Vers une recherche reproductible

Faire évoluer ses pratiques

Loïc Desquilbet, Sabrina Granger, Boris Hejblum,
Arnaud Legrand, Pascal Pernot, Nicolas Rougier Facilitatrice :
Elisa de Castro Guerra

2019-05-06

Table des matières

Préambule

Ce livre s'adresse à tout acteur de la recherche scientifique qui :

— se pose des questions sur la recherche reproductible ou
— souhaite améliorer ses pratiques.

L'objectif de cet ouvrage est de donner les outils pour comprendre les enjeux de la recherche reproductible et permettre à un chercheur de vérifier ou d'obtenir de nouveau les résultats d'un autre chercheur. Cet autre chercheur peut être lui-même car ce qui était clair à un moment donné ne l'est plus quelques semaines plus tard. Notre ambition est d'apporter des solutions avec un niveau de technicité varié, permettant à chacun d'améliorer sa pratique dans son domaine et selon ses moyens.

Le livre se structure en trois grandes parties :

— la première partie traite de la notion même de recherche reproductible selon différentes perspectives. Il s'agit de prendre pour point de départ des expériences "traumatiques" engendrées par des pratiques plus ou moins (moins en fait) propices à une recherche reproductible.
— la deuxième partie explicite les causes et les origines des différents problèmes auxquels chaque chercheur, quel que soit son degré d'expertise dans son domaine, peut être confronté un jour ou l'autre.
— la troisième partie présente enfin une sélection de solutions offertes par la recherche reproductible pour résoudre ou prévenir ces problèmes à l'échelle de chacun.

Première partie

Chroniques de la non reproductibilité

1

État des lieux

1.1 La crise de la reproductibilité de la science

La crise de la reproductibilité de la science est aujourd'hui un phénomène mondial et largement transdiciplinaire qui concourt à la défiance de la société à l'égard du monde de la recherche (**?**). Le sujet est ancien, mais la situation semble avoir atteint un point critique. Des études ont par exemple démontré qu'il n'était pas possible d'obtenir de nouveau les résultats d'études pré-cliniques ou cliniques (**?**) (**?**). Si la reproductibilité des résultats ne peut être considérée comme seul critère de la scientificité d'une recherche, cette crise suscite des interrogations au sein même de la communauté scientifique.

1.2 Définition(s) de "recherche reproductible" ?

Si la communauté scientifique perçoit ce que peut être une recherche reproductible dans son propre domaine, il s'avère *a priori* difficile de fournir une définition standard satisfaisante pour toutes les disciplines. De fait, parce que la notion de "même résultat" dépend fortement du domaine de recherche. Pour les uns, il suffira de confirmer la signifiance d'un effet, pour les autres, il s'agira d'obtenir le même résultat numérique bit pour bit. L'expression "recherche reproductible" apparaît dès 1992, lors du congrès de la "*Society of Exploration Geophysics*" : "*The first appearance of the phrase "reproducible research" in a scholarly publication appears to be an invited paper presented at the 1992 meeting of the Society of Exploration Geophysics (SEG), from*

the group of Jon Claerbout at Stanford (Claerbout & Karrenbach, 1992). [...] His idea of reproducible research was to leave finished work (an article or a thesis) in a state that allowed colleagues to reproduce the calculation, analysis and final figures by executing a single command. The goal was to merge a publication with its underlying computational analysis" (**?**). De nombreuses définitions de "recherche reproductible" sont fournies par Barba (**?**). Parmi celles-ci, nous retiendrons la suivante, issue de l'article de Vandewalle *et al.*(**?**) : ***"A research work is called reproducible if all information relevant to the work, including, but not limited to, text, data and code, is made available, such that an independant researcher can reproduce the results"*** (**?**). (Notre traduction : "Un travail de recherche est dit reproductible si toutes les informations qui concernent ce travail incluant, sans s'y limiter, le texte, les données, et le code de programmation, sont rendues disponibles de telle sorte que n'importe quel chercheur indépendant peut reproduire les résultats.")

1.3 Pourquoi la question de la reproductibilité est-elle devenue centrale dans les débats actuels ?

Le fait que les problèmes de reproductibilité occupent une telle place dans les débats actuels n'est pas tout à fait un hasard. Le numérique, sous des formes multiples, a largement investi tous les champs de la science et l'outil informatique occupe une place incontournable : stockage, formatage, archivage, indexation, analyse, modélisation, statistiques, environnements, précision, *etc*. Or, peu de chercheurs ont été formés (ou se forment) aux fondamentaux et aux bonnes pratiques liés aux outils informatiques. Cela peut amener à la publication de résultats fragiles (dans le sens "peu robustes") dans le meilleur des cas, et faux dans le pire des cas - mais ce n'est pourtant pas là que se situent les plus graves dangers pour la recherche.

1.4 Où l'on parle de recherche reproductible de manière pragmatique

Quel peut être le point commun entre : un archéologue en train d'effectuer une campagne de fouille, un biologiste préparant une nouvelle expérience dans son laboratoire, un numéricien finalisant une simulation de grande ampleur ? Tous sont exposés aux risques ~~drames~~ suivants, indépendamment de leur volonté de contribuer à l'accroissement des connaissances dans leurs domaines respectifs :

— envoyer à des collègues des données qui ne pourront pas être lues pour des raisons d'incompatibilité de formats,
— réaliser une simulation effectuée sur deux machines différentes et obtenir des résultats radicalement différents,
— se rendre compte qu'une donnée essentielle était stockée sur feu le disque dur (*requiescat in pace*),
— renoncer à une hypothèse prometteuse faute de pouvoir reproduire une de ses propres expériences

La liste pourrait s'allonger. Ne vous êtes-vous jamais posé les questions suivantes : "Suis-je vraiment sûr de mon analyse statistique ?", "Suis-je capable de recréer une figure conçue il y a 6 mois ?" Outre votre équipe de recherche, votre communauté scientifique et *in fine* le monde non académique, le premier bénéficiaire d'une recherche reproductible, c'est d'abord *vous*. Une recherche reproductible facilite en effet les tâches les plus quotidiennes, permet de garantir l'exactitude des méthodes et de documenter l'ensemble de la pratique scientifique. Cela peut-il constituer un gage de qualité de la recherche ? Non. Cela y participe, mais ne suffit pas. En effet, une recherche reproductible (au sens de l'ouvrage, "une recherche dont les résultats publiés peuvent être reproduits") n'est pas synonyme de "bonne" recherche : une mauvaise recherche peut tout à fait être reproductible (*spoiler alert* : *don't try at work !*).

1.5 Aperçu (très rapide) des causes d'une recherche non reproductible

Comme nous le verrons dans ce livre, les causes d'une recherche non reproductible sont très nombreuses. Le suspect habituel est la perte de données. D'autres causes s'avèrent plus difficiles à détecter : par exemple, le chaos numérique, aussi subtil à identifier que vecteur de troubles majeurs. Une fois de plus, il ne s'agit pas de développer une vision accusatoire des pratiques de recherche. L'impossibilité même de reproduire des résultats n'est pas engendrée par la malhonnêteté scientifique, mais s'avère bien plus souvent le fruit d'une forme de méconnaissance, de pratiques plus ou moins hasardeuses. Sous des dehors souvent anodins, les petits "braconnages" du quotidien ("Ça va passer") et autres rustines font le lit de la "dette technique" qui à terme, peut devenir insurmontable et peut condamner un laboratoire. Au travers de situations fictives mais hélas réalistes, nous verrons qu'à l'origine des problèmes de reproductibilité se trouve un ensemble de concepts fondamentaux qu'il est nécessaire de connaître. L'objectif n'est pas de les maîtriser totalement. La recherche reproductible n'exige pas d'adopter une logique du "tout ou rien". Il existe des solutions très simples à mettre en œuvre que tout un chacun peut s'approprier. D'autres solutions demanderont un peu plus de temps et d'énergie.

