

Git n'oublie pas, ce qui en est l'intérêt et le danger

Un dépôt se souvient de chaque version de chaque fichier qu'il a jamais contenu, ce qui est exactement ce qu'on veut pour du code et exactement ce qu'on ne peut pas permettre pour une liste de bénéficiaires. Supprimer le fichier dans un commit ultérieur ne le retire pas.

Pierre Robentz Cassion

Leçon 2 sur 8

Chaînes d'analyse reproductibles — Le projet sur le disque

Ce que couvre cette leçon

- Le commit qu'on ne peut pas reprendre
- Le `.gitignore` qu'on écrit avant le premier commit
- Ce que « données personnelles » signifie ici, précisément
- La vérification qui s'exécute avant le commit
- Que faire si c'est déjà arrivé
- Les commits sur un travail d'analyse
- Ce que devrait contenir un dépôt d'analyse de programme
- Rapportez-le en entier
- La suite

Le commit qu'on ne peut pas reprendre — Shell

```
git add data/raw/beneficiaries.csv
git commit -m "add survey data"
# ... realised three days later ...
git rm data/raw/beneficiaries.csv
git commit -m "remove data"
```

Le commit qu'on ne peut pas reprendre

- Le fichier est toujours dans le dépôt — Il est dans l'historique, dans chaque clone, sur le dépôt distant, et dans la...
- Sur un dépôt contenant des données de protection ou de santé, c'est une divulgation — Non une erreur à corriger plus...
- D'où une règle préventive plutôt que corrective — On ne peut pas dé-commiter des données personnelles; on ne peut que...

Le .gitignore qu'on écrit avant le premier commit — Exemple (suite)

```
# Raw data. Nothing under here is ever committed.
data/raw/*
!data/raw/.gitkeep
!data/raw/CHECKSUMS

# Derived data, regenerated by the pipeline
outputs/

# Credentials, in every form they arrive in
.env
*.pem
*_secret*
config.local.*

# Editor and language noise
.Rhistory
```

Le `.gitignore` qu'on écrit avant le premier commit — Exemple (suite)

```
.RData  
__pycache__/  
*.pyc  
.DS_Store  
.ipynb_checkpoints/
```

Le `.gitignore` qu'on écrit avant le premier commit

- Écrivez-le en premier, committez-le en premier — Un `.gitignore` ajouté après le premier commit est un `.gitignore...`
- Les négations comptent — `data/raw/*` exclut tout, puis `!data/raw/CHECKSUMS` laisse revenir exactement le fichier qui...

Ce que « données personnelles » signifie ici, précisément

Paraît anonyme	Ne l'est pas, parce que
Un identifiant de cas	Il se joint à un système de gestion de cas qui a le nom
Les coordonnées GPS d'un ménage	C'est une adresse
Date de naissance plus commune plus sexe	Trois champs identifient la plupart des gens dans une petite commune
La photographie d'un formulaire	Chaque champ qui y figure
Une colonne de notes en texte libre	Elle contient ce que le travailleur social a tapé

Ce que « données personnelles » signifie ici, précisément

- Le cours de protection a établi l'arithmétique — une ventilation à quatre entrées d'un jeu de données anonyme a produit. . .
- Le test n'est donc pas « y a-t-il une colonne de noms » — C'est de savoir si une combinaison quelconque de ce que. . .

La vérification qui s'exécute avant le commit — Shell

```
# .git/hooks/pre-commit -- or the pre-commit framework, or a CI step
if git diff --cached --name-only | grep -qE '^data/raw/'; then
    echo "Refusing to commit anything under data/raw/" >&2
    exit 1
fi
```

La vérification qui s'exécute avant le commit — En R

The same check as an R function, run by whatever CI you have.

La vérification qui s'exécute avant le commit

- Un crochet qui refuse vaut mieux qu'une politique qui rappelle — Le mode d'échec est quelqu'un de pressé à 18 heures,...
- Ajoutez une recherche des formes que prennent les secrets — parce qu'un jeton collé dans un script est l'autre cas. . .

Que faire si c'est déjà arrivé

- Un : signalez-le — Votre organisation a une procédure d'incident de données
- Deux : renouvelez tout ce qui était un identifiant — Un jeton commité est compromis définitivement ; le supprimer n'y. . .
- Trois : réécrivez l'historique — avec `git filter-repo` ou BFG, et forcez la poussée
- Quatre : supposez qu'il a été récupéré — Sur un dépôt public, supposez que le fichier a été téléchargé et copié, et. . .

Que faire si c'est déjà arrivé — Shell

```
git filter-repo --path data/raw/beneficiaries.csv --invert-paths
```

Que faire si c'est déjà arrivé

- La troisième étape est la seule qui ressemble à une correction et c'est la moins importante

Les commits sur un travail d'analyse

- Commitez le code et la sortie ensemble quand la sortie est commitée — Une figure et le script qui l'a produite. . .
- Écrivez le pourquoi dans le message — `corrige les noms de communes` est ce que le diff montre déjà

Les commits sur un travail d'analyse — Shell

```
git commit -m "Group four spellings of Nord-Ouest before computing coverage"
```

Ungrouped, the worst district splits into fragments of 727, 33, 22 and 20 households and none of them looks alarming. The 20-household fragment shows 0% open defecation, which a district table would report as a solved problem."

- Une branche par question d'analyse, non par journée — Une branche qui répond à « la couverture diffère-t-elle par. . . »

Ce que devrait contenir un dépôt d'analyse de programme — Exemple

project/	
.gitignore	committed first
README.md	what it produces, how to run it, what to know
uv.lock	the environment, pinned
src/	every transformation
tests/	the checks that fail loudly
data/	
raw/	.gitkeep and CHECKSUMS only
reference/	committed: thresholds, admin codes, lookup tables
outputs/	ignored

Ce que devrait contenir un dépôt d'analyse de programme

- Les tables de référence sont commitées et les données brutes non — et la ligne qui les sépare est de savoir si le...

Rapportez-le en entier — Exemple

Data handling in this analysis

No raw data is committed. `data/raw/` is excluded by `.gitignore`, with only a `CHECKSUMS` file tracked so the exports used can be identified.

A pre-commit hook refuses any staged file under `data/raw/` and a CI step scans for credential patterns.

Reference tables (WHO growth standards, IPC thresholds, admin codes) are committed; they describe no individual.

The repository is private and access is limited to the M&E team. Raw exports are held in [the organisation's storage], not here.

- La dernière ligne est celle qu'on oublie et celle qu'un audit demande — Dire où *sont* les données compte autant que...

- Du code en gestion de versions ne reproduit toujours pas un nombre si l'environnement diffère.

Lire la leçon complète, avec le code exécutable

```
https://data-analysis.cassion.dev/fr/cours/reproducible-workflows/  
repro-02-git-and-what-ever-goes-in
```