Université
de Paris

ifb
INSTITUT FRANÇAIS
DE BIOINFORMATIQUE

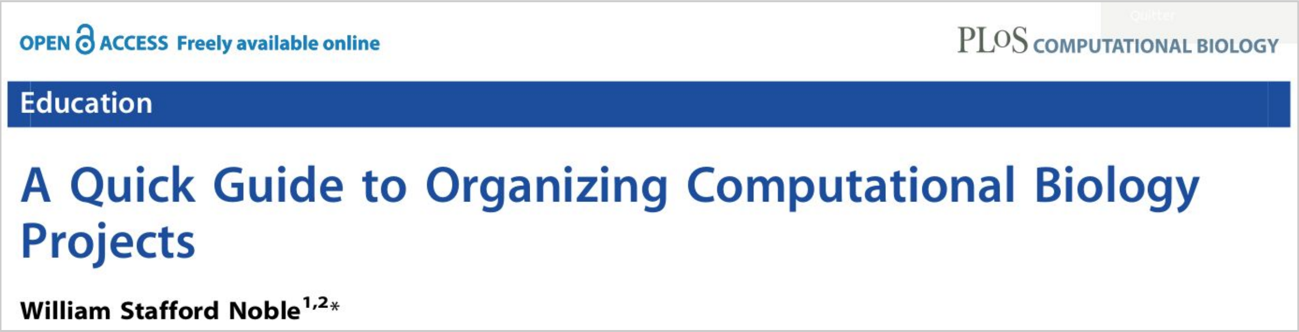
Bonnes pratiques pour organiser vos projets en bioinfo