1.6 La minute théologie négative : ce que vous ne trouverez pas dans cet ouvrage

Il n'est pas question dans cet ouvrage de traiter toutes les solutions pour garantir la reproductibilité de la recherche au sens de Randall et Welser (**?**) : ainsi, la question de la qualité de la recherche est hors périmètre du présent ouvrage. En effet, nous vous proposons plutôt de nous focaliser sur les solutions qui permettent communiquer des

résultats pouvant être reproduits de façon exhaustive. Ainsi, nous n'allons pas traiter des solutions à des problèmes qui nuisent à la qualité de la recherche, et en particulier :

— aller à la "pêche" aux résultats significatifs parmi tous les tests statistiques réalisés ("p-hacking") (**?**), (**?**), (**?**)
— générer une hypothèse de recherche *a posteriori*, c'est-à-dire après avoir obtenu un résultat significatif (« harking ») (**?**)
— sur-interpréter le résultat statistique qui est significatif (« Probability That a Positive Report is False ») (**?**), (**?**), (**?**), (**?**), (**?**)

Pour tous ces sujets cités précédemment, nous invitons le lecteur à se référer à la littérature :

The Seven Deadly Sins of Psychology : A Manifesto for Reforming the Culture of Scientific Practice (**?**)

"Why Most Published Research Findings Are False" (**?**)

"A manifesto for reproducible science" (**?**)

Statistics Done Wrong (**?**)

"A Guide to Robust Statistical Methods in Neuroscience" (**?**)

2

Retours d'expériences

Nous avons recueilli ci-dessous plusieurs témoignages fictifs de *personas* (**?**) incarnant différents acteurs de la recherche. Il s'agit de personnages inventés mais néanmoins vraisemblables car inspirés d'expériences réelles. Parce qu'ils se présentent sous une forme narrativisée, les *personas* permettent :

— à des chercheurs, d'appréhender concrètement différentes questions de recherche reproductible, de contextualiser ces enjeux dans un cadre quotidien ;
— à des personnels de soutien à la recherche, de mieux comprendre les questions auxquelles sont confrontés les chercheurs.

2.1 *Charles P., doctorant en sociologie*

"J'interroge des personnes en situation de précarité économique afin de recueillir leurs témoignages. Je consigne toutes mes notes dans un carnet relié et je retranscris les informations démographiques dans un tableur Calc afin de compiler quelques statistiques : pyramide des âges, répartition des sexes, *etc*. Dans le cadre de ce projet, je collabore étroitement avec un autre doctorant qui a le même directeur de thèse que moi. Il alimente lui aussi ce fichier Calc, que nous nous échangeons régulièrement par clé USB."

2.2 *Jeanne A., jeune méthodologiste en biostatistique*

"J'ai terminé mon Master 2 l'année dernière. Dans ce cadre, j'ai été formée à conduire des analyses statistiques sous R. J'analyse régulièrement des jeux de données fournis par des cliniciens qui travaillent au CHU. Les données sont au format Excel et me sont fournies directement par mail. Je travaille de manière relativement isolée pour effectuer ma tâche d'analyse de données : mes collègues sont en effet tous médecins ou biologistes.

J'ai récemment participé à la rédaction d'un article scientifique et nous venons de recevoir les commentaires des relecteurs : je dois modifier les couleurs d'une figure afin que celle-ci soit lisible en noir et blanc. Comme je n'arrivais pas à remettre la main sur mon script R ayant généré la figure en question, j'ai ré-écrit le programme correspondant. Le seul problème, c'est que cette nouvelle figure est un peu différente de la précédente et remet en cause les conclusions de l'article. Je ne comprends pas ce qui a pu se passer. J'aimerais changer mes pratiques pour rendre ma recherche plus reproductible, mais je ne sais pas par où commencer."

2.3 *Cindy D., stagiaire de Master en physique des matériaux*

"Il y a 3 semaines, j'ai commencé mon stage de Master 2. En fait, j'ai plutôt l'impression d'avoir pris l'autoroute de la souffrance. Ma principale occupation a été d'extraire des données à partir d'une série d'articles, qui donnait les points de caractéristique dans les PDF en fichiers supplémentaires : du *fun* à tous les étages. Je les copie-colle directement dans Excel. *Fun*[2]. Dans le tableur, je dois ensuite transformer les"." en ",". Comme je ne dispose pas des incertitudes d'estimation pour tous les articles

(seuls certains articles les incluent dans leurs résultats), je n'en tiens pas compte et ne rentre que les estimations.

C'est un travail assez rébarbatif (faites un stage qu'ils disaient), mais c'est le seul moyen de pouvoir ensuite analyser les données de la littérature. Lorsque ma collecte sera terminée, ~~je fête ça avec des amis~~ je calculerai le coefficient R^2 pour mes données et mon directeur espère pouvoir publier nos résultats."

2.4 *Long N., enseignant-chercheur en biologie*

"Pour interpréter mes données expérimentales sur les protéines fluorescentes, je me suis lancé dans la simulation moléculaire et j'ai installé un code réputé dans la littérature. La fluorescence est en compétition avec des mouvements intramoléculaires assez rares. Pour avoir une chance de les observer et de faire une statistique, je dois faire des simulations avec des pas de temps assez longs.

En analysant les résultats, j'ai détecté un mouvement intéressant, mais le pas de temps est trop long pour en observer les détails. J'ai donc repris les données stockées par le programme juste avant cet événement et j'ai relancé la simulation avec un pas de temps plus fin. Je n'ai jamais pu reproduire cet événement. En faisant de la bibliographie, j'ai réalisé que ce type de simulation était affecté par du "chaos numérique". Je n'ai plus confiance dans mes résultats."

2.5 *Mohammed B., ingénieur d'étude en calcul scientifique*

"En tant qu'ingénieur du labo, je suis reponsable de la maintenance du logiciel *pytR*, développé il y a 15 ans par un post-doctorant très doué. Il a depuis quitté le labo et personne

ne sait vraiment comment modifier le code de ce logiciel alors qu'une bonne partie de notre activité de recherche repose sur son utilisation. J'ai collé pas mal de rustines qui nous ont permis de tenir un certain temps face à l'évolution de nos infrastructures et de nos systèmes. Mais depuis la mise à jour de notre parc informatique il y a 6 mois, ce n'est plus possible. J'ai été obligé de garder une machine sous l'ancien OS pour pouvoir continuer à faire tourner le logiciel. Mes collègues n'ont pas l'air trop inquiets, mais ça m'angoisse parce que nous avons cumulé une grosse dette technique. Je ne sais pas ce qui va se passer quand cette vieille bécane va nous lâcher !"

2.6 *Christina Z., directrice de recherche*

"Je dirige une équipe de recherche depuis 3 ans et j'encadre actuellement 2 doctorants et 3 post-doctorants. Je suis un peu inquiète car l'un de mes doctorants a soutenu sa thèse la semaine dernière et part en post-doc aux États-Unis à la fin du mois. Ses derniers écrits sont très prometteurs, mais il reste 50% du travail à faire et l'article n'est pas encore rédigé. J'ai prévu de recruter un stagiaire pour prendre la suite mais cela risque de prendre énormément de temps : alors que l'exploitation des résultats est particulièrement délicate sur ce projet, ce doctorant documente très peu les étapes de son travail en dehors de ses manips. Pour aggraver la situation, tout doit être bouclé d'ici la fin de l'année car le financement du projet arrive à échéance.

Récemment, j'ai aussi aussi reçu des nouvelles d'un ancien camarade de thèse dont un article important s'est fait rétracter. Certaines conclusions de son papier ont été attaquées. Il n'a pas pu fournir les données expérimentales qu'il avait utilisées : l'article date d'il y a 5 ans et il a perdu toute trace du doctorant qui avait conduit les manips. L'éditeur de la revue a rétracté l'article, faute d'éléments tangibles pour faire valoir un éventuel

droit de réponse. La réputation de mon ancien camarade en a pris un coup.”

3

Symptômes d'une recherche non reproductible

Il existe de nombreuses situations où l'on souhaite reproduire des résultats de recherche :

— on peut vouloir vérifier que la méthode mise en œuvre il y a quelques mois par un stagiaire, par un doctorant, ou par soi-même, donne les mêmes résultats avant de poursuivre l'étude ;
— on doit obéir (hé oui, c'est bien la dure vie de chercheur) aux demandes d'un relecteur de modifier une figure ou de tester l'impact de paramètres non envisagés dans l'étude initiale ;
— on souhaite vérifier que les méthodes "maison" font aussi bien, voire mieux, que celles des équipes concurrentes.

Nous sommes persuadés que vous avez d'autres situations en tête. Le point commun de toutes ces situations est qu'elles peuvent être l'occasion de surprises (très) désagréables que nous appellerons les "symptômes d'une recherche non reproductible". En complément des *personas*, nous vous livrons ci-dessous quelques témoignages, certes fictifs mais non moins vraisemblables, qui illustrent les dits symptômes.

