IL FAUT GÉRER PUBLIQUEMENT LES DONNÉES SCIENTIFIQUES

LE 1 SEPTEMBRE 2010 ***ANDRE VELLINO (TRAD. MARTIN CLAVEY)***

L'organisation de la publication des données scientifiques est importante pour rendre plus simples les vérifications mais aussi et surtout pour sauver les données qui ne sont pas reproductibles.

Andre Vellino est chercheur au **NRC Canada Institute for Scientific and Technical Information** et professeur invité à l'**École de l'Information de l'Université de Ottawa**.

Les données de recherches scientifiques sont sans nul doute un composant central du cycle de vie de la production de la connaissance. D'une part, les données scientifiques sont essentielles à la **corroboration (ou la falsification)** des théories. Mais aussi, l'accès ouvert à ces données est tout aussi important pour que le processus de validation scientifique soit mis en œuvre (comme cela a été récemment démontré par la controverse du « **ClimateGate** » et la récente affaire sur **les données de recherche de la cognition chez les primates de Marc Houser**). L'accessibilité publique des données permet une relecture ouverte par les pairs et encourage la reproductibilité des résultats.

De ceci découle l'importance des pratiques de gestion des données dans les bibliothèques scientifiques du XXIème siècle : le traitement éditorial, l'accès et la préservation des données de recherches scientifiques vont devenir cruciaux pour le futur du discours scientifique.

Certaines données ne sont pas stockées

Il est vrai que les institutions de recherches scientifiques à grande échelle gèrent des données de références depuis longtemps. Dans beaucoup de disciplines, les centres de données dans ces institutions ont rassemblé un bon nombre de bases de données contenant les fruits d'années de recherches. Par exemple, **GenBank**, la base de données de séquences génétiques du National Institutes of Health (NIH), et une collection annotée et globale de toutes les séquences d'ADN accessible publiquement (plus de 150 000 séquences).

Toutefois, d'autres types de données rassemblées par les scientifiques sont soit éphémères soit très dépendants du contexte et ne sont préservés à long terme pour le bénéfice de recherches futures ni par les institutions ni individuellement par les chercheurs. Ceci n'est pas si important pour les données reproductibles (soit expérimentalement, soit par simulation). Mais beaucoup de données, **comme celles concernant la concentration du pétrole et sa dissipation dans l'eau du golf** du Mexique en 2010, sont uniques et non-reproductibles.

Comme je l'ai indiqué dans un **billet précédent**, l'émergence de méthodes de référencement unique pour les jeux de données comme celles de **DOI** implémentées par les partenaires de **Datacite**, permettrait d'aider à résoudre certains problèmes subit par les petits jeux de données orphelins et inaccessibles. La combinaison de politiques de dépôts de données par les agences de financement de la recherche scientifique (comme **NSF** aux États-Unis et **NSERC** au Canada) et la reconnaissance par les pairs des universités envers l'effort intellectuel effectuer dans la création de données va, dans le futur proche, augmenter tant le nombre de publications de données que leur référencement pour correspondre à la situation actuelle avec les publications savantes.

En parallèle, l'émergence du mouvement de « **l'accès ouvert aux données** » et d'autres initiatives qui augmentent la disponibilité des données générées par les gouvernement et les institutions gouvernementales (dont la **NASA**, le NIH, et la **Banque mondiale**) sont en bonne adéquation avec **les principes de l'OCDE** [PDF]. On y trouve dans celui-ci, incidemment, une liste longue et convaincante des bénéfices économiques et sociaux qui découle du libre accès des données scientifiques.

Un mouvement mondial émerge

Les **États-Unis**, le **Royaume-Uni**, et **l'Australie** sont les fers de lance dans l'effort fait pour rendre les données de recherches scientifiques accessibles. Par exemple, aux États-unis, le récent **rapport** [PDF] du Conseil national de science et technologie (NSTC) rendu au Président Obama détaille une stratégie complète pour promouvoir l'accès aux données numériques et leur préservation.

Ses rapports et initiatives montrent qu'il existe un mouvement mondial pour réaliser et articuler les visions de plusieurs organismes concernés par le traitement et l'archive de données qui s'est créé pendant la première décennie du XXIème siècle (voir **To Stand the Test of Time** [PDF] et **Long Lived Scientific Data Collections** [PDF]).

