Gérer les données d'un travail de recherche

JcB

07/03/2015

Contents

Les données (data) sont le corolaire indiscociable de tout travail de recherche. Elle peuvent exister avant de poser la question de recherche (travail rétrospectif) ou apparaitre après que la question soit poseé (travail prospectif). La nature et la pertinence des données à recueillir relève de la QR. On s'intéresse ici à la gestion de ces données.

La gestion des données soulève les questions suivantes:

- comment stocker et organiser mes données ?
- comment rendre mes données persistantes ?
- comment partager mes données avec un statisticien?
- comment exploiter statistiquement mes données ?
- comment s'inscrire dans une démarche de recherche reproductible ?

1 Organiser ses données

Votre travail de recherche est important et des données bien organisées faciliteront le travail du statisticien et de vos pairs qui doivent pouvoir reproduire et évaluer vos résultats (ref[1]).

1.1 Cycle de vie des données

- données brutes (raw data)
- données désorganisées (messy data)
- donnée exploitables (tidy data)

1.2 Formats de fichier [1]

Certains formats de fichiers sont plus adaptés à la conservation que d'autres. Par exemple, certains formats sont propriétaires, ce qui signifie que leur structure ne est pas accessible au public pour la documentation ou la réplication. Si vous enregistrez vos données dans un tel format et que la société qui le détient cesse d'exister, il peut être impossible de les récupérer. Idéalement les données:

- doivent rester accessibles dans le futur (20 ans)
- dans un format non propriétaire
- respecter des standards ouverts, documentés
- être commun, utilisé par la communauté de recherche pertinent ou des communautés;
- utiliser des polices de caractère ouvertes et standard (ASCII, Unicode);
- ne pas dépendre d'un support unique pour ˆetre affichées ou manipulées.
- si possible ni cryptées, ni compresées.

Les standards ouverts sont par exemple:

- texte: ASCII ou Unicode (UTF8)
- Pour l'audio: MPEG-4, pas Quicktime;
- Pour les images: TIFF ou JPEG2000 ou PNG, (pas GIF ou JPG)
- Pour données structurées: XML, JSON ou RDF (pas SGBDR)
- traitement de texte: libre office (pas Word)
- tableur: libre office (pas excel)

1.3 Nommage des fichiers [1]

Le nom donné aux fichiers est très important car c'est par cet identifiant que les données sont lues par un logiciel statistique et que l'on peut savoir si la version utilisée est la bonne. Certaines disciplines ont défini des standard très élaborés (DOE's Atmospheric Radiation Measurement (ARM) program).

Voici quelques recommandations:

- Utilisez un système de nommage des fichiers qui reste cohérent.
- la lecture du nom doit ˆetre suffisament explicite pour rester comréhensible dutant plusieurs années. Par exemple: projet_questionaire_lieu_date_version.ext
- Évitez ces symboles dans les noms de fichiers: "/ : *? "<> [] & $. Ces caractères ont une signification particulière dans certains systèmes d'exploitation d'ordinateur qui pourrait entraîner une mauvaise lecture ou la suppression de ces fichiers.
- évitez les caractères accentués qui sont mal supportés par les logiciels statistiques
- Utilisez le caractère soulignement (_), pas des espaces pour séparer les termes.
- Essayez de garder les noms de dossiers courts (15-20 caractères ou moins) et descriptif de ce qui est à l'intérieur.
- Essayez de garder les noms de fichiers de moins de 25 caractères.
- Inclure les dates dans les noms de fichiers ou de dossiers; utiliser le format AAAA-MM-JJ.
- Si vous ne utilisez pas le versionnage automatique, inclure un numéro de version à la fin du nom de fichier tel que v01. Changez ce numéro de version chaque fois que le fichier est enregistré. Ne utilisez pas d'étiquettes confusion: la révision, finale, final2, etc.
- Pour la version finale, remplacer le dernier mot pour le numéro de version. (Ceci est particulièrement important si les fichiers sont partagés.)
- Les noms de fichiers doivent contenir qu'un seul point, avant l'extension de fichier (par exemple name_paper.doc PAS name.paper.doc)

1.4 Sauvegardez vos données [1]

Il est généralement conseillé de faire trois copies de vos données. Par exemple, vous pourriez avoir le jeu original, une copie sur un disque dur externe local, et une autre copie sur un disque externe situé ailleurs. Ces copies doivent avoir des localisations géographiques différentes. Assurez-vous que le système de back-up est fiable et que les données sont réllement récupérables. Le "Cloud" est une solution envisageable (Google drive, Bit Bucked).

1.5 Sécurisez vos données

En général le cryptage des données n'est pas conseillé. Cependant il peut ˆetre indispensable pour certaines données sensibles (données nominatives, personnelles, recommandation de la CNIL) de les crypter. Gardez des mots de passe et les clés sur le papier (2 exemplaires), et dans un fichier numérique crypté comme PGP (Pretty Good Privacy). Ne comptez pas sur le cryptage 3ème partie seul.

Il est également recommandé de stocker des données dans un format non compressé. Si vous avez besoin d'économiser de l'espace, nous recommandons de limiter la compression à votre copie de sauvegarde.

2 Utiliser un tableur [2]

Les données sont généralement conservées sur 2 types de support, les tableurs et les bases de données. Il est aisé de passer de l'un à l'autre. Le tableur est la solution unervesellement adoptée du fait de sa disponibilité immédiate et de son utilisation intuitive. Cependant un certain nombre de règles doivent ˆetre respectées.

La première (et très importante): un tableur doit servir à saisir des données et ne servir qu'à cela. La plupart des ennuis commencent lorsqu'on pense qu'un tableur peut tout faire, en particulier des statistiques. Il faut abandonner cette idée et se placer uniquement sous l'angle de la saisie des données. Pour faire des statistiques, il y a des logiciels pour celà.