3.1 J'ai perdu mes données ou mon code de programmation !

Après la publication d'un des mes articles, un collègue souhaite collaborer avec moi pour tester de nouvelles hypothèses sur le jeu de données que j'ai présenté. Malheureusement :

— le disque dur sur lequel j'archivais les données brutes a crashé, ou bien,

— j'ai effacé (vous repassez par la case départ) les données qui prenaient beaucoup de place sur mon ordinateur, puisque l'article était publié.

Que pourrais-je répondre à ce collègue tout en gardant un semblant de dignité ? Non reproductibilité : 1 / Tranquillité d'esprit : 0

3.2 Mes résultats ont changé !

Il y a quelques mois, j'ai soumis un article au journal. Depuis, j'ai continué de travailler sur mon code de traitement des données. Un *reviewer* me demande de compléter quelques figures, ce qui nécessite pour moi de faire de nouveau l'analyse des données. Malheureusement, la version actuelle du code, dans laquelle j'ai amélioré les algorithmes, ne donne plus tout à fait les mêmes résultats que ceux de l'article.

Dois-je annoncer au *reviewer* qu'une partie des résultats a changé ?

3.3 Mon code ne marche plus !

Ayant réussi à décrocher une ANR, je m'offre un nouvel ordinateur, avec la version la plus récente du système et des logiciels. Pour profiter au mieux des performances de cette machine, je recompile avec enthousiasme mon code de simulation : ma simulation prendra dix fois moins de temps qu'avec ma précédente machine, me dis-je. Mon excitation retombe subitement en voyant que :

— le compilateur génère des erreurs, ou bien
— le code recompilé démarre et se plante après quelques secondes.

Ma première idée (un peu la mort dans l'âme) est de récupérer mon

vieil ordinateur, mais l'informaticienne du labo l'a déjà reformaté pour le passer à un stagiaire. Et là je me dis : "Mais pourquoi moi ?"

3.4 Mon nouveau doctorant n'observe pas les mêmes effets que son prédécesseur !

L'an passé, un de mes doctorants a soutenu brillament sa thèse après avoir obtenu des résultats remarquables, que nous avons publiés dans un excellent journal. Il a trouvé un post-doctorat à l'étranger sur un sujet sensiblement différent pour élargir le spectre de ses compétences. Mon nouveau doctorant doit repartir de ces résultats pour améliorer l'efficacité de notre processus. Cela fait maintenant un an qu'il échoue à reproduire les observations de son prédécesseur, alors qu'il suit scrupuleusement (et j'ai vérifié) le protocole établi. Dois-je mettre fin à sa thèse pour incompétence, le lancer sur un autre sujet et abandonner cet axe de recherche, ou envisager de rétracter l'article de son prédécesseur car tout d'un coup pris d'un certain doute ?

3.5 Mais fallait-il vraiment écrire toute cette tartine ?

Pour mon stage de M2, je souhaite montrer un lien entre la délétion d'un gène chez la souris et la survenue de problèmes neurolocomoteurs. J'ai vu de nombreux articles évaluer la présence de ces problèmes chez la souris, mais aucun ne m'a particulièrement attiré. En revanche, toute cette revue de la littérature m'a donné une super idée pour une telle évaluation comportementale, reposant sur un protocole d'observations certes complexe mais que je jugeais génial. Je n'ai pas voulu perdre de temps (nous ne pouvions bénéficier de ce lot de souris que pendant un mois). J'ai obtenu d'excellents résultats et les ai présentés en réunion d'équipe, tout excité à l'idée de rédiger mon premier article que j'allais soumettre dans une excellente

revue. Après ma présentation, mon chef d'équipe m'a demandé de lui montrer le descriptif rédigé de mon protocole si "génial". Je lui ai répondu que je n'avais pas voulu prendre le temps de rédiger quelque chose qui était dans ma tête : quel intérêt d'écrire toute cette tartine pour soi, quand tout est si clair dans sa tête ? Il m'a répondu "Ok, donc tu peux oublier ton premier papier sur tes souris".

Deuxième partie

Sources de non reproductibilité

4

Acquisition de l'information

Dans une démarche de recherche, la première étape est bien souvent l'aquisition d'information, que ce soit à partir de la collecte de nouvelles mesures expérimentales ou à partir de données déjà publiées.

4.1 Absence de standardisation de la collecte des informations

Dans la grande majorité des cas, la production de résultats issus d'une recherche passe par la collecte d'informations. Ces informations sont recueillies sur des "unités" (une unité pouvant être une pièce mécanique, un être vivant, *etc*.). Ce que l'on entend par "informations" sont les caractéristiques de chaque unité qui fait l'objet de la recherche ; par exemple : la résistance à la traction d'un matériau, la concentration en glucose d'une personne atteinte de diabète, ou bien encore son âge, son poids, *etc*.

Si cette collecte des informations n'est pas standardisée, la personne qui collecte des informations sur un individu/unité un jour n°1 ne le fera potentiellement pas de la même façon le jour n°2 si elle doit réitérer l'opération. Or, si elle ne procède pas à l'identique, la valeur de l'information du jour n°2 sera différente de celle du jour n°1, non pas parce que l'information a changé au cours du temps (ce qui est possible, et éventuellement acceptable – *cf*. ci-dessous), mais parce que la méthode a changé.

Ainsi, comment s'assurer d'une recherche reproductible si celle-ci se fonde sur des informations dont la valeur varie en fonction des

modalités de collecte appliquées ? Quelle peut-être la valeur, voire la fiabilité, des données issues d'un mode de collecte qui n'est pas stabilisé pendant toute la durée de l'étude ?

4.2 Modification des données après une première collecte

Un autre problème conduisant à une recherche non reproductible se produit lorsque des informations recueillies sur une unité sont modifiées après une première collecte, sans que ces modifications ne soient tracées. Dans ce cas-là, les analyses statistiques qui seront conduites sur les informations modifiées ne fourniront évidemment pas les mêmes résultats que celles conduites sur les informations initiales. S'il n'y a aucun moyen de revenir aux informations initiales et/ou de savoir quelles sont les informations qui ont été modifiées, votre recherche devient par conséquent non reproductible.

Les solutions pour éviter de perdre ainsi la trace de la modification sont évoquées dans le chapitre **??**.

4.3 Une collecte d'informations non répétable ou non reproductible

Dans cette section, en vue d'examiner les impacts de l'étape de la collecte de données, nous allons utiliser une définition particulièrement précise des termes "reproductibilité" et "répétabilité", en utilisant leur définition métrologique.

4.3.1 Quelques définitions issues du *Vocabulaire International de Métrologie*

Nous avons décidé de choisir les définitions proposées en 2012 dans la dernière version du *Vocabulaire International de Métrologie (VIM)* (**?**)

car elles représentent l'effort le plus récent de normalisation dans ce domaine ; document téléchargeable ici : (**?**).

La **fidélité** de mesure s'entend comme : "l'étroitesse de l'accord entre les indications ou les valeurs mesurées obtenues par des mesurages répétés du même objet ou d'objets similaires dans des conditions spécifiées."

La **répétabilité** est la fidélité de mesure dans les conditions de mesures suivantes : "conditions qui comprennent la même procédure de mesure, les mêmes opérateurs, le même système de mesure, les mêmes conditions de fonctionnement et le même lieu, ainsi que des mesurages répétés sur le même objet ou des objets similaires pendant une courte période de temps."

La **reproductibilité** est la fidélité de mesure dans les conditions de mesures suivantes : "conditions qui comprennent des lieux, des opérateurs et des systèmes de mesure différents, ainsi que des mesurages répétés sur le même objet ou des objets similaires."

4.3.2 Impact d'une absence de répétabilité ou de reproductibilité dans la collecte des informations

Si la collecte d'une information n'est pas "répétable" au sens du *VIM* défini ci-dessus (**?**), les conditions d'une recherche reproductible ne peuvent alors pas être remplies : vous n'obtiendriez pas les mêmes résultats à partir d'informations collectées sur des unités identiques, évalués dans les mêmes conditions par un même opérateur. Si la collecte des données n'est pas "reproductible" d'un opérateur à un autre au sens du *VIM*, les conditions d'une recherche reproductible ne sont pas non plus remplies : personne d'autre que vous ne pourrait obtenir les mêmes résultats sur des unités identiques évaluées dans les mêmes conditions.

