

Data Management (from day 0)

Egon Willighagen (@egonwillighagen)

3 April 2014, Masterclass RDM in NL

Day 0: data plan

Before you start doing an experiment, you get a lab notebook.

(Some universities already require electronic lab notebooks!)

Day 1: the electronic lab notebook

- Version Control System
 - Allows backups
 - Allows annotation
 - Dated changes

The screenshot displays a Subversion (SVN) client window. The top menu bar includes 'File', 'Edit', 'View', and 'Help'. The main window is divided into several panes:

- Commit Log:** A list of recent commits with details such as the author (egonw), the commit message (e.g., 'Added CNMR data for Buma data set'), and the timestamp (e.g., '2004-05-27 13:42:46').
- File Details:** A pane showing the SHA1 ID of the selected file (5dceaf069ebef805a3d85492ed1d6b245dbdde4a) and the row number (1275 / 1364).
- Search:** A search bar with the text 'commit containing:'.
- Diff View:** A pane showing the differences between the current version and the previous one. It includes the author (egonw), the commit message, and the diff output for the file 'projects/svrqsar/hnmr/AllesVanHarm.R'. The diff shows the addition of a new file with mode 100644 and index 0000000..32661b3. The diff content includes R script code for reading a table and creating a matrix.
- Comments:** A pane showing the commit message for the selected commit: 'projects/svrqsar/hnmr/AllesVanHarm.R'.

Day 2: be careful what data you use

- Availability in 4 years?
 - Your Library/University has a copy?
- Can you read the format?
- Can you copy the data and share (e.g. with collaborators)?
- What if the journal you publish in requires you to share data?

Day 3: store *everything*

- Experiments
 - Description
 - Results (images, measurements, ...)
- Written output
 - Reports, papers, presentations

```
egonw@elitebook:~/var/Projects/hg/runrepos/
├── art.aux
├── art.bbl
├── art.blg
├── art.log
├── art.pdf
├── art.ps
├── art.tex
├── images
│   ├── boxplot.png
│   ├── dendro4.png
│   ├── flowchart.dia
│   ├── flowchart.png
│   ├── flowchart.ps
│   ├── preferred_bits.png
│   └── sammon.png
├── jref.log
├── list_refs.pl
├── Makefile
├── make_refs.pl
├── refs.bib
├── refs.bib.extra
├── report.pl
└── stats.sh
```

```
egonw@elitebook:~/var/Projects/hg/eNanoMapper$ tree
.
├── 20130315
│   ├── eNanoMapper-130315-Merging.docx
│   ├── eNanoMapper-130315-NJ-WP3.doc
│   ├── eNanoMapper-130315-Sect3Update-13u25.doc
│   ├── eNanoMapper-130315-Sect3Update-14u25.doc
│   ├── eNanoMapper-130315_WP5_ch.doc
│   └── eNanoMapper-130315_WP5_initial_ch.doc
├── 20130316
│   ├── eNanoMapper-130316.docx
│   ├── eNanoMapper-130316_EW_1209_BH.docx
│   └── eNanoMapper-130316_EW_1209.docx
```