DUBii 2020

Hélène Chiapello, Pierre Poulain

https://du-bii.github.io/module-1-Environnement-Unix/

Deux références



OPEN ACCESS Freely available online

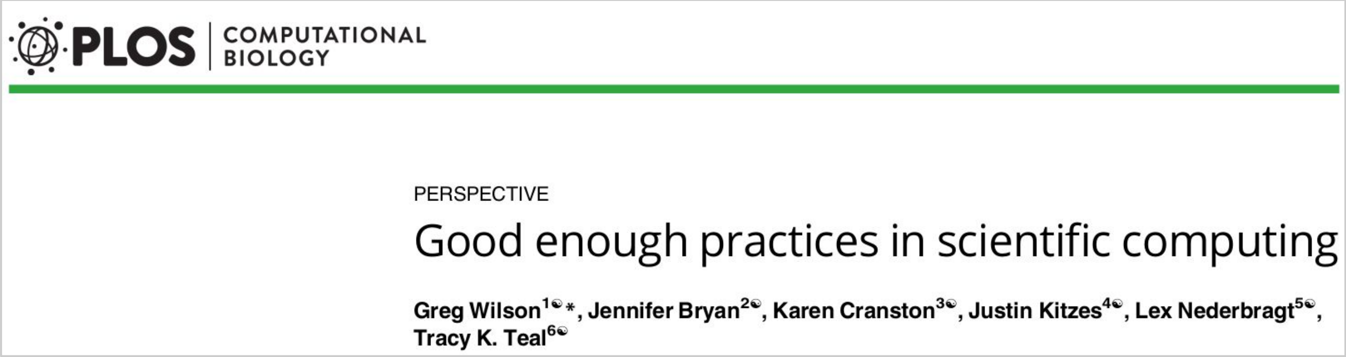
PLoS COMPUTATIONAL BIOLOGY

Education

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble[1,2]*

Noble, PLoS Comput Biol, 2009
DOI 10.1371/journal.pcbi.1000424



PLOS | COMPUTATIONAL BIOLOGY

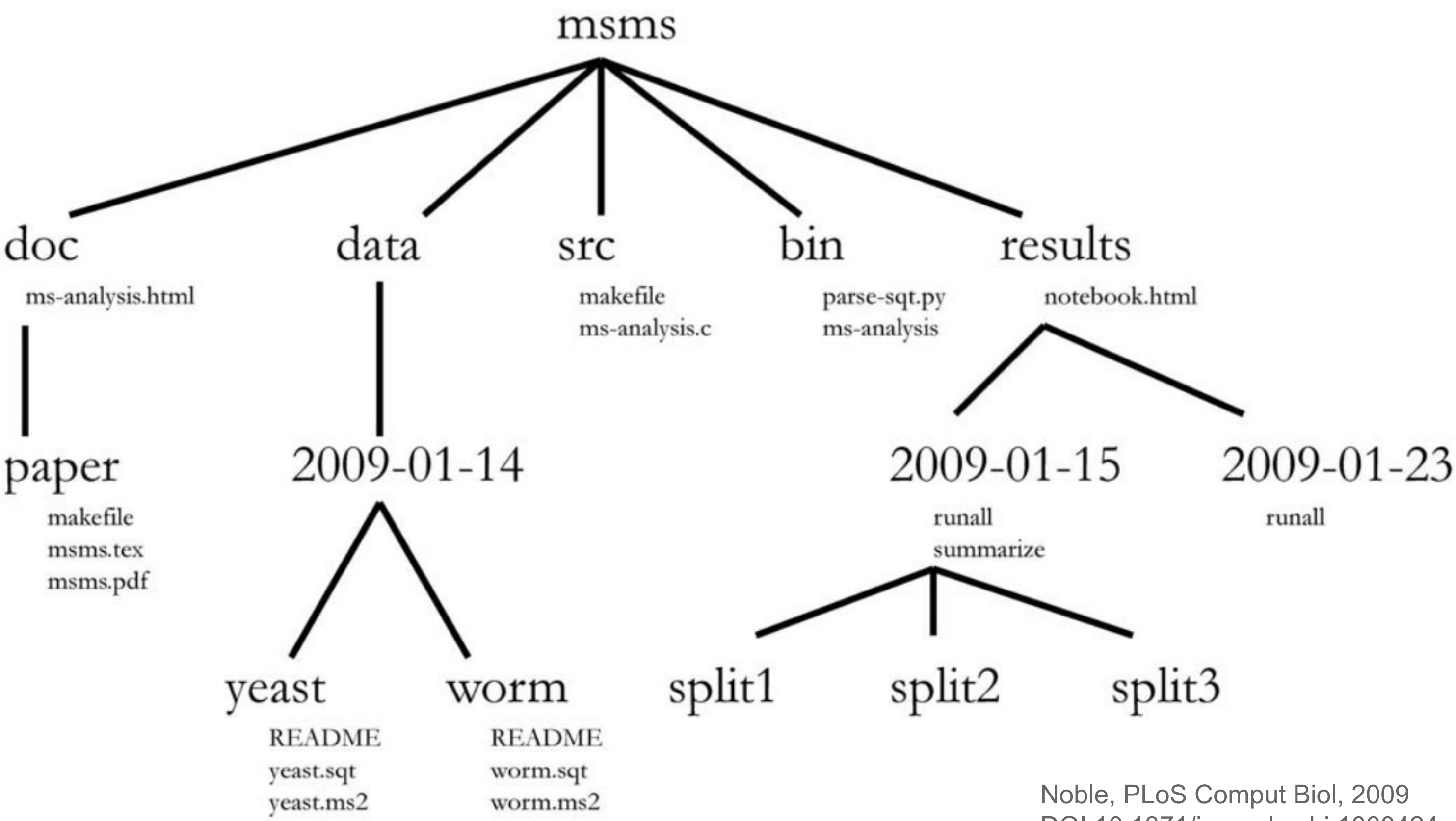
PERSPECTIVE

Good enough practices in scientific computing

Greg Wilson[1☯]*, Jennifer Bryan[2☯], Karen Cranston[3☯], Justin Kitzes[4☯], Lex Nederbragt[5☯], Tracy K. Teal[6☯]

Wilson, PLoS Comput Biol, 2017
DOI 10.1371/journal.pcbi.1005510

Un exemple d'organisation



Noble, PLoS Comput Biol, 2009
DOI 10.1371/journal.pcbi.1000424

Noms de fichiers et répertoires

Pas d’espace
_ ou - pour séparer les « mots »