4.4 Collecte des données à plusieurs

Supposons que vous ne soyez pas la seule ou le seul à collecter les informations pour votre étude. Deux questions se posent alors :

— la première, déjà abordée ci-dessus, concerne la standardisation de la collecte des informations : si cette collecte n'est pas standardisée, votre collègue et vous n'obtiendrez potentiellement pas les mêmes valeurs des informations collectées lorsque vous évaluez pourtant les mêmes unités.
— la seconde concerne l'outils de partage de l'information : dans quel document, sur quel support, allez-vous collecter les données, pour garantir que vous et votre collègue n'allez pas effacer ou affecter les informations collectées par l'autre ?

4.5 Collecte des données de la littérature

Nous envisageons maintenant le cas d'une étude qui dépend d'informations collectées dans la littérature. Dans ce genre de cas, une intervention manuelle est souvent nécessaire pour constituer la base de données.

Considérons d'abord le cas, fréquent dans certains domaines (et *a priori* favorable), où les données d'intérêt sont dans le fichier PDF d'un article ou de son supplément. Lorsqu'on effectue un copier/coller d'une partie de fichier PDF vers un éditeur de texte, les sources de contrariété sont multiples et dépendent largement du logiciel utilisé pour afficher le fichier PDF. Les désagréments les plus courants sont :

— une impossibilité éventuelle de gérer correctement des tables complexes avec des cellules vides ou une table pivotée ;

— la présence de renvois bibliographiques sur certains éléments du tableau ;
— la gestion du signe moins (" - "), qui est souvent récupéré comme un tiret (" — ") ou demi-tiret, ne pouvant alors pas être interprété par les codes de calcul.

Après extraction des données, une étape de correction manuelle est donc souvent indispensable et constitue en elle-même une source potentielle d'erreur, en plus de ne pas toujours être effectuée de façon traçable. La récupération de données à partir d'images (OCR) présente des problèmes similaires.

Et pour le chercheur aventureux, copier/coller les données collectées dans un tableur peut introduire une couche supplémentaire de surprises (transformation de nombres ou d'identifiants en dates, par exemple) (**?**).

4.6 Que faire?

Les solutions pour faire face aux problèmes évoqués dans ce chapitre, dépendent du collecteur de données, mais également de l'émetteur.

Le collecteur de données pourra se reporter aux solutions présentées dans les chapitres **??** et **??** pour automatiser et tracer au maximum le processus de collecte, d'autant plus que le volume de données est important et/ou si la tâche est répétitive.

L'émetteur de données pourra se reporter aux solutions présentées dans les chapitres **??** et **??** dédiés aux bonnes pratiques d'archivage de données dans des formats ouverts et lisibles par la machine.

NB : en tant que chercheurs, nous sommes souvent émetteurs de données pour d'autres chercheurs. De fait, nous devrions intégrer autant que possible cet aspect dans nos bonnes pratiques de partage de résultats.

5

Gestion des données

La perte de données à tous les niveaux d'un processus de recherche est une cause majeure de non reproductibilité. Cela peut aller du simple accident matériel, comme par exemple le *crash* d'un disque, au problème de méthode, comme l'absence d'une politique de sauvegarde ou de règles élementaires de documentation (métadonnées).

5.1 Intégrité et curation des données

Voici un *scenario* catastrophe classique quand il est question d'intégrité des données : alors qu'un éditeur vous demande de mettre à disposition les données brutes sous peine de ne pas publier votre article pourtant accepté, les données associées ont été effacées ou égarées. Quel que soit le degré d'intégrité scientifique du chercheur, si des doutes sur la validité des données émergent, ne pas être en mesure de fournir les données constitue pour lui un handicap difficilement surmontable. Par ailleurs, l'absence de sauvegarde des données est considérée comme une négligence professionnelle.

Il existe des *scenari* encore plus insidieux où l'intégrité des données peut être compromise sans que vous vous en rendiez compte. Par exemple : vous recevez vos données avec une certaine précision mais vous sauvegardez ces données avec une précision moindre. Vous serez alors confronté à une perte d'information irréversible : une partie de l'information s'est littéralement évaporée. De même, dans le cas de résultats produisant un déluge de données (comme par exemple le *Large Hydron Collider*) et devant l'impossibilité de tout sauvegarder, il faut sélectionner les données à sauvegarder, sachant que les autres

seront irrémédiablement perdues. Une mauvaise décision initiale peut se révéler catastrophique pour peu que vous ayez besoin de ces données à une étape ultérieure.

Enfin, si vous ne vous êtes pas assuré du contrôle d'accès sur vos données, quelqu'un peut venir les modifier par inadvertance et à votre insu, changeant ainsi les conclusions de vos analyses.

5.2 Traçabilité de la source des données

Quand bien même l'intégrité des données aurait été assurée, l'absence d'information descriptive sur la source des données (métadonnées) peut causer de nombreux problèmes. Vos données sont disponibles mais impossible de comprendre ce qu'elles représentent exactement. Par exemple : des données sont collectées dans la littérature, mais les références bibliographiques ne sont pas mentionnées ou s'avèrent lacunaires. Un problème pouvant être perçu comme formel constitue en réalité un manque de traçabilité portant atteinte à la reproductibilité. Le bibliothécaire avait donc raison.

5.3 Indexation des données

Lorsque vous manipulez de très larges volumes de données (en termes de nombre d'échantillons) il devient tout à fait possible de perdre, non pas les données, mais l'accès à ces données. Imaginez : vous avez utilisé un nommage particulier des fichiers pour indiquer par exemple la nature de la donnée (*well done*!) mais vous avez égaré le fichier expliquant les règles de nommages (*too bad*). Alors que vous possédez l'intégralité de vos données, vous vous trouvez incapable les utiliser.

5.4 Codages et unités

Lorsque vous sauvegardez des données sur un support informatique, il est important de comprendre qu'un certain nombre de choix sont effectués de façon automatique et sans possibilité de contrôle de votre part. Ces choix dépendent étroitement de l'architecture matérielle de votre ordinateur. Par exemple, en ce qui concerne la représentation des nombres en virgule flottante, certaines machines vont lire la représentation binaire de gauche à droite alors que d'autres le feront de droite à gauche (*endianess*). Si vous travaillez toujours avec le même type de machine, vous n'aurez pas de problème jusqu'au jour où vous changerez de machine et observerez alors des valeurs complètement erratiques, vous laissant à penser que vos données auront été compromises.

Plus généralement, stocker ou transmettre des données numériques sans en préciser les unités ni les conventions de codages associées constitue un vecteur important de risques, notamment si un tiers désire les réutiliser. Cela fut le cas pour la sonde "*Mars Climate Orbiter*" qui s'est désintégrée à la surface de Mars en raison d'une communication entre un système de mesure anglo-saxon (émission) et un système métrique (réception) (**?**).

5.5 Obsolescence des données

Dans certains cas, les données ont été sauvegardées, leur intégrité est parfaite, on peut les retrouver très facilement et pourtant, elles s'avèrent inutilisables. Comment expliquer ce paradoxe ? Les données sont généralement sauvegardées dans un format pouvant être ouvert ou fermé (propriétaire). Or si le format est fermé, vous ne pouvez pas contrôler l'évolution de ce format. Prenez par exemple un fichier Word créé il y a une vingtaine d'années, pouvez-vous encore

le lire aujourd'hui ? Votre version de Word vous assure-t-elle une compatibilité avec ce format obsolète ? Vous avez répondu par la négative à ces questions ? Considérez alors les données comme inutilisables.

5.6 Que faire ?

Privilégier des formats ouverts (chapitre **??**), assurer un archivage pérenne des données et leur associer des métadonnées pertinentes sur des serveurs institutionnels ou publics (chapitres **??** et **??**) constitue actuellement l'une des meilleures manières de se prémunir contre la perte de données.

6

Programmation et calcul

Les problèmes inhérents au calcul et aux codes associés partagent des similarités avec les difficultés liées aux données, par exemple la non disponibilité. Toutefois, les questions de calcul ont leurs spécificités du fait de leur nature opératoire : il s'agit d'exécuter ce code afin d'obtenir un résultat. Or, c'est lors de cette étape d'exécution que vont surgir un certain nombre de complications que l'on peut classer en deux grandes catégories :

— d'une part, celles qui empêchent d'obtenir un résultat
— d'autre part, celles qui rendent un résultat différent voire faux.