Day 4: Analyse data directly from a repository

```
mart = biomaRt::useMart(biomart="snp", dataset="hsapiens_snp")
brca1 = c("rs16940","rs16941", "rs16942", "rs799916", "rs799917")
data = biomaRt::getBM(attributes=attribs, filters=c("snp_filter"),
  values=brca1, mart=mart)

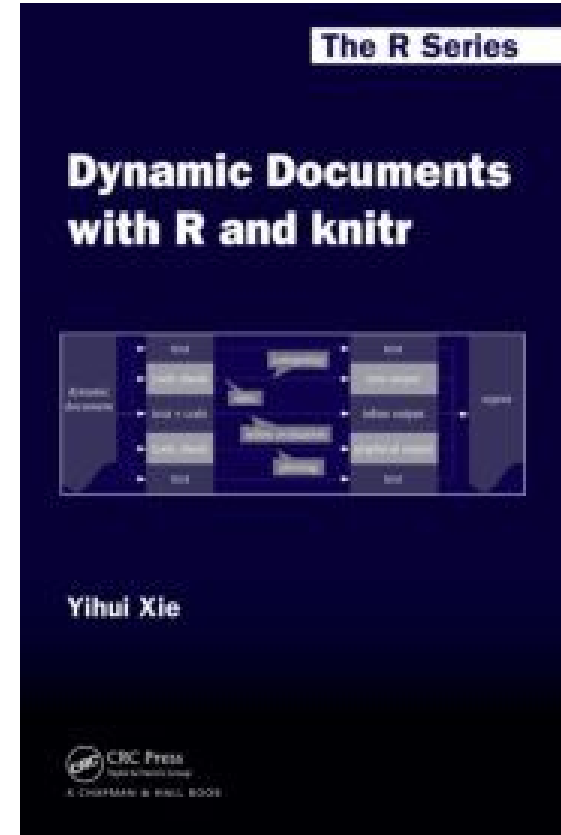
results = sparql.remote(
  "http://rdf.farmbio.uu.se/chembl/sparql", paste(
    "SELECT DISTINCT ?predicate ?object WHERE {",
    " ?assay <http://www.w3.org/2000/01/rdf-schema#label> \"CHEMBL615603\" ;",
    " ?predicate ?object . }"
  ))
```

Willighagen E. (2014) Accessing biological data in R with semantic web technologies. PeerJ PrePrints 2:e185v3. 10.7287/peerj.preprints.185v3

Day 4: Analyses inside your report

<p>We can also produce plots
(centered by the
option
`fig.align='center'`):
</p>

```
<!--begin.rcode html-cars-scatter,  
message=FALSE, fig.align='center'  
library(ggplot2)  
plot(mpg~hp, mtcars)  
qplot(hp, mpg, data=mtcars)  
+geom_smooth()  
end.rcode-->
```




<http://yihui.name/knitr/>

Day 5: Large Repositories

- Uniprot, ChEMBL, Gene Ontology
 - Is there a deposition workflow?
- Growing repositories
 - WikiPathways
- Set up a new database (paper+1)
 - e.g. DrugMet
 - Problem: what about small data?

- Journal driven
 - CSD
 - PDB

 BETA
[pathway](#) [discussion](#) [view source](#)

DNA Damage Response (Homo sapiens)

Jonas Hummel, Alexander Pico, Stan Gaj, Martijn van Iersel, et al.

search

navigation

- Home
- Help

pathway

- Create
- Browse
- Wish List
- Download
- Web service API

overview

- Recent Changes

Day 5: Database Seeds

- Set up a new database (paper += 1)
 - e.g. DrugMet



Navigation

[Main page](#)
[Project Planning](#)
[Recent changes](#)
[Random page](#)
[Help](#)

Semantic tools

[SPARQL Endpoint](#)
[RDF Import](#)

Toolbox

[What links here](#)
[Related changes](#)
[Special pages](#)
[Printable version](#)
[Permanent link](#)
[Browse properties](#)

Administration

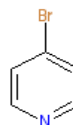
[ARC2Admin](#)

Page [Discussion](#) [Read](#) [View source](#) [View history](#) [Go](#) [Search](#)

4-bromopyridine

4-bromopyridine

InChIKey [BSDGZUDFPKIYQG-UHFFFAOYSA-N](#)
DOI [10.1021/ed050p510](#)



Facts about 4-bromopyridine ⓘ

[RDF feed](#) ⓘ

CHEMINF	000200	BSDGZUDFPKIYQG-UHFFFAOYSA-N ⓘ
Equivalent URI	http://drugmet.rilspace.org/resource/4-bromopyridine ⓘ ⓘ	
HasPKaValue	An experimental pKa value from 10.1021/ed050p510 for 4-bromopyridine ⓘ ⓘ	
IsDiscussedBy	Paper with DOI 10.1021/ed050p510 ⓘ	
Label	4-bromopyridine ⓘ	
Original URI	http://drugmet.rilspace.org/resource/ ⓘ ⓘ	
SubClassOf	CHEMINF 000000 ⓘ	

