Recherche reproductible chez Inria: quelle stratégie et quels outils?

Rémi Gribonval (EPC PANAMA, Rennes)
Emmanuel Vincent (EPI PAROLE, Nancy)

merci à tous les collègues qui ont apporté leur pierre
à ces quelques éléments de réflexion via des discussions stimulantes

Exemples célèbres

- **Théorème de Fermat**
  - Marge des *Arithmétiques* de Diophante, 1621
- **Mémoire de l'eau**
  - Nature, 1988
- **Cellules STAP**
  - Nature, 2014
- Et en informatique / mathématiques appliquées ?

Reproductibilité: de quoi parlons-nous ?

- **J’ai lu un article aux résultats intéressants mais:**
  - ✓ 1) je n'ai pas accès à tout ou partie des données
  - ✓ 2) l'étape X de l'algorithme n'est pas décrite dans l'article
  - ✓ 3) il manque la valeur du paramètre Y
  - ✓ 4) le critère d'évaluation n'est pas exactement spécifié
  - ✓ 5) je n'ai pas le temps de tout ré-implémenter!

… en pratique, souvent une combinaison des 5.

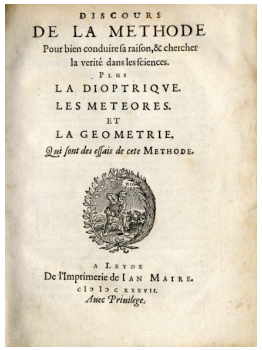
Plan de la présentation

- **Pourquoi ?**
  - Enjeux de la reproductibilité
  - Quelques initiatives en « sciences computationnelles »
- **Pourquoi pas ?**
  - Les (nombreux) obstacles à la reproductibilité
- **Comment ?**
  - Quelques pistes pour lever certains de ces obstacles

Pourquoi ? Enjeux

Enjeux de la reproductibilité

- **Socle fondamental de la méthode scientifique**
  - Fondements philosophiques (Descartes 1637, Popper 1934)


DISCOURS
DE LA METHODE
Pour bien conduire sa raison, & chercher la verité dans les sciences.
PLUS
LA DIOPTRIQVE.
LES METEORES.
ET
LA GEOMETRIE.
Qui sont des essais de cette METHODE.
A LEYDE
De l'Imprimerie de IAN MAIRE.
Avec Privilege.

- **Accélérateur de la recherche**
  - Facilite la reprise de travaux antérieurs
  - Accélère le progrès scientifique global
- **Facteur d'excellence**
  - Augmente la **visibilité** et la citation par les pairs
  - Articles avec des données en ligne ont **70% de citations en plus**

  Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308.
- **Garde-fou sociétal / déontologique / éthique**
  - Evite rejet dû au manque de transparence (climat, OGM, …)
- **Obligation contractuelle**
  - Obligatoire dans les projets NSF depuis 2011, bientôt à l'ANR

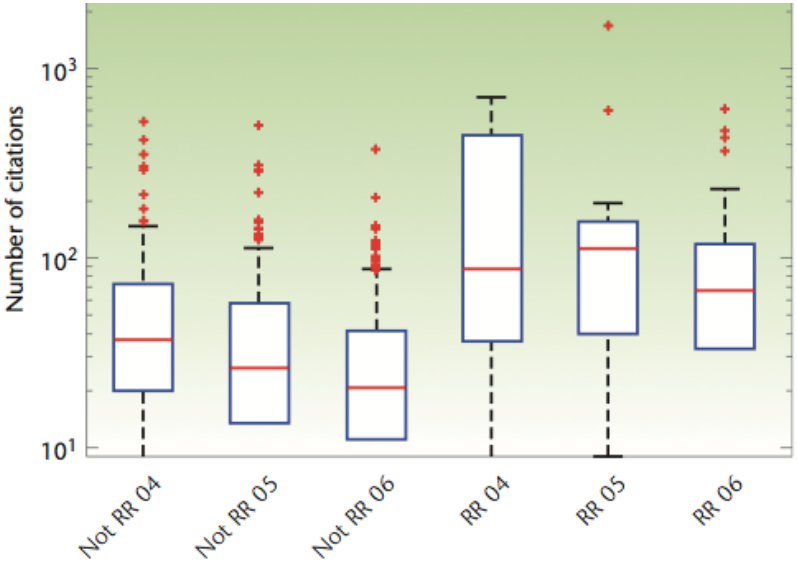
Quelques initiatives en “sciences computationnelles”

- Reproductibilité = article **+ code + données**
- **Campagnes et outils d’évaluation**
  - Données, tâches et critères de performance en traitement du signal audio (E. Vincent, R. Gribonval)
  - Frameworks logiciels COCO pour l’optimisation black-box (N. Hansen), OSA pour la simulation de réseaux (O. Dalle)
- **Plateformes de stockage**
  - Researchcompendia.org
- **Serveurs d’exécution**
  - Serveurs d'exécution pour la géométrie discrète (B. Kerautret), pour l'analyse de documents (B. Lamiroy)
  - Ipol.im
  - Runmycode.org

Reproductibilité et impact

Impact: Citations vs Code Sharing CiSE

Analysis of citations vs code availability for TIP 2004-2006 papers



P. Vandewalle, *Code sharing is associated with research impact in image processing*, IEEE Comp. in Science and Engineering, 2012.