Si le premier type de problème est ennuyeux (euphémisme), le second type de problème est d'autant plus grave qu'il est difficile à détecter (effroi intense).

6.1 Le code n'est pas disponible

En guise de préambule, débutons par une liste non exhaustive des cas où l'on n'a tout simplement pas ou plus accès au programme à exécuter :

— **Les logiciels propriétaires ou la loterie de la licence d'exploitation** : votre équipe/structure a cessé de payer la licence. Variante : ce logiciel est disponible dans l'université d'un collègue mais pas dans l'établissement où vous travaillez actuellement. Autre variante : vous avez accès au logiciel, mais seul un nombre restreint de personnes peut y accéder

en même temps, *via* un système de jetons. De fait, vous vous retrouvez à devoir attendre un bon moment avant d'y arriver.

— **Un seul code vous manque et tout est dépeuplé** : le code a été développé "en interne". Il arrive trop souvent qu'à la suite d'un *crash* disque, d'un vol d'ordinateur portable, du départ du développeur principal, que l'on n'ait simplement plus accès au logiciel. C'est souvent le résultat d'une politique (ou d'une absence de politique) de sauvegarde ou de partage d'informations au sein d'une équipe.

— **Le numéro que vous avez demandé n'est plus attribué** : assez souvent, il s'agit d'un code développé "en externe" (dans une autre équipe de recherche par exemple) que l'on souhaite ré-exécuter, par exemple pour avoir un point de comparaison ou bien pour vérifier si on obtient bien des résultats similaires avec une autre méthode. En général, on cherche alors le code sur le web mais il est assez courant que l'URL indiquée dans l'article ne soit plus accessible car le développeur a depuis quitté l'équipe où il travaillait et que sa page web a été supprimée ou complètement restructurée. Ce problème est connu sous le nom d'*URL decay* (**?**) ou de *Link Rot* (**?**).

— **Cachez ce code que je ne saurais voir** : enfin, les auteurs du code peuvent tout simplement ne pas souhaiter le partager, par exemple parce qu'ils jugent qu'il n'est pas montrable en l'état (pas ou peu commentaires, structure horrible cachant des erreurs) ou encore pour conserver ce qu'ils considèrent comme un avantage compétitif.

Si cette question vous intéresse, vous pouvez lire les travaux de Collberg, Proebsting et Warren : (**?**) (**?**). Les auteurs étudient les causes d'incapacité à réexécuter du code dans la communauté de recherche *Computer Systems*, pourtant très au fait des aspects logiciels. Vous y trouverez de nombreux témoignages (assez drôles si c'était sans conséquences !) issus d'une étude de terrain ; vous pourrez notamment lire les excuses les plus couramment utilisées pour justifier une incapacité à donner accès au code derrière une publication. Vous pouvez aussi consulter "Re-run, Repeat, Reproduce, Reuse, Replicate : Transforming Code into Scientific Contributions"(**?**).

6.2 Comment lance-t-on ce code ? ("Allô Houston ?")

Lorsque l'on fait de la recherche, il est courant de devoir développer soi-même un code pour répondre à un besoin spécifique. Que ce soit un "gros" code ou un petit script, on prend rarement le temps de rédiger une documentation "externe" (à destination des utilisateurs) puisque le code est principalement utilisé par les membres de l'équipe que l'on croise tous les jours. Mais lorsque l'on revient quelques mois plus tard, pour ré-exécuter un de ses propres calculs ou bien que l'on essaye de repartir du travail de quelqu'un d'autre (qui a quitté le laboratoire ou n'y a même jamais travaillé), il est courant de ne pas (ou plus) savoir comment il avait été lancé. Avec quels paramètres, quels fichiers d'entrées, quelles variables d'environnement, *etc.* ? La moindre erreur sur les paramètres conduira à des résultats différents voire à un *crash*. Et malheureusement pour vous, le "vous" d'il y a 6 mois ne répond pas au *mail*. Enfin, et comme nous le verrons par la suite, il existe bien d'autres raisons qui peuvent conduire à ces deux symptômes.

6.3 Comment fonctionne ce code ? *Lost in translation*

Si on n'est plus sûr des paramètres utilisés, on peut vouloir chercher à comprendre d'où vient le problème en inspectant le code … si tant est qu'on ait accès au code source bien sûr. Or, les logiciels propriétaires ou les logiciels disponibles uniquement sous forme binaire rendent toute inspection de ce type impossible. Mais admettons que vous ayez réussi à inspecter les sources et que vous ayez les compétences pour le comprendre (*a minima*, un langage de programmation que vous connaissez).

Les codes de recherche, développés pour des besoins spécifiques, sont souvent des prototypes et il est rare de prendre le temps de rédiger une documentation "interne" (à destination des développeurs). Et

quand bien même il y aurait des commentaires, encore faut-il qu'ils soient compréhensibles donc *a minima* en anglais. Par ailleurs, ils doivent aussi correspondre à la réalité : quand un code évolue vite, on ne prend pas toujours le temps d'actualiser au fur et à mesure les commentaires et la documentation. Si ces critères ne sont pas réunis, les commentaires risquent davantage de vous induire en erreur que de vous aider.

Un célèbre dicton en informatique dit : "*Programs must be written for people to read, and only incidentally for machines to execute*." C'est une citation d'Harold Abelson [1] tirée de son livre *Structure and Interpretation of Computer Programs* publié en 1979 (**?**). Commenter, c'est une chose, mais lorsque l'on cherche à comprendre un programme, on se rend vite compte qu'il est indispensable que les noms de variables et de fonctions aient été bien choisis, que le code ait été été bien structuré avec des fonctions au rôle clairement défini, sans quoi le code devient totalement incompréhensible (ce qui est précisemment l'objet du concours "Obfuscated C" (**?**). De même, lorsque qu'il s'agit d'un code conséquent réparti dans de nombreux fichiers, une mauvaise convention de nommage des fichiers ou bien l'usage d'une structure de fichiers absconse empêchent toute tentative de compréhension.

Enfin, même si le code est relativement compréhensible, il est possible que des *bugs* (des erreurs de programmation) soient à l'origine de vos malheurs mais comment les trouver ?

6.4 Quelle version du code ?

Nul n'est parfait et les *bugs* sont donc courants, même chez les programmeurs les plus chevronnés. Il se peut que le *bug* à l'origine de vos problèmes provienne de la version d'un logiciel actuellement installé sur la machine. Pour corriger ce *bug*, on peut vouloir mettre à jour le logiciel. Mais quelle version a été utilisée dans le passé et quelle est la version actuelle ? Et comment savoir si c'est effectivement

1. https://en.wikipedia.org/wiki/Hal_Abelson

la cause de la différence observée ? La mise à jour n'introduirait-elle pas de nouveaux *bugs* ? L'idéal serait peut-être de revenir à une version plus ancienne mais comment faire ? Quelle est la version la plus récente que je puisse utiliser ?

Enfin, cette nouvelle version sera-t-elle toujours compatible avec mon ordinateur ? Et si je repars du code source, arriverai-je à le recompiler ?

6.5 L'environnement de calcul ou le paradigme des poupées russes diaboliques

Plus le langage que vous utilisez est de haut niveau, plus il est probable qu'il dissimule une grande complexité. Même le script le plus anodin dépend en général d'une large hiérarchie de bibliothèques que l'on a du mal à imaginer. À titre d'exemple, lorsqu'en Python vous souhaitez faire un petit graphique, il est courant de charger la bibliothèque `matplotlib` avec un simple :

```
import matplotlib
```

Or cette bibliothèque est fournie par un "paquet" qui, sur la machine d'un des auteurs s'appelle, `python3-matplotlib`. Lorsque nous cherchons à en savoir plus sur ce paquet, voilà ce que nous obtenons :

```
Package: python3-matplotlib
Version: 2.1.1-2
Depends: python3-dateutil, python-matplotlib-data (>= 2.1.1-2),
python3-pyparsing (>= 1.5.6), python3-six (>= 1.10), python3-tz,
libjs-jquery, libjs-jquery-ui, python3-numpy (>= 1:1.13.1),
python3-numpy-abi9, python3 (<< 3.7), python3 (>= 3.6~),
python3-cycler (>= 0.10.0), python3:any (>= 3.3.2-2~), libc6 (>=
2.14), libfreetype6 (>= 2.2.1), libgcc1 (>= 1:3.0), libpng16-16 (>=
1.6.2-1), libstdc++6 (>= 5.2), zlib1g (>= 1:1.1.4)
```