Pas de caractères spéciaux

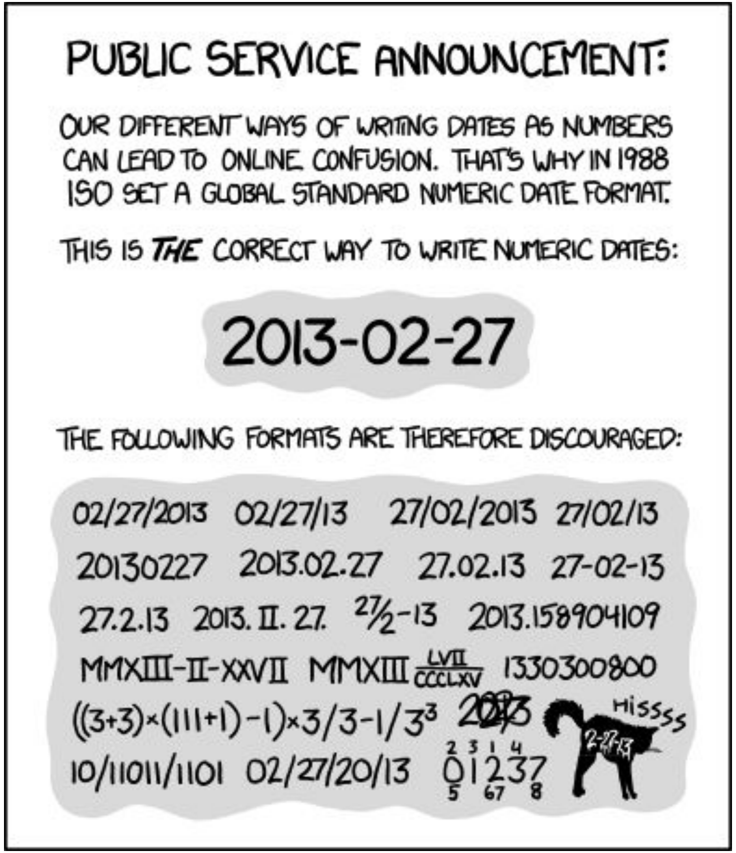
Format de date

ISO 8601 ?

Format de date

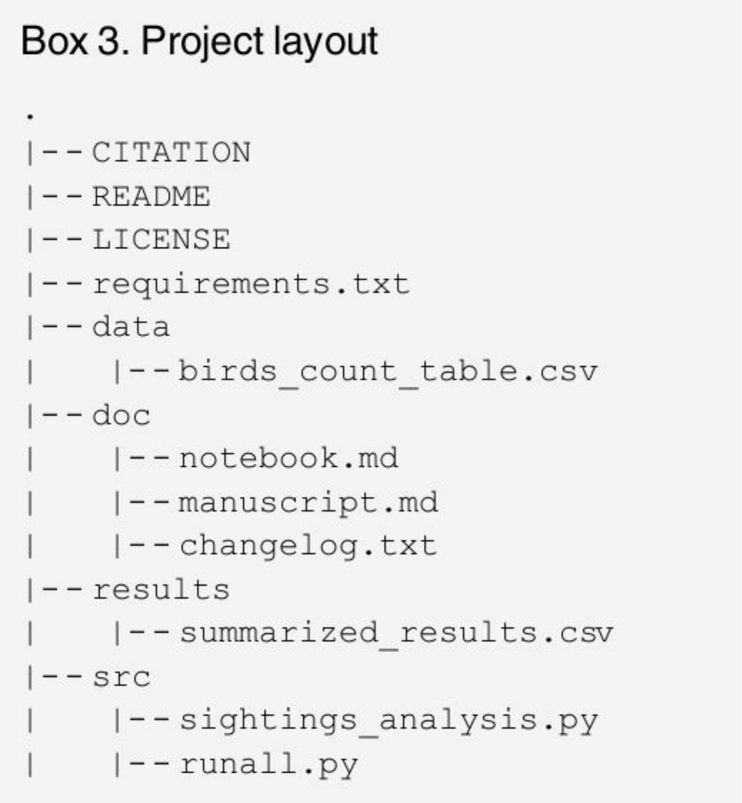
So what's your idea of a perfect date?
YYYY-MM-DD
I find other formats a bit confusing

Mahdi Yusuf / @myusuf3
https://mobile.twitter.com/myusuf3/status/865722106071453696


PUBLIC SERVICE ANNOUNCEMENT:
OUR DIFFERENT WAYS OF WRITING DATES AS NUMBERS CAN LEAD TO ONLINE CONFUSION. THAT'S WHY IN 1988 ISO SET A GLOBAL STANDARD NUMERIC DATE FORMAT.
THIS IS THE CORRECT WAY TO WRITE NUMERIC DATES:
2013-02-27
THE FOLLOWING FORMATS ARE THEREFORE DISCOURAGED:
02/27/2013 02/27/13 27/02/2013 27/02/13
20130227 2013.02.27 27.02.13 27-02-13
27.2.13 2013. II. 27. 27/2-13 2013.158904109
MMXIII-II-XXVII MMXIII LVII/CCCLXV 1330300800
((3+3)×(111+1)-1)×3/3-1/3³ 2013
HISSSSS
2-27-13
10/11011/1101 02/27/20/13 01237
2 3 1 4
5 67 8