S. Lampa + me
CC-SA, but data CC0

Day 5: National Repositories

Dutch Dataverse Network >

POWERED BY THE **Dataverse Network™** PROJECT
 v. 3.3

ChEMBL-RDF data Dataverse



Log In

Log in with DVN account

CHEMBL-RDF V13.5

hdl:10411/10279

Version: 1 – Released: Thu Aug 29 19:06:18 CEST 2013

CATALOGING INFORMATION

Data & Analysis

Comments (0)

Versions

i If you use these data, please add the following citation to your scholarly references. [Why cite?](#)

Data Citation

Egon Willighagen, 2013-08-29, "ChEMBL-RDF v13.5", <http://hdl.handle.net/10411/10279> v1 [Version]

Citation Format Print ▼

Publications

Willighagen, E., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A., Tkachenko, V., Hastings, J., Chen, B., Wild, D., 2013. The ChEMBL database as linked open data. Journal of Cheminformatics 5 (1), 23
ID: DOI:10.1186/1758-2946-5-23

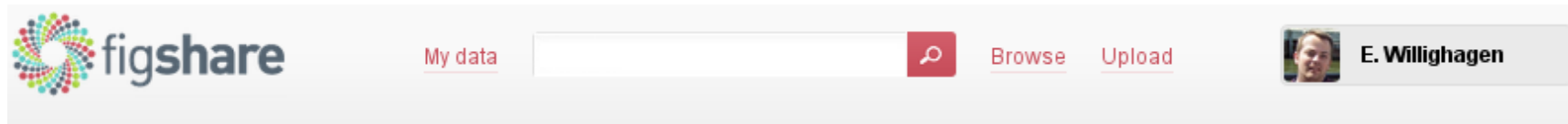
Data Citation Details ▼

Title ChEMBL-RDF v13.5

Study Global ID hdl:10411/10279

Authors Egon Willighagen

Day 5: Small Data @ FigShare



ChemPedia as RDF

Edit article

```
prefix dc: <http://purl.org/dc/elements/1.1/>
prefix cp: <http://rdf.openmolecules.net/chempedia/onto#>
<http://chempedia.com/substances/4-9789-8142-2454> dc:identifier "4-9789-8142-2454"
<http://www.w3.org/2002/07/owl#sameAs> <http://rdf.openmolecules.net/?InChI=1S/C26H44O/c1-17(2)6-5-7-18(3)21-10-11-12-13-14-15-16>
<http://chempedia.com/substances/6-3551-8801-2109> dc:identifier "6-3551-8801-2109"
<http://www.w3.org/2002/07/owl#sameAs> <http://rdf.openmolecules.net/?InChI=1S/C10H20O/c1-9(2)5-4-6-10(3)7-8-11>
<http://chempedia.com/substances/6-3551-8801-2109> cp:hasNaming <http://chempedia.com/substances/6-3551-8801-2109/naming0> a cp:Naming;
  cp:hasName "R-(+)-β-Citronellol";
  cp:hasStatus "OK";
  cp:hasScore "1".
<http://chempedia.com/substances/6-3551-8801-2109> cp:hasNaming <http://chempedia.com/substances/6-3551-8801-2109/naming0>
```

73
views

1
shares

cites
coming
soon

Published on 14 Apr 2013 - 12:33 (GMT)