C'est ici la version 2.1.1-2 qui est présente et, pour

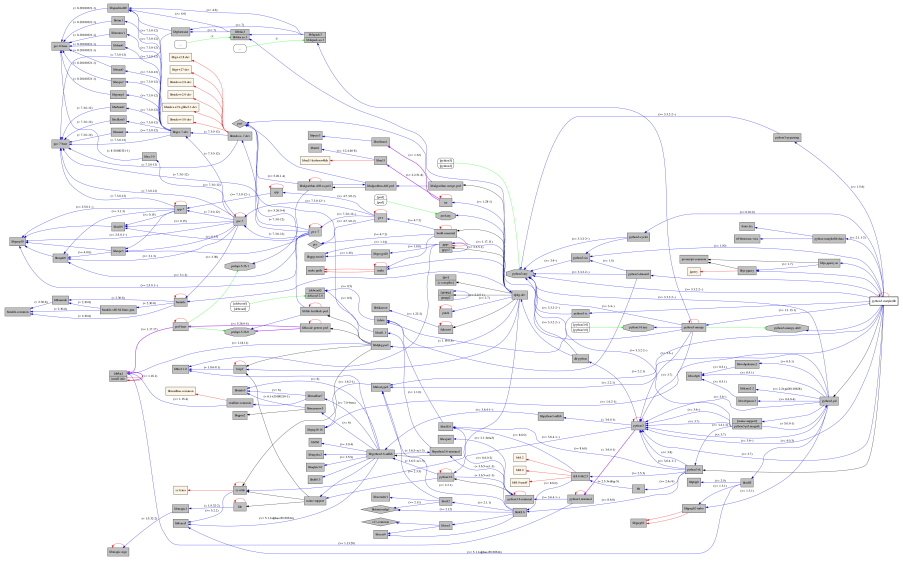

FIGURE 6.1 : Dépendances de Matplotlib sous Debian obtenues avec debtree

l'installer, il a fallu installer les paquets `python3-dateutil`, `python-matplotlib-data`, `python3-pyparsing`, *etc.* C'est ce qu'on appelle les "dépendances". Mais pour ces paquets dépendent eux-mêmes d'autres paquets. Lorsque l'on récupère l'ensemble des paquets nécessaires avec leurs dépendances, voici le graphe qu'on obtient, *cf.* Fig. 6.1 (alerte paracétamol) :

Vous remarquerez dans les dépendances que la version n'est pas précisément indiquée mais qu'il faut par exemple une version supérieure de `python3-pyparsing` qui soit au moins 1.5.6. Mais si des *bugs* peuvent être introduits, comment être sûr que votre code fonctionnera de la même façon avec les versions `1.5.6`, `1.5.7`, ..., sachant que nous en sommes maintenant au moins à la version `2.2.0`.

En résumé, tout code, aussi petit soit-il, possède tout un arbre de dépendances qui sont le plus souvent cachées. Ce code s'exécute donc dans un environnement donné et une différence, même insignifiante, de cet environnement peut conduire à des résultats différents, c'est-à-dire à des problèmes de non reproductibilité.

Si vous pensez que ce problème n'est que théorique, nous vous invitons à lire Gronenschild et ses co-auteurs (**?**) qui étudient l'influence de la version de MacOSX et de FreeSurfer, un logiciel permettant de mesurer l'épaisseur corticale et le volume de structures neuroanatomiques.

6.6 Le chaos numérique

Les nombres manipulés par ordinateur ne sont pas des nombres réels, avec une précision infinie, mais des nombres dits "à virgule flottante" qui n'obéissent pas exactement aux mêmes règles que celles que l'on nous enseigne à l'école. Par exemple, si vous demandez à, à peu près n'importe quel ordinateur si `0.1*3==0.3` ou si `3-2.9==0.1` il vous répondra très certainement `FALSE` (Faux) dans les deux cas. Cela est dû au fait que la représentation au format binaire de ces nombres, en apparence simple, n'est pas exacte. Beaucoup de machines à calculer ont une représentation interne en base `10` un peu différente, or, nous n'avons pas été habitués de manière précoce à intégrer ce genre de problème, sauf peut-être pour des nombres du genre $1/3 \approx 0,3333333$. Lorsque l'on programme, il faut donc faire très attention à cette subtilité qui joue des tours dès que l'on veut comparer deux nombres.

Un autre problème au prime abord surprenant, mais probablement plus simple à comprendre, est la non associativité des opérations. Si avec les nombres réels, il va de soi que $(a + b) + c = a + (b + c)$, ce n'est pas le cas avec les nombres en virgules flottantes. Par exemple, `(1e-10+1e10)-1e10` vaut `0` alors que `1e-10+(1e10-1e10)` vaut `1e-10`. Même une simple moyenne peut donc devenir problématique et, non, il ne suffit pas de trier les nombres avant de les additionner pour résoudre le problème.

Comme vous utilisez vraisemblablement un ordinateur parallèle (même votre téléphone a maintenant plusieurs cœurs de calcul), il est possible que la somme $a_1 + \cdots + a_n$ ne soit pas calculée comme vous

l'imaginez (i.e., $(((((((a_1 + a_2) + a_3) + \cdots + a_n))$, mais en plusieurs parties (i.e., $((((a_1 + a_2) + \cdots + a_{n/2}) + ((((a_{n+1} + a_{n+2}) + \cdots + a_n))$, chaque cœur de votre processeur réalisant une des sommes partielles, la somme finale étant faite à la fin. Les résultats peuvent changer par le simple fait de passer d'un ordinateur à un autre alors que les machines semblent avoir le même environnement. Les ordinateurs peuvent ne pas avoir exactement le même nombre de cœurs. Les cœurs d'un ordinateur n'allant pas toujours exactement à la même vitesse, un code un peu optimisé ajustera la taille des sommes partielles pour terminer le calcul le plus rapidement possible et le résultat du calcul variera donc d'une exécution sur l'autre alors que rien n'a changé ! Mais alors, comment décider lequel de ces différents résultats de calculs est le "bon" ?

Toutes ces petites imprécisions de calcul peuvent hélas rapidement devenir très problématiques lorsque le système sous-jacent correspond par exemple à la discrétisation d'une équation différentielle. Le calcul est alors très sensible aux conditions initiales et l'accumulation des imprécisions peut amener à une catastrophe (voir notamment *The Patriot Missile Failure* (**?**)).

Il y a de nombreux articles décrivant ce genre de cauchemars : (**?**) (**?**)

Vous pouvez vouloir lire le classique *What Every Computer Scientist Should Know About Floating-point Arithmetic* [2] (**?**) ou encore les travaux de Stodden et ses collègues (**?**). Pour une présentation de ces problématiques et de quelques solutions, vous pouvez aussi vouloir regarder ce séminaire sur la reproductibilité numérique (**?**).

6.7 Que faire ?

Pour résoudre ces problèmes, des solutions sont abordées aux chapitres **??**, **??**, **??** et **??**.

2. https://www.itu.dk/~sestoft/bachelor/IEEE754_article.pdf

7

Communication des résultats

Nous allons voir dans ce chapitre qu'une recherche peut devenir non reproductible s'il existe une mauvaise utilisation des résultats de l'étude au moment de la rédaction d'un article : il peut s'agir d'une mauvaise sélection de votre part des résultats, d'un choix inapproprié du format de présentation ou d'une transformation de ces données.

7.1 Une mauvaise sélection des résultats

Les résultats fournis par un logiciel peuvent contenir de si nombreuses informations qu'il faille opérer une sélection parmi elles. En d'autres termes, vous pouvez être amené à devoir identifier les informations pertinentes pour la question de recherche faisant l'objet de votre article.

Confronté à cet amoncèlement, il peut vous arriver de mal sélectionner l'information pertinente : votre sélection à la souris a "oublié" quelques caractères en début ou en fin de séquence à sélectionner, par exemple. En outre, si cette information est complexe et difficilement compréhensible par vos collaborateurs parce que vous êtes seul spécialiste du domaine, alors cette erreur de sélection sera répercutée dans l'article et persistera après le processus de revue si les relecteurs ne la repèrent pas.

7.2 Transformation des résultats

Une autre erreur pouvant survenir à partir des résultats fournis par un logiciel est celle d'une "transformation" (bien entendu involontaire) de ces données. Cette modification délétère peut très facilement se produire si vous devez saisir de nouveau dans votre article les sorties résultats du logiciel. Une erreur de frappe est si facilement arrivée !