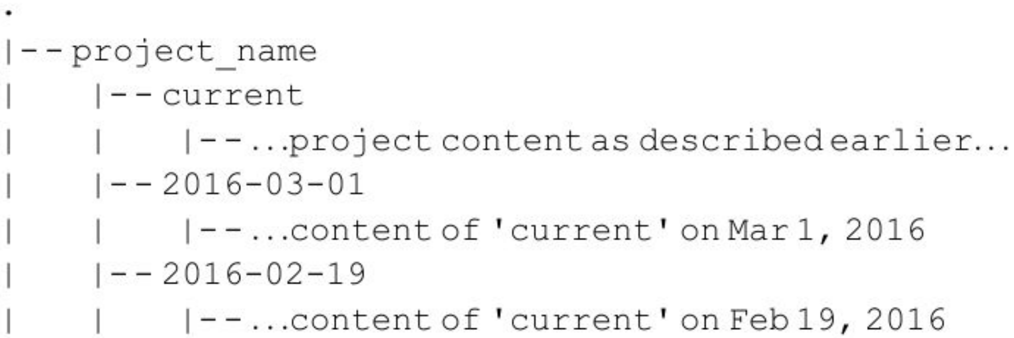
XKCD, ISO 8601
https://xkcd.com/1179/

Un autre exemple d’organisation



Box 3. Project layout

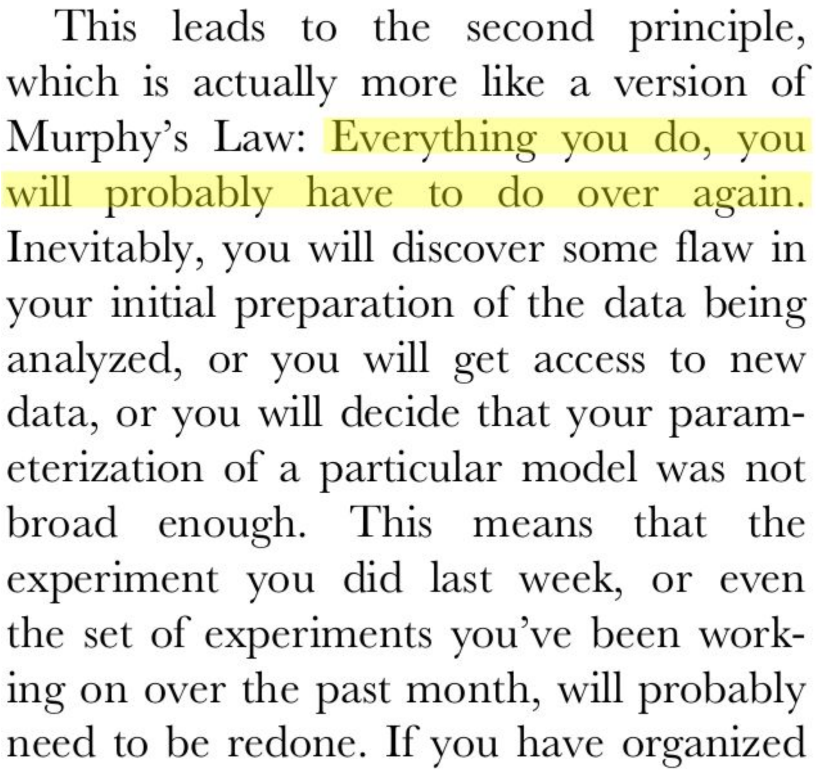
```
.
|--CITATION
|--README
|--LICENSE
|--requirements.txt
|--data
|    |--birds_count_table.csv
|--doc
|    |--notebook.md
|    |--manuscript.md
|    |--changelog.txt
|--results
|    |--summarized_results.csv
|--src
|    |--sightings_analysis.py
|    |--runall.py
```



```
.
|--project_name
|    |--current
|    |    |--...project content as described earlier...
|    |--2016-03-01
|    |    |--...content of 'current' on Mar 1, 2016
|    |--2016-02-19
|    |    |--...content of 'current' on Feb 19, 2016
```

