC

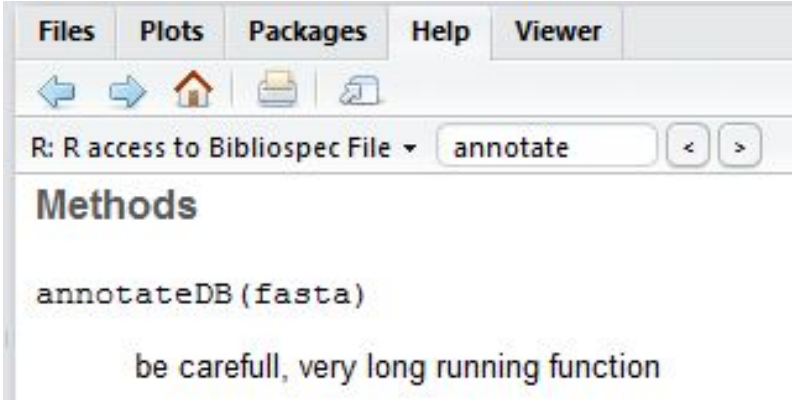
BRUKER


SORRY

R / Shiny/ SQLite - the perfect team?

SQLite can NOT:

- speed up computing with **R**
- Improve **tooling** for **R**
  - **SQLite** would make debugging more difficult if R Studio debugger would work with R reference classes.



Quo vadis scientific application development?

The prototypers (engineers and scientists) are into R, Mathematica, and Python. It might be fair to say that they were **moving from R** to the other two. Then along came **Hadley Wickham** on a one-man crusade to drag R into the 21st century.

Things like dplyr, ggplot2, data.table, and a few other new packages have completely changed the face of R, much to the chagrin of the old guard and delight of users.

http://bruceeckel.github.io/**2015/02/15**/why-not-go-there/

SQLite can simplify integration of R applications with application written in other languages.

2009 - GO
2012 - TS, Julia
2016 - Kotlin, .Net Core 1.0
2017 - Python 3.6 introduces Jit compilation in CPython https://github.com/Microsoft/Pyjion

Acknowledgments

Proteomics group at **FGCZ**

FGCZ b-fabric

Can Türker

Ugur Gürel

Universität Zürich[UZH] ETH zürich