Un deuxième exemple de transformation des résultats est celui de l'amélioration d'une figure. Si vous trouvez que la figure que vous obtenez à partir d'un logiciel n'est pas satisfaisante, vous pouvez décider de la retravailler manuellement, par exemple en ajoutant une courbe interpolant des points, alors que les valeurs intermédiaires n'ont pas de sens. Cette manipulation est dangereuse, car le traitement de l'information n'est plus produit par une chaîne d'instructions validée et automatisée, mais par vous-même – et vous n'êtes pas infaillible.

Une transformation involontaire des résultats peut aussi se produire si vous collaborez sur une étude et si vous devez intégrer les résultats d'analyses réalisées par votre collaborateur, sans en comprendre toutes les subtilités. Un exemple que l'on rencontre malheureusement fréquemment dans la littérature concerne les résultats d'analyses statistiques. Vous avez demandé à votre collègue, spécialiste des statistiques, de traiter certaines de vos données, et il vous envoie ses résultats que vous ne comprenez pas dans le détail. Il peut alors facilement arriver que, par défaut de compréhension, vous retranscriviez mal ou partiellement les résultats dans l'article. De telles erreurs de saisie peuvent passer totalement inaperçues si les relecteurs n'ont pas non plus les compétences statistiques requises pour interpréter ces résultats.

7.3 Présentation des résultats

Une forme très courante de perte d'information est liée à l'absence ou à la forme inappropriée des incertitudes associées aux résultats d'une mesure physique ou virtuelle, ou d'une étude statistique. Très souvent, l'absence d'incertitude (par exemple, l'absence de valeur d'écart-type) empêche une comparaison de résultats, ou bien l'absence de matrice de corrélation entre les paramètres incertains d'une étude empêche leur réutilisation. Consulter : "How Measurements of Rate Coefficients at Low Temperature Increase the Predictivity of Photochemical Models of Titan's Atmosphere" (**?**).

Même lorsque des efforts ont été faits pour publier les informations adéquates, des problèmes d'arrondi peuvent anéantir la réutilisabilité des données. Une mauvaise sélection du nombre de chiffres significatifs à reporter dans un résultat numérique peut tout à fait se produire. Pour une question de "présentation" (taille d'une table de résultats), vous pourriez juger qu'un seul chiffre significatif peut suffire. Mais si vos résultats sont nécessaires à la réalisation d'autres recherches, pour permettre par exemple des simulations, les erreurs engendrées dans ces autres travaux pourront être amplifiées : une petite erreur initiale peut conduire à une erreur très importante en bout de course – cf. « Chaos numérique » dans le chapitre **??**.

A titre d'illustration, la matrice de variance-covariance publiée par le CODATA en 2002 pour l'ajustement des constantes fondamentales, arrondie pour être présentable dans les annexes de l'article, s'est avérée inutilisable pour des travaux ultérieurs (**?**).

7.4 Que faire ?

La réutilisabilité des résultats d'une étude doit être une priorité. Pour cela, les données doivent être mises à disposition de futurs utilisateurs

(Chapitre **??**), dans un format lisible par la machine (Chapitre **??**), en utilisant un processus automatisé limitant les interventions manuelles (Chapitre **??**).

Troisième partie

Solutions de la recherche reproductible

8

Le temps des changements ?

Nous espérons que les anecdotes du chapitre **??** vous ont semblé vraisemblables, mais vous souhaitons de ne pas trop vous reconnaître dans ces situations de crise ! Nous essayons de vous proposer d'intervenir de manière curative, mais aussi préventive.

8.1 Quand mettre en œuvre les bonnes pratiques ?

Il existe des solutions applicables à court comme à plus long terme. Si le déploiement de certaines solutions nécessite des compétences techniques avancées, la reproductibilité de la recherche demeure peut-être avant tout une question culturelle, dans la mesure où il s'agit de faire évoluer des pratiques parfois solidement ancrées. La crainte de perdre du temps en modifiant ses repères est une problématique majeure, d'où l'approche pragmatique du présent ouvrage. En outre, comment accompagner l'adoption de nouvelles pratiques à l'échelle collective, qu'il s'agisse d'un laboratoire ou d'une communauté disciplinaire ? Par exemple, pour un directeur d'unité ou d'équipe, il pourrait être tentant d'imposer tout ou partie des méthodes décrites en vue d'une plus grande efficience, mais faire évoluer les pratiques quotidiennes de ses collègues constitue un exercice délicat (**?**) (**?**). De fait, une approche très progressive qui valorise l'implication directe des individus a davantage de chances d'aboutir (ou moins de risques d'être rejetée, selon que vous voyez le verre à moitié plein ou à moitié vide).

Au fil de l'ouvrage, nous décrivons des situations de crise, moins par goût pour les histoires d'horreur que par conviction que

l'identification des problèmes couramment rencontrés constitue l'une des meilleures options pour aborder la question de la reproductibilité.

8.2 Concrètement, que changer et comment s'y prendre ?

Bonne nouvelle : il est très peu probable que votre recherche soit à 100% non reproductible. La palette des solutions à votre disposition est variée ; en fonction de vos compétences et de vos urgences, vous pouvez choisir d'agir sur :

— la collecte et la gestion des données,
— le code et sa robustesse,
— les environnements logiciels,
— les sauvegardes,
— les versions et les archives,
— les licences.

Bonne nouvelle bis : le présent ouvrage a été conçu pour être hautement "*cherry-picking-proof*". Selon l'état de vos pratiques actuelles, les chapitres peuvent donc être lus dans l'ordre qui vous conviendra le mieux, même si nous conseillons bien sûr de tous les lire, voire de diffuser cet ouvrage auprès de vos collègues et étudiants (fin de la minute d'autopromotion).

9

Documenter ses pratiques

Cette partie expose une sélection de solutions aux problèmes soulevés dans le chapitre **??**.

9.1 Rédiger un protocole de collecte des informations

Comme nous l'avons vu plus haut (chapitres **??** et **??**), l'information doit être collectée de façon "standardisée". Cela implique de rédiger un protocole décrivant la façon dont l'information est collectée, et ce, donnée par donnée. Par exemple : chez une personne atteinte d'un diabète, vous recueillez la concentration du glucose dans le sang. Vous devez dès lors rédiger le protocole permettant d'obtenir la valeur de la concentration en glucose, en précisant notamment dans le protocole le volume de sang prélevé, à quel moment de la journée est-il prélevé, *etc*. Ce protocole doit non seulement être rédigé mais il doit par ailleurs être approuvé par votre équipe de recherche, ou *a minima*, par les personnes de votre équipe disposant des compétences pour juger de la qualité de ce protocole.

9.2 Partager le protocole de collecte des données

Idéalement, le protocole de collecte des informations devrait faire partie de l'article publié qui relate les résultats de votre recherche. Dans un monde idéal encore, les revues devraient systématiquement

demander aux auteurs de rédiger une telle partie dans l'article afin que d'autres chercheurs puissent éventuellement confirmer les résultats de votre étude. A noter : il ne s'agit pas forcément de rédiger *in extenso* le protocole de collecte des informations dans l'article. Il s'agit plutôt de se demander si, sur la base de votre descriptif, un tiers pourrait appliquer votre protocole de collecte de données et obtenir *in fine* des résultats similaires aux vôtres.

9.3 Tenir un cahier de laboratoire

Modifier des informations ne constitue pas en soi un problème. Les difficultés surviennent lorsque ces modifications n'ont pas été documentées, tracées. Une solution pour pallier ce problème consiste à rédiger un cahier de laboratoire. A l'origine, un cahier de laboratoire est un document physique dans lequel on consigne toute modification d'information au stylo (indélébile) : on indique la raison de la modification de la valeur d'une information, la date de la modification, et l'auteur de la modification. Même si le cahier physique a tendance à disparaître des laboratoires au profit de carnets de laboratoire numériques, la logique reste la même : toute modification des informations initiales doit être renseignée, en indiquant la raison de la modification, la date, et l'auteur de la modification.
Le tableur est-il mon ami pour effectuer ce type de tâche? Pas vraiment. Plutôt pas du tout, dans la mesure où la modification d'une valeur dans une cellule peut être réalisée sans laisser de trace.

