cahiers numérique

Damien Belvèze

A partir de la présentation d'Aurélien Tabard [[@TabardOSW2021Cahiersnumeriques2021]]

[[20210623_notebooks_tabard.pdf]]

intérêts des carnets de laboratoire : - documenter pour soi-même - documenter pour assurer la [[reproductibilité]] de l'expérience (cahiers computationnels comme les Jupyter Notebooks)

Induisent deux styles d'écriture différents

choisir le bon produit en matière de cahier numérique

- bien définir le périmètre des utilisateurs de la solution : à quel niveau on installe l'outil pour que le maximum de chercheurs en dispose.
- ergonomie : le passage du papier au numérique doit se faire de manière très fluide.
- gouvernance des données : conservation locale des données. devrait nous détourner des solutions en ligne. [[Mbook]] et [[ElabFTW]] nous permet de mettre l"outil en local
- modalités d'export en cas de migration de solution ou bien pour installer la solution en local ou changer de serveur
- coût de la solution (par exemple : Mbook :120 euros par utilisateur/an) ; difficilement accessible pour de petites institutions qui n'ont pas la masse critique pour négocier

Une solution libre [[ElabFTW]] installée par exemple sur un serveur de Lyon (projet Datacc)

Le CNRS pourrait fournir des lignes budgétaires aux labos pour qu'ils puissent se doter de cahiers numériques. Le CNRS fournira une liste de critères pour choisir sa solution mais ne fera pas de recommandation (chaque discipline a ses propres besoins)

cahiers hybrides

thèse d'Aurélien Tabard sur les carnets de labo (expérience de terrain auprès des biologistes entre 2005 et 2009) pas de norme, peu de recul sur le sujet. Test de la solution PRISM dans différents laboratoires

PRISM

PRISM permet de conserver une version numérique de notes manuscrites sur le papier (couplé à un OCR).

intégration de pages web intégration de flux (mails) cahier à dimension chronologique

cahier papier : contraintes matérielles liées à la structure de la page. Les cahiers papier poussent vers davantage de discipline dans la prise de notes quotidiennes (on ne peut pas rajouter facilement un paragraphe) cahier numérique : possibilité de rajouter des paragraphes, notes réécrites régulièrement.

Originellement il y a beaucoup de cahiers pour un même projet de labo.

intérêt réflexif de la prise de notes.

Adapter les cahiers au terrain

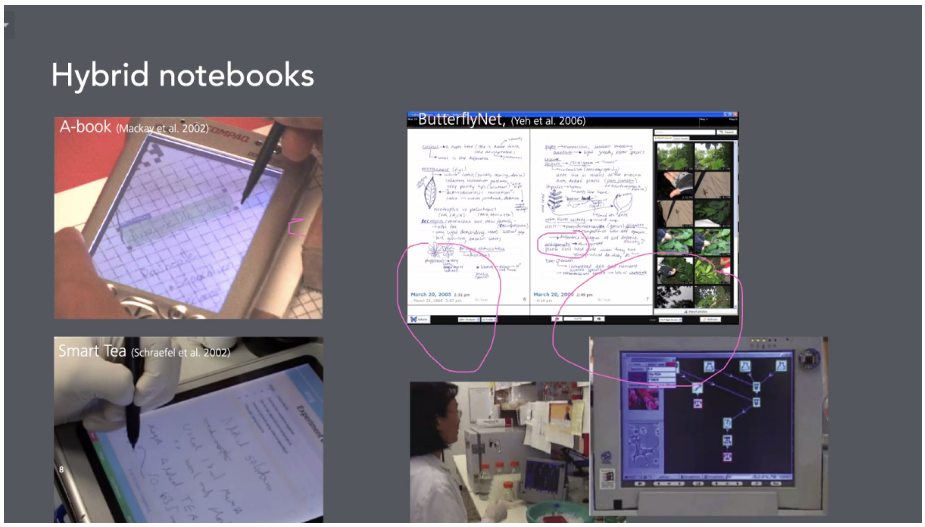

Figure 1: cahiers de terrain

aide à la capture et à l'accès aux informations

surface interactive sur la paillasse + caméra pointée sur la paillasse. Bouton pour prendre des snapshots facilement. système de prise de notes collaboratives.

Avec le [[numérique]] on peut basculer d'une prise de notes individuelles à une activité de prise de notes collaboratives.

cahiers computationnels

mélange de texte, d'analyse, de code et de visualisations. approche de la programmation lettrée : fournir suffisamment d'explications sur son code pour le rendre compréhensible et partageable.

