UNE DÉMARCHE QUALITÉ AU SERVICE DE LA PRÉPARATION DE DONNÉES DE RECHERCHES EN SCIENCES SOCIALES.

Cédric Gendre, Christophe Bontemps
pour le groupe de travail “démarche qualité” du CATI INRA CITISES.
US-ODR INRA

7èmes journées de recherches en sciences sociales

INRA SFER CIRAD

Angers, 12-13 décembre 2013

Résumé. Les données sont des composantes stratégiques de nombreuses recherches. Avant d'entrer dans le processus de recherche proprement dit, ces données font l'objet de différents traitements (ou prétraitements) décrits au travers des étapes dites du « cycle de vie » des données. La fiabilité et la traçabilité de ces étapes, autrefois souhaitables, sont désormais nécessaires non seulement par souci d'efficacité, mais aussi parce que la profession l'exige. Par ailleurs, les données traitées peuvent être mises à disposition de différents partenaires qui, au même titre que la recherche, sont en droit d'exiger une qualité et une reproductibilité des processus et des traitements qui ont servi à l'élaboration de ces données. Compte tenu de l'investissement que représentent ces données, il apparaît indispensable d'en assurer la pérennité et de mettre en place des procédures visant à conserver ce patrimoine. Aussi la mise en œuvre d'une démarche qualité de la préparation des données pour la recherche est-elle indispensable. Elle vise à promouvoir l'homogénéisation des pratiques et garantir une qualité des processus de traitement. Nous proposons ici trois outils pour en promouvoir la diffusion :

- une charte au travers de laquelle le gestionnaire des données s'engage à respecter un certain nombre de principes
- un guide des bonnes pratiques qui, pour chacune des étapes du cycle de vie des données, détaille différents outils et leur mise en œuvre pratique
- un pense-bête qui permet une auto-évaluation en repérant les étapes qui mériteraient d'être améliorées.

Mots-clés. Qualité, réplication, traçabilité, data management

1 Introduction

Après qu'un collègue ou un lecteur de votre article vous ait posé une question sur un résultat que vous avez obtenu, vous passez un temps considérable pour trouver les fichiers des programmes que vous avez utilisés pour le produire. Sans parler du temps pour comprendre ce que vous aviez vraiment fait dans ces programmes pour produire ce fameux résultat.

Parce que ce type de situations est familier, nous assistons depuis plusieurs années à un changement important dans les pratiques de la communauté scientifique. Cette dernière tend à promouvoir la « reproductibilité de la recherche » sous l'impulsion du mouvement « *Reproducible Research*[1] ». Suivi par des éditeurs de grandes revues (*Review of Economics and Statistics*, *Journal of Applied Econometrics*, etc..), ce processus se propage en amont, accentuant le contrôle sur toutes les étapes du processus de recherche menant à publication[2]: vérification des sources, vérification ou contrôle des résultats tirés de ces données, stockage dans des archives publiques ou privées, développement de plateformes.

Ce phénomène a conduit différentes institutions et universités (e.g. *National Science Foundation*, *American Statistical Association*, *Australian National University,* …) à proposer des règles ou des principes (*data management policies, ethics policies, data management plans, etc...*) s'appliquant aux chercheurs et ingénieurs traitant des données.

Au-delà de ces incitations institutionnelles, des études tendent à prouver que les articles permettant

1 Voir par exemple la première rencontre autour de la recherche reproductible organisée à l'université d'Orléans en avril 2012 (http://www.fdpoisson.fr/cascimodot/doc/RRRR/R4-050412.php) ou le site Reproductible Research (http://reproducibleresearch.net/).

2 Ainsi de nombreuses revues (comme *l'European Review of Agricultural Economics*, *Economics Letters* ou *Sociologie du travail* par exemple ont adhérées au *Comittee On Publication Ethics* (COPE) et suivent les recommandations de cet organisme concernant les problématiques liées aux données (cf http://publicationethics.org/category/keywords/data-manipulation-/-falsification).

un accès aux données sont plus cités que les autres[3]. Enfin, avec le développement de l'informatique scientifique et des possibilités de collaborations à distance entre chercheurs et ingénieurs au sein d'un même projet, des méthodes et des outils se sont développés. Cette nouvelle façon de travailler est en train de s'imposer à tous les praticiens, et permettra à terme d'améliorer l'ensemble du processus de recherche, d'en améliorer l'efficacité et d'en augmenter la portée. Nous proposons ici des préconisations afin de faciliter la diffusion des outils et des méthodes menant à une meilleure maitrise des processus de gestion des données

2 Des préconisations pour qui ?

Ces préconisations s'adressent à toute personne qui traite des jeux de données quotidiennement, soit dans le cadre de la gestion de systèmes d'information, soit dans le cadre de travaux scientifiques. Les préconisations se centrent sur le traitement initial des données, dans cette zone généralement peu visible (le nuage de la figure 1), en amont du processus d'analyse proprement dit.
Les ingénieurs débutants et les doctorants y trouveront une aide pour structurer et organiser leurs travaux à partir de données brutes et y prendront de bonnes habitudes, puisque cette étape se reproduit dans chaque projet. Les ingénieurs et chercheurs confirmés pourront s'en servir pour analyser leurs pratiques en regard d'un objectif de traçabilité et de reproductibilité de leurs travaux et y découvrir de nouveaux outils simplifiant leur approche de ces traitements.

