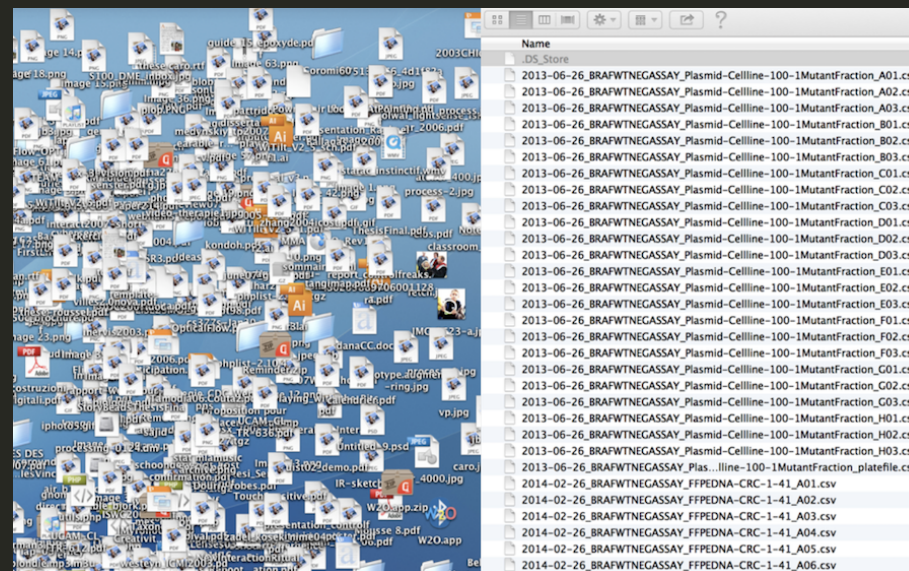


Le stockage de données

Facundo Muñoz
facundo.munoz@cirad.fr



Comment sont typiquement stockées vos données ?

...

1. Quel système de stockage ?

- ☐ Base de données relationnelle (E.g. MySQL, Oracle, MS Access)
- ☐ Feuille de calcul (E.g. MS Excel, LibreOffice Calc)
- ☐ Fichiers de texte (E.g. CSV, JSON)

Comment sont typiquement stockées vos données ?

7 Responses 02:38 Average time to complete **Active** Status

1. Quel système de stockage ?


● Base de données relationnelle ...	0
● Feuille de calcul (E.g. MS Excel...	6
● Fichiers de texte (E.g. CSV, JSON)	1
● Autre	0



Quels problèmes avez-vous rencontré concernant les
fichiers de données ?

Quelques problèmes courants

- J'ai plusieurs **copies des données** mais je ne suis pas sûr quelle est la dernière version.
- J'ai les données, mais je ne suis pas sûr d'avoir la **dernière version**.
- J'ai un fichier de données mais je ne me souviens pas de ce qu'il **contient**.
- J'ai des données, mais je ne les **retrouve** pas.
- Je ne sais pas comment interpréter certaines des **variables**.
- J'ai corrigé des erreurs et on m'a envoyé une **nouvelle version** basée sur les données originales



Pertes de temps
Erreurs inaperçues

Waste. Of Time.