Filesize is 327.92 KB

Categories

- Organic Chemistry
- Cheminformatics

Day 5: Scientific dissemination

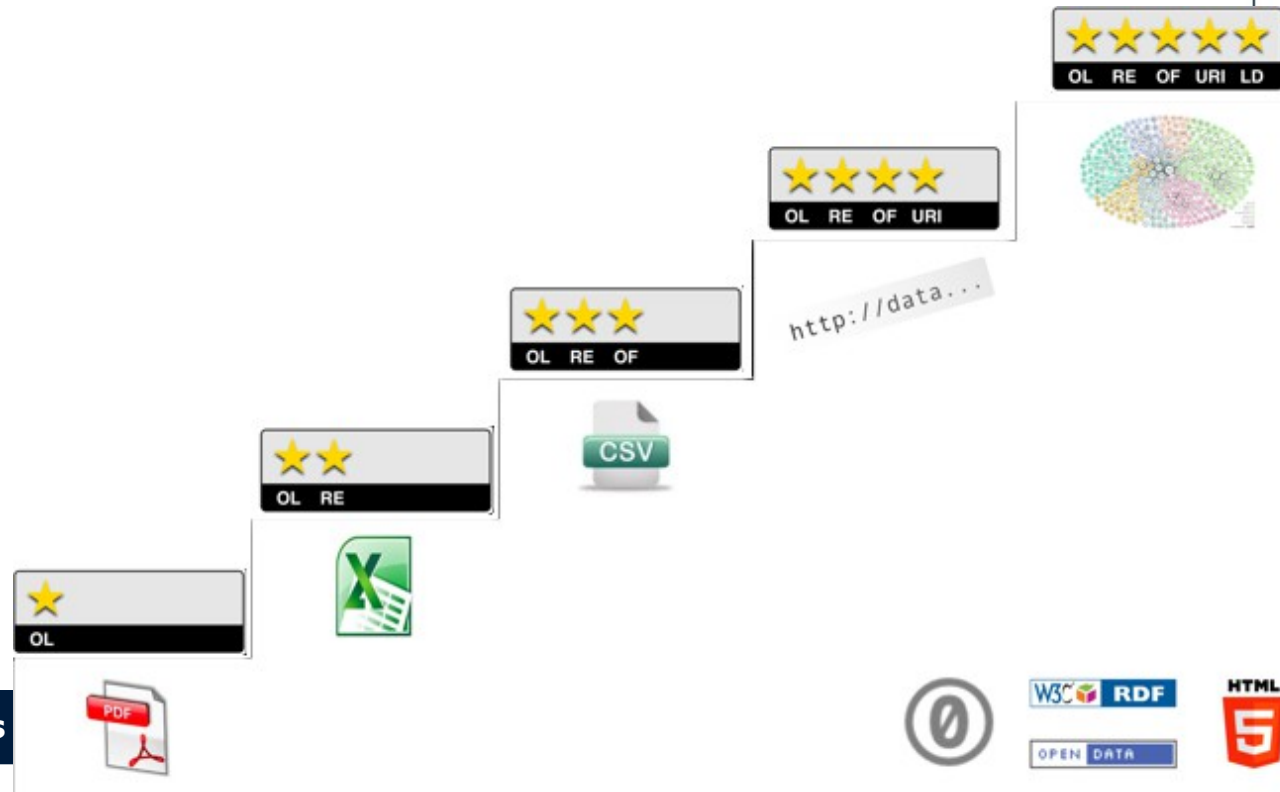
- Data sharing: **copyright**
 - Can data be copyrighted?
 - Data Source: you, lab mates, others?
 - Ownership
- Data sharing: **license**
 - Do you want your data reused?
 - And be modified (format!)?
 - Commercial use?

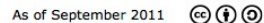
Terms of Use ▼

[Collapse Terms of Use \[-\]](#)

Day 6: Format? Why not SemWeb?

- 5 Star Open Data (5stardata.info)
open available, reusable, open format,
URIs (ontologies etc),
linked
data





Day 7: are people using your work?



← back to profile

ChemPedia as RDF

(2013) *figshare*.

highly viewed by scholars

 70 figshare views 

97 - 100 percentile  of datasets published in 2013

highly discussed by public

 3 tweets 

97 - 100 percentile  of datasets published in 2013

discussed by scholars

 1 figshare share 

97 - 100 percentile  of datasets published in 2013

discussed by public

 1 blog post 

97 - 100 percentile  of datasets published in 2013



DataCite

Helping you to find,
access, and reuse data

What is DataCite?

We are a not-for-profit organisation formed in London on 1 December 2009. Our aim is to:

- establish easier access to research data on the Internet
- increase acceptance of research data as legitimate, citable contributions to the scholarly record
- support data archiving that will permit results to be verified and re-purposed for future study.

These goals are laid down in the DataCite [statutes](#).

Day 8: back to step 0

- Take feedback (“peer review”), study new uses
- Plan your next study



CC-BY
frankensteinnn@flickr