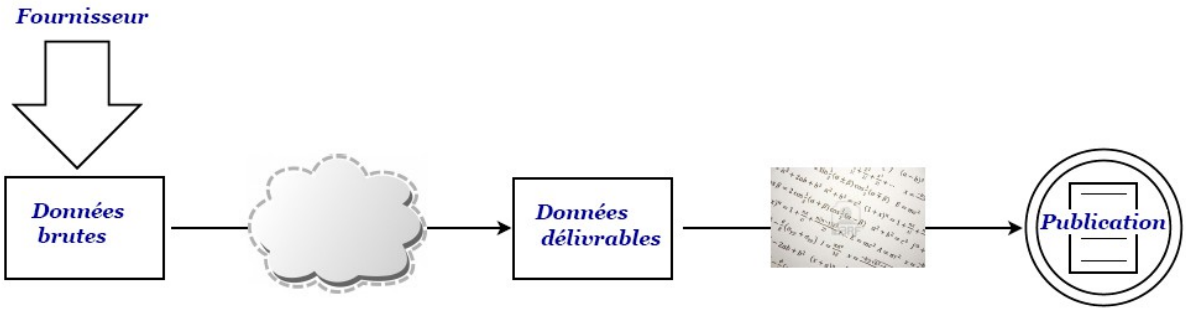

Figure 1 : Processus de recherche

3 En quoi consiste la démarche ?

La démarche qualité de gestion et de préparation des données s'inscrit tout au long du cycle de vie des données (voir figure 2), c'est à dire de la réception des données brutes[4] à leur utilisation dans les projets de recherche ou de leur mise à disposition aux utilisateurs (données délivrables), en passant éventuellement par des étapes de vérification, normalisation et enrichissement.
Les données évoluent au cours du temps, et passent de main en main dans un cycle dont une étape est représentée schématiquement par la figure 2. Il est à noter que ce processus se répète parfois, les fournisseurs pouvant se succéder les uns aux autres.

2 Exemple de références bibliographiques

La nécessité de produire des résumés clairs et bien référencés a été démontrée par Achin et Quidont (2000). Le récent article de Noteur (2003) met en évidence . . .

[3] Voir la page de Gary King à ce sujet http://gking.harvard.edu/pages/data-sharing-and-replication.

[4] Même si les données sont parfois collectées, nous nous plaçons ici à l'étape de réception des données, c'est à dire que nous considérons les données déjà disponibles.

Bibliographie

[1] Auteurs (année), Titre, revue, localisation.
[2] Achin, M. et Quidont, C. (2000), Théorie des Catalogues, Editions du Soleil, Montpellier.
[3] Noteur, U. N. (2003), Sur l'intérêt des résumés, Revue des 0rganisateurs de Congrès, 34, 67–89.

Bibliographie

[1] American Statistical Association (1999). "*Ethical Guidelines for Statistical Practice*." ASA. http://www.amstat.org/about/ethicalguidelines.cfm

[2] ANU Data Management Manual "*Managing Digital Research Data at the Australian National University*"(2012) http://information.anu.edu.au/training_and_skills_development/information_literacy/resources/ANU_DM_Manual-v11.09.20_v2.pdf

[3] Dublin Core Metadata Initiative Web site. http://dublincore.org/

[4] Eurostat (2005), « *Code des bonnes pratiques de la statistique européenne pour les services statistiques nationaux et communautaires* » http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-77-07-026/FR/KS-77-07-026-FR.PDF

[5] Groupes de travail communs UNECE/Eurostat/OCDE sur les métadonnées statistiques (METIS), 2009. *Modèle générique du processus de production statistique* (version 4.0)

[1] Inter-university Consortium for Political and Social Research (ICPSR), 2012. *Guide to social science data preparation and archiving. Best practice troughout the data life cycle* (5th ed.). Ann Arbor, MI. http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Lyman, Peter and Hal R. Varian, (2003) "*How Much Information?"* University of California., Berkley http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/

Pienta, Amy M., Alter George , Lyle Jared (2010) *The Enduring Value of Social Science Research: The Use and Reuse of Primary Research Data* Inter-university Consortium for Political and Social Research (ICPSR) , Population Studies Center. Working paper. http://deepblue.lib.umich.edu/bitstream/2027.42/78307/1/pienta_alter_lyle_100331.pdf

Strasser, Carly, Cook, Robert; Michener, William; & Budden, Amber. (2012). "*Primer on Data Management: What you always wanted to know*". DataOne Working paper. http://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

TGE ADONIS, 2010. *Le guide des bonnes pratiques numériques. Entrepôt OAI-PMH* (version1) http://www.tge-adonis.fr/sites/default/files/ressourcesdoc/pdf_guide_oai10_vf.pdf

TGE ADONIS, 2011. *Le guide des bonnes pratiques numériques.* (version2) http://www.tge-adonis.fr/sites/default/files/ressourcesdoc/guide_des_bonnes_pratiques_v2.pdf

Van den Eynden V., Corti L., Woollard M., Bishop L., Horton L. 2011. *Managing and sharing*

data. Best practice for researchers (3rd ed.) Universityt of Essex. http://www.data-archive.ac.uk/media/2894/managingsharing.pdf

Vardigan, M., Heus, P. and Thomas, W. (2008) *Data documentation initiative: Toward a standard for the social sciences*. Int. J. Digital Cur., 3, 107-113.