Wilson, PLoS Comput Biol, 2017
DOI 10.1371/journal.pcbi.1005510

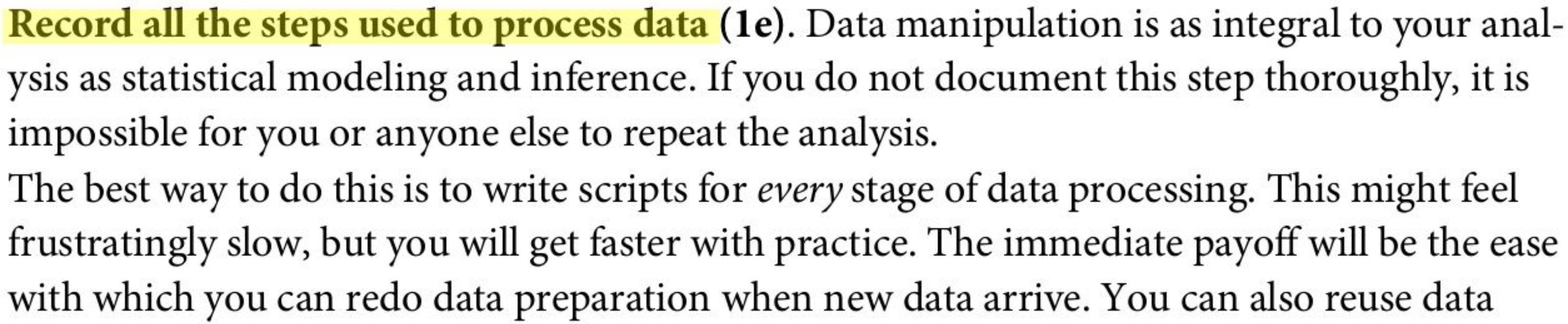
Quelques conseils



This leads to the second principle, which is actually more like a version of Murphy's Law: Everything you do, you will probably have to do over again. Inevitably, you will discover some flaw in your initial preparation of the data being analyzed, or you will get access to new data, or you will decide that your parameterization of a particular model was not broad enough. This means that the experiment you did last week, or even the set of experiments you've been working on over the past month, will probably need to be redone. If you have organized

Noble, PLoS Comput Biol, 2009
DOI 10.1371/journal.pcbi.1000424

Quelques conseils



Record all the steps used to process data (1e). Data manipulation is as integral to your analysis as statistical modeling and inference. If you do not document this step thoroughly, it is impossible for you or anyone else to repeat the analysis.

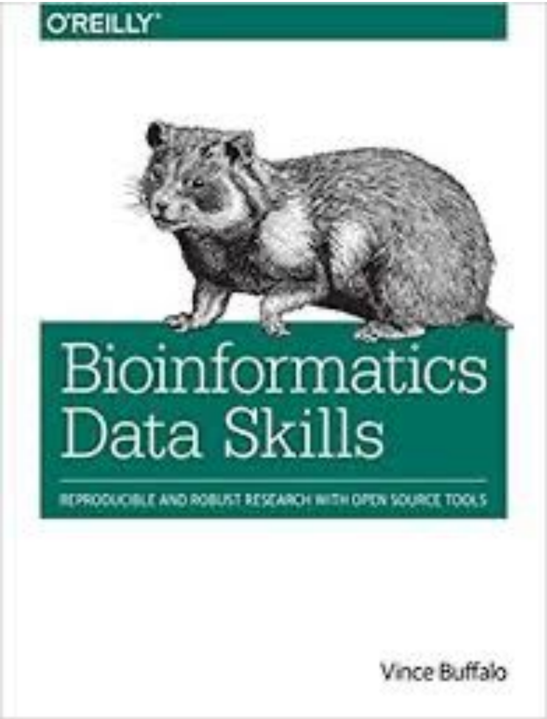
The best way to do this is to write scripts for *every* stage of data processing. This might feel frustratingly slow, but you will get faster with practice. The immediate payoff will be the ease with which you can redo data preparation when new data arrive. You can also reuse data

Wilson, PLoS Comput Biol, 2017
DOI 10.1371/journal.pcbi.1005510

Des conseils, encore !

Adopter des pratiques **robustes** et **reproductibles**

- Code
  - ○ Lisible
  - ○ Documenté
  - ○ Utiliser des librairies existantes dès que c’est possible
  - ○ Versionné et partagé
- Données
  - ○ Versioning
  - ○ Fichier de données : en read-only
  - ○ Plans de Gestion de Données (PGD)
- Code + données + résultats
  - ○ Gestionnaires de workflows
  - ○ Notebook


O'REILLY
Bioinformatics Data Skills
REPRODUCIBLE AND ROBUST RESEARCH WITH OPEN SOURCE TOOLS
Vince Buffalo