Quelques principes

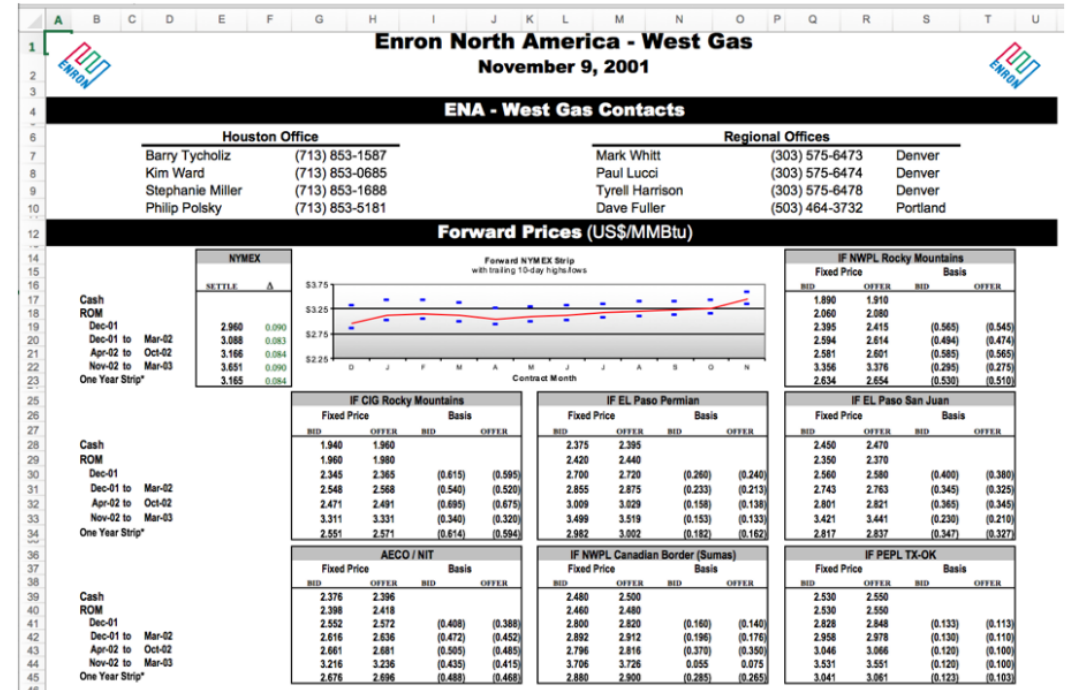
1. ~~Limiter~~ Éviter la **duplication** des données
2. Gérer les **versions**
3. **Documenter** les données (méta-données)
4. Adopter une convention pour les **noms des fichiers**

Entamez un document de *guidelines* à respecter

Feuilles de calcul

Combinent :

- stockage de données
- entrée de données
- visualisation (tables, format)
- analyses (formules, conditions, résultats, résumés, etc.)
- figures



Les besoins pour l'**entree**, le **stockage** et la **visualisation** de données sont fondamentalement différents

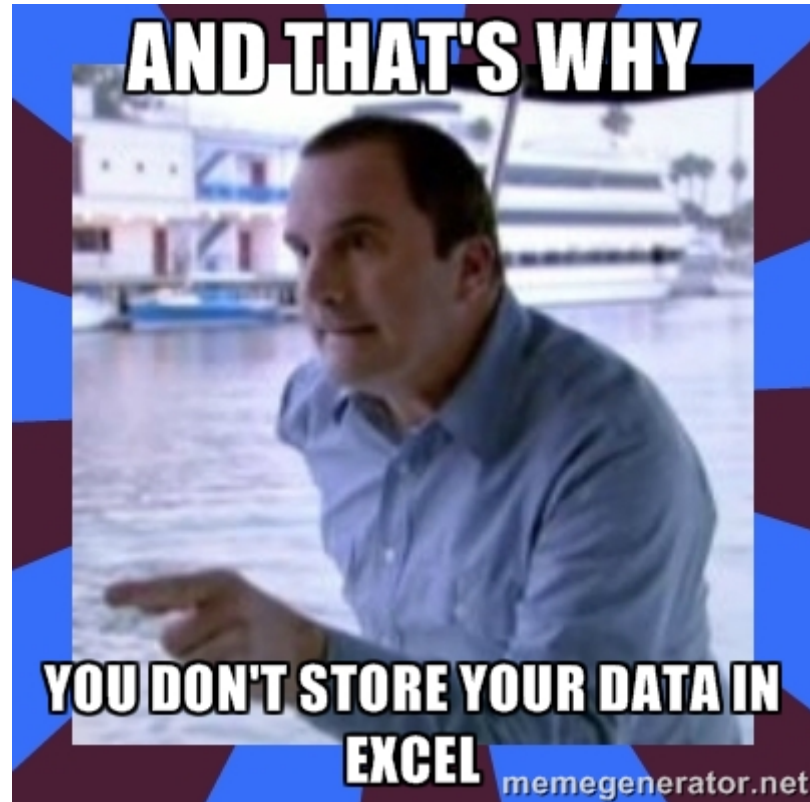
Histoires d'horreur

- MS Excel interprète les dates et les garde internement comme un nombre... avec des conventions différents pour Mac et Windows



- MS Excel interprète *automatiquement* certains textes comme des dates. E.g. le symbole pour le gène "Oct-4" peut être écrasé sans notification.