[[Jupyter]] est encore le plus utilisé. ça permet de faire des analyses reproductibles pour avoir du code et des analyses vérifiables, du moins dans le discours.

Etude faite en 2017 : récupération de tous les cahiers disponibles sur Github (plus d'un million de cahiers) Sélection de 150 cahiers connectés à des expériences scientifiques. Entretiens avec 150 chercheurs pour comprendre leur usage de ces cahiers.

structuration des cahiers

plusieurs cellules : titre, texte (markdown) dont headers, code ([[Python]] par exemple), sortie graphique La plupart des cahiers en 2017 avaient entre 10 et 50 cellules.

La plupart des cahiers avaient 100 lignes de code.

un quart des cahiers en 2017 étaient dénués de texte.

Les cahiers Jupyter sont fréquemment utilisés comme tutoriels dans des cours de machine learning.

L'intérêt d'un Jupyter notebook est de présenter une organisation linéaire de l'information (notamment du code). On ne va pas se servir de ce type d'outil pour créer un logiciel.

questions relatives à la reproductibilité

recherche dans des repositories de github des cahiers qui avaient dans leur fichier READ.me un lien comportant [[DOI]] ou Arxiv

sélection de 145 cahiers liés à 50 projets

dont une quarantaine utilisés pour des tutoriels

50 [[articles scientifiques|articles scientifiques]] qui ne disposaient que du code en lien avec les figures : montrer que les figures ont été faites directement à partir des données sans risque de problème de conversion opaque.

Seul 77% des cahiers comportaient du texte, souvent réduits par ailleurs aux headers.

Entretiens complémentaires avec 15 chercheurs de San Diego (Californie)

Les cahiers sont toujours considérés par les personnes interrogées comme des documents personnels, des documents qu'on rédige pour soi. On les met à disposition à des collègues proches mais guère plus. Il est courant de demander la permission avant de consulter le cahier d'un collèque. Quand on fait de la recherche on est dans un moment de fragilité par rapport aux standards de la science ([[intégrité scientifique]]) tels qu'ils sont énoncés (et pas forcément autant pratiqués).

Impression générale que ses cahiers sont bordéliques : inconfort entre ce qu'on aimerait avoir dans son cahier et ce qui s'y trouve réellement. Cette documentation est souvent faite mais de manière laborieuse à des moments spécifiques. Selon l'objectif de départ (perspective d'illustration de cours ou de communication) on a des écritures différentes : soit on documente pour soi, soit on explique pour les autres.

Ces moments de nettoyage des cahiers, ce sont aussi des moments réflexifs importants pour le chercheur.

difficultés d'usage

La première génération d'outils n'était pas mature ce qui a découragé les premiers utilisateurs à s'en servir. Notamment les mises à jour de Java rendaient leur usage assez difficile.

conclusion

tensions entre les [[usages]] : - explication ou documentation - exploration individuelle ou communication - assez peu d'aspect narratif dans ces cahiers.

Tension entre exploration de données et explication, cette tension va freiner la progression de la narration, l'une des conditions de la [[reproductibilité]] de la science.

Est-ce qu'on est train de documenter une enquête en construction ou bien un résultat final ?

Est-ce qu'il s'agit de reproduire ça pour soi-même ou bien pour prouver le bien-fondé de la démarche ?

Est-ce qu'on le fait pour d'autres membres de son équipe ou bien pour les [[Révision par les pairs|reviewers]] ?

La reproductibilité est-elle un idéal pour l'ensemble de la science ? Pour les Sciences Humaines, ne vaut-il pas mieux insister sur l'esprit critique démontré par le chercheur au cours ?

On peut fabriquer de la fraude avec ses cahiers. Ce thème de la reproductiblité est très connecté à la concurrrence entre chercheurs. Le problème fondamental c'est ce problème de compétition. Le carnet de labo, en tant qu'artefact technique, n'est pas un rempart vraiment satisfaisant par rapport à la crise de la reproductibilité de la science qui a pour origine surtout un problème social que rencontrent les chercheurs aujourd'hui.

Le cas des cahiers numériques à Lyon 1

Présentation de DATACC (2019-222)

Les cahiers sont adossés à des outils de traitement de données plus performants qu'excel ou calc

La bibliothèque propose trois logiciels

cahiers numériques propriétaires : Mbook et Sciformation cahier numérique libre : eLabFtW (E-lab for the world)

Data's shameful neglect

Les chercheurs sont assez souvent livrés à eux-même pour la conservation ou la communication des données entre chercheurs La plupart des établissements de l'ESR en France n'ont pas de plan de gestion de données, la problématique ne survient qu'avec le dépôt d'un projet ANR

Bibliographie