Reproducible Research

Tiré d'une présentation de P.Vandewalle au Workshop « Reproductibilité en traitement du signal », GDR ISIS, 2014

Pourquoi pas ? Obstacles

Stratégiques
Légaux
Techniques

Objections souvent entendues

- **Aspects “stratégiques”:**
  - ✓ je veux garder une longueur d’avance
    - ✦ Code et données = avantages à ne pas donner à la concurrence
    - ✦ Préserver sa capacité à breveter, à transférer
  - ✓ ça demande du travail
    - ✦ Code pas propre = ne pas se ridiculiser …
    - ✦ Publication reproductible pas beaucoup plus valorisée que publication normale (recrutements, promotions)
- **Aspects légaux / éthiques**
  - ✦ Propriété intellectuelle
  - ✦ Droit à l’image et protection de la vie privée
  - ✦ Ethique (données biomédicales)

Obstacles

- **Aspects techniques / pratiques**

  - ✓ **Capacité de stockage** (données notamment)
    - ✦ Aspects techniques et financiers (si archivage volumineux)
  - ✓ **Pérennité**
    - ✦ **Durée de vie de l'URL** (personne / équipe / projet européen …)
    - ✦ Dépendance à **un environnement logiciel ou un format**
      - Langage (ex: version de Matlab),
      - Plate-forme (pour un exécutable binaire)
      - Format de fichier
    - ✦ Cf. aussi notion de "*software sustainability*"
      - http://www.software.ac.uk/about « développement » durable ;-)

  - ✓ **Maintien du lien** article / code/ données

Comment ? Pistes

Des outils existent!

Tools for Computational Science

- Dissemination Platforms:
  - ResearchCompendia.org
  - IPOL
  - Madagascar
  - MLOSS.org
  - thedatahub.org
  - nanoHUB.org
  - Open Science Framework
  - RunMyCode.org
- Workflow Tracking and Research Environments:
  - VisTrails
  - Kepler
  - CDE
  - Galaxy
  - GenePattern
  - Paper Mâché
  - Sumatra
  - Taverna
  - Pegasus
- Embedded Publishing:
  - Verifiable Computational Research
  - Sweave
  - Collage Authoring Environment
  - SHARE

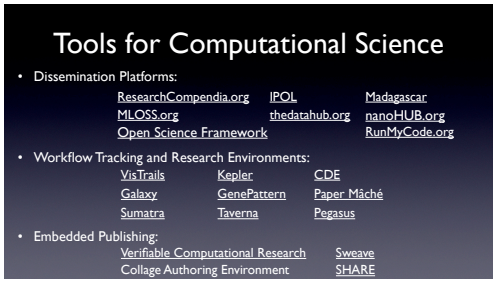
Tiré d'une présentation de V . Stodden au Workshop « Reproductibilité en traitement du signal », GDR ISIS, 2014

Fragmentation des ressources

- **Beaucoup d'outils**

  ✓ Traduisent mouvement de fond

  ✓ … mais spécifiques à telle ou telle communauté



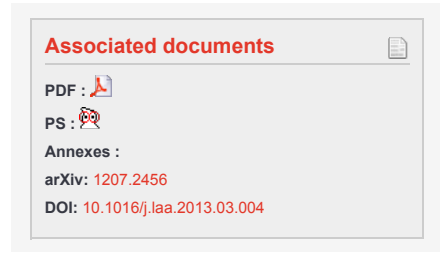
- **Analogie avec l'Open Access:**

  ✓ À une époque pas si lointaine, assuré uniquement via multitude de pages personnelles, pages d'équipe …

  ✓ Aujourd'hui, **outils institutionnels mutualisés:**

Infrastructures institutionnelles = soutien à la reproductibilité!

- **Quelques pistes immédiates:**
  - ✓ Pérennité URL + liens article-code-données
    - ✦ S'appuyer sur HAL !
    - ✦ Déjà possible (à la marge) via « ANNEXES »:
  - ✓ « Archivage » de données, de code
    - ✦ S'appuyer sur HAL ?
    - ✦ Faire évoluer types de fichiers acceptés ?
    - ✦ Mettre en valeur annexes reproductibles comme DOI ?



- **Plus ambitieux :**
  - ✓ Robustesse à l'environnement logiciel
    - ✦ S'appuyer sur grappes de machines virtuelles de type « plateforme de développement continu » ?
    - ✦ Lien avec EPI-journaux « logiciel environné » / article exécutable
- **Défis**: modèle économique, licences, modération …

Mais aussi (et surtout?)

- **Reproductibilité au cœur de la méthode de travail**
  - Développer du logiciel et des données distribuables dès la conception
  - Formations au développement, intégration au rapport annuel
- **Valorisation de la reproductibilité**
  - Dans recrutements et promotions?
- **Profiter du développement d'épi-journaux**
  - Voir http://episciences.org/ & exposé de Laurent Romary
  - Y valoriser la reproductibilité "à la IPOL"
- ***Etre raisonnablement ambitieux***
  - On ne règlera pas tout du jour au lendemain

Conclusion

- **Mouvement de fond vers la reproductibilité**
  - ✓ Nécessite accès ouvert à article **+ code + données**
- **Légitimité et responsabilité d'Inria**
- **Piste immédiate (fragmentation des ressources):**
  - ✓ S'appuyer sur HAL en le faisant évoluer
- **Réflexion plus large lancée**
  - ✦ Articles (P. Guitton, T. Viéville, M.H. Comte)
  - ✦ Groupe de travail du comité des projets d'Inria Nancy (initié en mai 2014)
  - ✦ Workshops spécialisés (G. Fursin, L. Nussbaum)
  - ✦ EPI-journaux (C. Kirchner, L. Romary)
- **Vos idées, contributions, remarques, critiques ?**
  - ✓ **Organiser une journée fin 2014**

Merci de votre attention !

remi.gribonval@inria.fr
http://team.inria.fr/panama