Un étude de 2016 à trouvé ce type d'**erreurs dans 20 %** des listes des gènes publiées.



Vous allez utiliser les outils que vous **maîtrisez**, pas forcément ceux dont vous avez **besoin**

Vous pouvez les utiliser, mais pensez aux **principes** de gestion de données

Noms des fichiers (et variables)

*‘There are only two
hard things in
Computer Science:
cache invalidation and
naming things.’*

- *Phil Karlton*

"FINAL".doc



FINAL.doc!



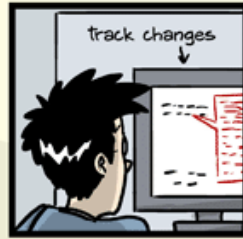
FINAL_rev.2.doc



FINAL_rev.6.COMMENTS.doc



FINAL_rev.8.comments5.
CORRECTIONS.doc



FINAL_rev.18.comments7.
corrections9.MORE.30.doc



FINAL_rev.22.comments49.
corrections.10. #@\$%WHYDID
ICOMETOGRADSCHOOL????.doc

JORGE CHAM © 2012

No

- myabstract.docx
- Les noms de fichiers d'Agnès utilisent les espaces et la ponctuation.xlsx
- figure I.png
- fig 2.png
- JW7d^(2sl@nepassupprimerWx2*.txt

Oui

- 2014-06-08_abstract-for-sla.docx
- les-fichiers-dagnes-vont-mieux.xlsx
- fig01_nuage-points-taille-vs-interet.png
- fig02_histogramme-participation.png
- 1986-01-28_donnees-brutes-projet-code.txt

3 principes pour les noms (de fichiers)

1. Lisibles par une machine
2. Lisibles par les humains
3. Fonctionne bien avec l'ordre d'affichage


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. $27\frac{1}{2}$ -13 2013.158904109
 MMXIII-II-XXVII MMXIII $\frac{LVII}{CCCLXV}$ 1330300800
 $((3+3) \times (111+1) - 1) \times 3 / 3 - 1 / 3^3$ 2013
 10/11011/1101 02/27/20/13 $\begin{matrix} 2 & 3 & 1 & 4 \\ 0 & 1 & 2 & 3 & 7 \\ 5 & 6 & 7 & 8 \end{matrix}$ 

Excellents noms de fichiers

-  2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
-  2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
-  2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
-  2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_platefile.csv
-  2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
-  2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
-  2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
-  2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv

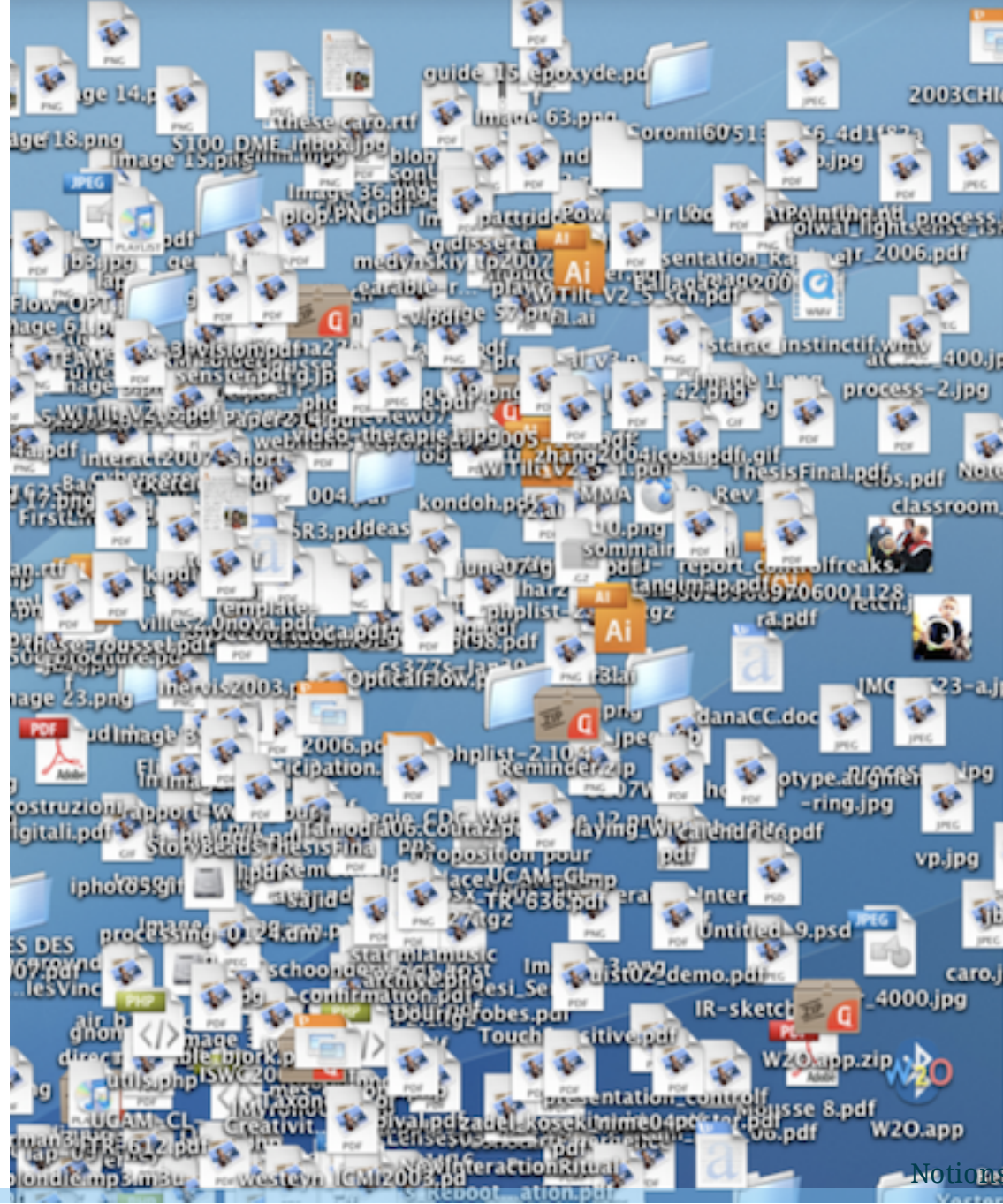
datacarpentry

Nommer des fichiers (et des variables)

Des outils d'aide pour développer une **nomenclature** de fichiers adaptée à vos besoins :

- Kristin Briney (2020) [File naming convention worksheet](#)
- [Minnesota Historical Society](#)
- [University of Edinburgh, Records Management](#)

Organiser ses fichiers



Name

.DS_Store

2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_A03.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_B03.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_C03.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_D03.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_E03.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_F03.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_G03.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H01.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H02.csv
 2013-06-26_BRAFWTNEGASSAY_Plasmid-Cellline-100-1MutantFraction_H03.csv
 2013-06-26_BRAFWTNEGASSAY_Plas...line-100-1MutantFraction_platefile.csv
 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A01.csv
 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A02.csv
 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A03.csv
 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A04.csv
 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A05.csv
 2014-02-26_BRAFWTNEGASSAY_FFPEDNA-CRC-1-41_A06.csv

Mon structure de projet basique

```
.  
├── data/  
├── doc/  
├── reports/  
├── src/  
└── Readme.md
```

Principes

1. Je ne modifie **jamais** les fichiers dans data
2. Toute la **documentation** à lire dans doc (e.g. description des données, articles, etc.)
3. Le travail d'**analyse** se passe dans src
4. Les **résultats** dans reports

Références

- Kristin Briney (2020) [File naming convention worksheet](#)
- Data Carpentry (2018) [lesson on file organisation](#)
- Karl W. Broman & Kara H. Woo (2018). Data organisation in Spreadsheets. *The American Statistician*, 72:1, 2-10, DOI: [10.1080/00031305.2017.1375989](#)

Merci!

Diapositives créées à l'aide du package R **xaringan**.

En s'appuyant sur **remark.js**, **knitr**, et **R Markdown**.



Ce(tte) œuvre est mise à disposition selon les termes de la **Licence Creative Commons Attribution - Partage dans les Mêmes Conditions 4.0 International**.