Au Canada, d'autres rapports similaires comme **la Consultation sur l'accès aux données de recherches scientifiques** [PDF] et la **Stratégie canadienne de l'information numérique** soulignent également le besoin d'un conseil national pour la préservation de l'information numérique, y compris les ensembles de données scientifiques. Malgré beaucoup de discussions, les efforts systématiques dans dans la gestion des données scientifiques canadiennes ne sont qu'à leurs premiers stades. Alors que les données dans certains domaines sont bien préservées et bien gérées, comme en sciences de la terre (avec **Géogratis**) et en astronomie (avec le **Canadian Astronomy Data Centre**) qui ont des communautés d'utilisateurs scientifiques spécialisés et dont les besoins sont bien compris, les besoins en gestion de données des scientifiques esseulés dans des petits groupes de recherches mal financés sont soit impossible à trouver soit déjà perdu.

Un obstacle au traitement éditorial bibliographique effectif des jeux de données est l'absence de normes communes. Il n'y a pour l'instant « aucune règle de publication, de présentation, de citations ou même de catalogue de jeux de données » (OCDE) [Green, T (2009), "**We Need Publishing Standards for Datasets and Data Tables**", OECD Publishing White Paper, OECD Publishing]

La Passerelle vers les données scientifiques de l'Institut canadien de l'information scientifique et technique (ICIST) ainsi que d'autres sites nationaux (tel que **British National Archives of Datasets**) qui rassemble des informations à propos des jeux de données, utilisent des normes bibliographiques (tel que **Dublin Core**) pour représenter les méta-données. L'avantage est que ces normes ne dépendent pas du domaine et sont déjà suffisamment riches pour exprimer les éléments clés dont on a besoin pour archiver et retrouver les données. Toutefois, ces normes de méta-données développées pour la bibliothéconomie traditionnelle, ne sont pas (encore) suffisamment riches pour récupérer complètement la complexité des données scientifiques venant de toutes les disciplines, comme je l'ai fait valoir dans un **précédent billet**.

Lorsqu'on décide de la faisabilité de la création d'un dépôt de données, une des inquiétudes majeures est le coût lié au dépôt, au traitement et à la préservation à long terme des données de recherches. Typiquement, les coûts dépendent de nombreux facteurs incluant la façon dont les phases caractéristiques (planification, acquisition, mise à disposition, organisation, traitement, archivage, stockage, préservation et les services d'accès) sont déployées (voir les rapports du JISC "Keeping Research Data Safe" **Part 1** and **Part 2**). Les coûts associés à différentes collections de données varient aussi considérablement selon la rareté ou la valeur des données et selon exigences pour l'accès au fil du temps.

Un point à noter venant des rapports "Keeping Research Data Safe" est que :

« les coûts d'archivage des activités (archivage, stockage et préservation, planification et actions) sont une toute petite proportion des coûts globaux et sont très légers comparés aux coûts d'acquisition/traitement et d'accès des données."

En bref, la gestion de bibliothèques de données est critique pour l'avenir de la science et les coûts de la technologie nécessaire pour cette gestion sont la moindre de nos préoccupations.

Cet article est une traduction d'un billet publié sur **Synthèse** par Andre Vellino

Illustration FlickR CC : **scott ogilvie**, **hexodus**

CARROLL B. MERRIMAN
le 16 novembre 2011 - 19:51 • SIGNALER UN ABUS - PERMALINK

I actually found your site through a list on a different blog, and have enjoyed perusing several of your posts. Though Il faut gÃ©rer publiquement les donnÃ©es scientifiques » OWNI, News, Augmented has been my most lovedso far! I have added your feed and aim on coming back to see what you write about next.

VOUS AIMEZ 0 VOUS N'AIMEZ PAS 0 LUI RÉPONDRE

2 pings

La difficile accessibilité des données scientifiques « meridianes le 26 juillet 2011 - 7:26

[...] gouvernementales (dont la NASA, le NIH, et la Banque mondiale) sont en bonne adéquation avecles principes de l'OCDE [PDF]. On y trouve dans celui-ci, incidemment, une liste longue et convaincante des bénéfices [...]

Les scientifiques découvrent que seuls 20% des données climatiques sont accessibles « Les moutons enragés le 27 septembre 2011 - 2:09

[...] gouvernementales (dont la NASA, le NIH, et la Banque mondiale) sont en bonne adéquation avecles principes de l'OCDE [PDF]. On y trouve dans celui-ci, incidemment, une liste longue et convaincante des bénéfices [...]