9.4 Collecter les données de façon répétable ET reproductible

Le protocole de collecte d'une information doit être défini de telle sorte qu'elle soit "répétable" et "reproductible" au sens du *VIM* (**?**) (cf. chapitre **??**). Tout d'abord, si vous collectez une information

deux fois sur un individu dans un intervalle de temps restreint, les deux valeurs de l'information doivent être les mêmes, ou du moins être suffisamment voisines pour être "considérées" comme identiques (condition de répétabilité). La collecte des informations doit être reproductible entre opérateurs (on parle de "reproductibilité inter-opérateurs"). Pour garantir une bonne répétabilité et reproductibilité inter-opérateurs, il faut donc avoir standardisé au maximum le protocole de collecte des données, en s'appuyant sur un instrument de mesure capable de recueillir les informations avec le minimum d'erreurs possibles. Des indicateurs statistiques, tels que les coefficients de concordance, permettent de quantifier la "répétabilité" ou la "reproductibilité inter-opérateurs" d'une méthode de mesure.

9.5 Pour en savoir plus

Pour en savoir plus sur les cahiers de laboratoire, nous vous invitons à consulter le site du CNRS (**?**). Pour en savoir plus sur la quantification de la répétabilité et de la reproductibilité de la façon dont on collecte les données, vous pouvez consulter le document suivant : "Guide pratique de validation statistique de méthodes de mesure : répétabilité, reproductibilité, et concordance" (**?**).

10

Formater et structurer l'information

Dans ce chapitre, nous allons traiter des formats et structures de fichiers numériques, bien que ces concepts aient aussi leur importance en dehors de l'outil numérique.

10.1 Comment structurer mes informations ?

Les chercheurs peuvent être amenés à travailler sur des données de nature très variée. Si on peut spontanément penser à des nombres, les chercheurs peuvent aussi travailler sur des images ou du texte. Le plus souvent, les chercheurs travaillent en fait sur une "collection" de données liées les unes aux autres. La relation entre ces données est essentielle.

Par exemple, il peut être commode de rassembler des informations sur des patients sous forme d'une "table" comportant une ligne pour chaque individu et une colonne pour chaque type d'information (nom, âge, sexe, nature du médicament administré, dose, taux de glycémie après 1 heure, taux de glycémie après 5 heures). Dans ce cas, le nom et le type de chaque colonne sont souvent considérés comme des "métadonnées". Ce n'est pas la seule façon de représenter ce type d'information. On pourrait préférer avoir une entrée (une ligne) pour chaque mesure (et non pour chaque patient) en ajoutant une colonne indiquant quand la mesure a été réalisée, et une autre indiquant par qui elle a été réalisée.

Un autre exemple de structure est celui d'un arbre généalogique. Au prime abord, il semblerait naturel d'utiliser une "hiérarchie",

mais la vie étant faite de surprises (décès, divorces, re-mariages, voire mariages entre cousins, "Luke, je suis ton père", *etc*.), cette représentation va rapidement s'avérer inadaptée pour concevoir un arbre généalogique clair.

Un dernier exemple d'information dont la conservation est primordiale : il s'agit du protocole expérimental généralement consigné dans un cahier de laboratoire. On structure souvent cette information de façon chronologique avec des annotations sémantiques : qui ? quand ? où ? pourquoi ? dans quel contexte ? *etc*. Dans tous les cas, la question du lien entre ce cahier et les données archivées doit être posée et résolue.

Bref, même si une table, une hiérarchie ou un texte libre peut sembler la solution la plus naturelle, il est vraiment important de bien réfléchir à la façon la plus adaptée de structurer vos informations en fonction du traitement que vous allez vouloir réaliser car cela risque de considérablement affecter ce que vous allez pouvoir faire de vos données.

10.2 Quel format choisir pour enregistrer et stocker des informations ?

L'enjeu autour du format pour la recherche reproductible est double :

— assurer l'interopérabilité,
— minimiser les risques d'erreur de manipulation.

Le chercheur doit donc avoir en tête les bonnes pratiques établies dans sa communauté et s'assurer que son choix de représentation permette, voire facilite, la réutilisation de ses données et de ses résultats. Le nom des fichiers est également un point important du formatage et sera abordé dans le chapitre **??**.

La recherche reproductible vise à limiter drastiquement les interventions manuelles dans le flux de production des résultats. Dans le choix d'un format d'enregistrement et de stockage des

informations, l'objectif est de garantir la "lisibilité par la machine". On devrait donc dire "lisible par toutes les machines" avec en tête, les spécificités des différents systèmes d'exploitation qui peuvent devenir problématiques pour certains formats.

On distingue 3 grands types de formats :

— les formats fermés/propriétaires pour lesquels le risque de perte de lisibilité n'est pas maîtrisé par l'utilisateur, et qui nécessitent que d'autres disposent également du logiciel nécessaire (parfois coûteux) pour pouvoir réutiliser les données.
— les formats codés, illisibles par l'humain, tels que les formats binaires ou de description de page tel que le PDF, qui nécessitent une étape de décodage, et qui peuvent parfois mal supporter la transition entre les systèmes d'exploitation et les architectures matérielles.
— les formats texte (tels que .csv pour les tableaux) qui sont lisibles par les humains comme par les machines, très interopérables, et dont les modifications peuvent être enregistrées par les outils de suivi de version (voir le chapitre **??**).

Par exemple, pour des tables de données simples, mieux vaut privilégier les formats .csv ou .tsv plutôt que les versions plus ou moins propriétaires ou spécifiques à un tableur (.dot, .xls, .xslx, ...) qui peuvent parfois contenir des informations très difficiles à lire pour la machine (cellules colorées, cellules fusionnées, *etc*.).

Dans la mesure du possible, nous préconisons de privilégier les formats texte lisibles à la fois par l'humain et par la machine et d'éviter les formats propriétaires et codés. Toutefois, ce n'est pas toujours possible. Dans ce cas, il est préférable d'utiliser les standards de sa communauté plutôt que des formats exotiques.

Pour des données plus complexes, hétérogènes, de type hiérarchique, des formats adaptés, ouverts, et interopérables existent comme par exemple .yml, .json, ou .xml pour des formats textes et .hdf5 ou .fits ou pour des formats binaires.

10.3 La présentation des résultats numériques

Lors du stockage de données numériques, il est primordial d'éviter la perte ou l'érosion de l'information. Ceci implique, outre une documentation exhaustive précisant les unités et la provenance des résultats, de gérer correctement leur représentation numérique (**?**).

10.3.1 Nombre de chiffres significatifs

Les calculs numériques sont effectués avec une précision finie, et il faut donc choisir le nombre de chiffres significatifs à reporter dans une table de données ou de résultats. Dans un fichier de résultats, il est tentant d'inclure tous les chiffres significatifs dont on dispose. Mais cela n'est pas necessairement souhaitable (cela peut conduire à une inflation inutile des tailles de fichiers), d'autant plus que ce n'est pas nécessairement la précision dont on dispose réellement : par exemple, R n'affiche pas tous ses chiffres significatifs avec sa commande `print()`.

10.3.2 Incertitude

Les informations devraient idéalement toujours être accompagnées d'une incertitude. Cela s'applique à la fois aux mesures (qu'elles soient physiques ou virtuelles), ainsi qu'aux résultats d'analyse (par exemple des estimations) (**?**).

L'incertitude peut servir de guide pour choisir le nombre de chiffres significatifs. Par exemple la recommandation en métrologie (**?**) est d'arrondir (par excès) l'incertitude à deux chiffres significatifs, et de reporter le résultat au même niveau décimal. Par exemple, si le résultat de mesure vaut 1.23456789 et l'incertitude vaut 0.00456, on reportera 1.2346 avec une incertitude de 0.0046. En outre, on évitera dans un tableau les notations du type 1.2346(46) ou 1.2346 ± 0.0046, qui peuvent fragiliser la lecture automatique par une machine.

Une attention particulière doit être portée à certains objets afin de respecter leurs propriétés intrinsèques. Par exemple, les éléments d'une matrice de variance-covariance doivent être arrondis de manière à s'assurer que celle-ci reste définie-positive (en exigeant par exemple que la plus petite valeur propre de la matrice garde deux chiffres significatifs). Voir "Definition" (3.21)(**?**).

Les valeurs et les incertitudes devraient être présentées dans des colonnes séparées.

10.4 Pour en savoir plus

En ce qui concerne la structuration des données, une approche assez populaire consiste à utiliser le plus possible une structure de tableau. Nous vous recommandons de lire ce document sur le sujet : *Tidy data* (**?**).