DONNÉES LIBÉRÉES, CHERCHEURS DÉBRIDÉS, SOCIÉTÉ IMPLIQUÉE

LE 8 DÉCEMBRE 2010 ***ANTOINE BLANCHARD***

Si les chercheurs avaient déjà pris conscience de la nécessité de libérer leurs publications, ils avaient négligé leurs données. Les scientifiques s'aperçoivent qu'il faut organiser leur diffusion.

Jusqu'à très récemment, les données étaient les parents pauvres de la recherche scientifique, particulièrement en biologie. S'accumulant dans les laboratoires et les centre de séquençage du génome, isolées sur le disque dur des chercheurs, elles étaient invisibles et difficilement accessibles — enfouies sous la montagne d'articles scientifiques auxquels elles contribuent à donner naissance. Et si les acteurs ont pris conscience de la nécessité de libérer leurs publications, notamment à travers l'**accès libre aux résultats de la recherche** (*open access*), ils avaient encore négligé leurs données.

Données libérées

Mais que sont exactement ces données ? Ce sont les résultats des expériences ou observations menées par les chercheurs, préludes à de nouvelles découvertes, à de nouvelles théories explicatives et à des publications scientifiques. Depuis la Révolution scientifique, elles étaient conservées consciencieusement afin de garantir la transparence inhérente à la recherche, mais rarement mises en commun. À ce titre, les publications scientifiques faisaient état des résultats de l'analyse de ces données sans jamais mettre les données elles-mêmes, brutes, à disposition. Ou quand elles le faisaient, elles étaient "emprisonnées" dans le format étroit et inutilisable du papier. Seuls des contacts informels entre chercheurs permettaient de s'échanger des données qu'Internet, désormais, permet de mettre à disposition de tous d'un seul clic.

Avec Internet, nous assistons en effet au développement de l'e-Science. Sous ce terme se cache un phénomène général d'appropriation par les chercheurs des technologies de l'information afin de **démultiplier les possibilités d'analyse, de stockage, de publication et de partage des données** mais aussi des articles, et autres résultats de la recherche ; sous ce nom se cache aussi des programmes institutionnels de promotion de cette nouvelle ère de la science, comme en Allemagne ou en Grande-Bretagne (**National e-Science Center**). C'est ainsi que l'on voit grossir les rangs des bases de données accessibles sur Internet, des chercheurs qui stockent leurs articles en accès-libre sur le serveur de leur laboratoire, des logiciels libres (open source) qui permettent d'exploiter ou d'analyser les résultats d'une puce ADN ou d'une analyse phylogénétique.

Ce mouvement a véritablement décollé lorsque les généticiens du Consortium international de séquençage du génome humain se sont réunis aux Bermudes en février 1996 pour convenir, ensemble, du mode de publication de la séquence du génome humain. Ils se mirent d'accord sur la diffusion automatique (et si possible sous 24h) des séquences de plus de mille bases, la publication immédiate des séquences annotées, le tout en accès libre. Ces "principes des Bermudes" régissent depuis lors (avec quelques modifications) les projets de séquençage financés sur fonds publics et sont les fondements de la publication de données génomiques sur Internet.

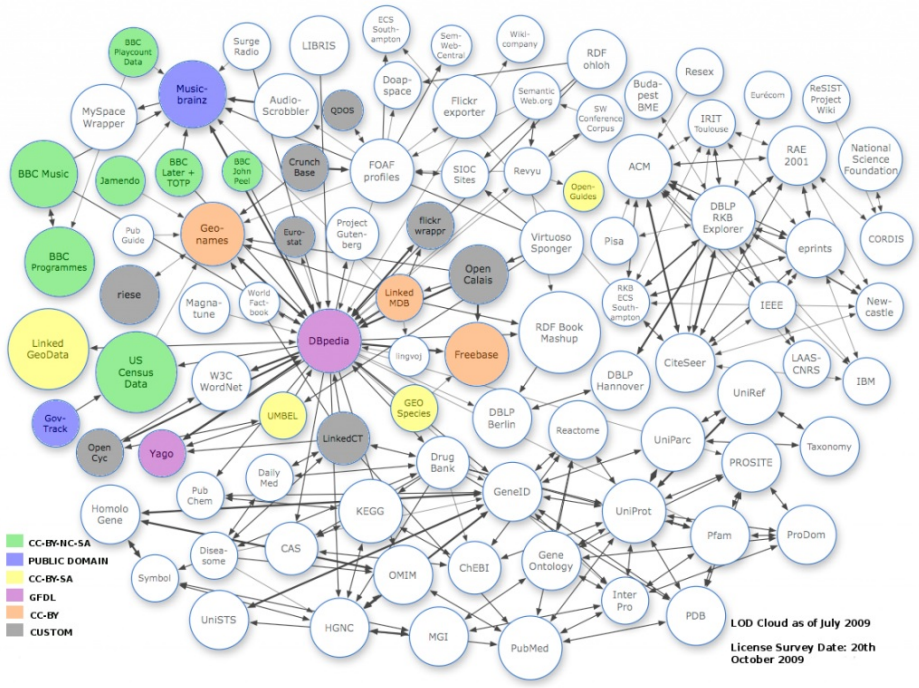
Dans la perspective de l'e-Science, les données autant que les articles doivent être ouverts, libérés. C'est d'une importance croissante avec la collaboration d'équipes à travers le monde entier, dans des projets pharaoniques comme l'étude du virus de la grippe aviaire H5N1. D'autant que les données nécessaires sont rapidement volumineuses, comme l'ont appris à leurs dépens les chercheurs en génétique, neurosciences et plus largement des sciences de la complexité.

Parmi elles, la biologie intégrative cherche à corréler les données issues de différents niveaux d'observation et donc à croiser des données tierces pour les mettre en perspective, les fouiller par des méthodes de fouille de données (data mining) etc. Cette nécessaire libération des données se joue à trois niveaux : l'accessibilité des données, qu'encouragent notamment les revues qui imposent aux auteurs de déposer leurs données — moléculaires, génétiques, écologiques… — dans les bases de données disponibles sur Internet, et qui s'opposent parfois à l'appropriation de droits de propriété intellectuelle ; l'inter-compatibilité des nombreuses bases de données et l'existence de métadonnées standardisées permettant de connaître leur origine, le contexte de leur obtention, etc., défendue par les bio-informaticiens ; et enfin, l'absence de frein légal à la réutilisation des données, par exemple en utilisant une licence libre de type **Science commons** comme le fait UniProt, la plus grande base de données sur les protéines.

Chercheurs débridés

Les chercheurs, acteurs de ce mouvement, en sont aussi les premiers bénéficiaires et ont y vite vu l'occasion de débrider leur créativité, comme le montrent de nombreux exemples récents. En 2005, le lancement de la base de données de structures chimiques ouverte **PubChem** gérée par la National Library of Medicine américaine, a augmenté la visibilité et l'accessibilité de ce type de données sur Internet. Une aubaine pour le chercheur qui dispose ainsi d'un offre plus grande, en terme de contenu et de moyen d'accéder à l'information qui l'intéresse. Ou encore, la mise à disposition du logiciel d'images satellitaires et données géographiques Google Earth — qui couvre toute la planète avec une excellente ergonomie et précision — a inspiré de très nombreuses utilisations originales. Le journaliste de Nature Declan Butler a par exemple construit une **représentation spatiale des foyers de grippe aviaire** (fichier .kml), en croisant la puissance de Google Earth avec les données de l'Organisation mondiale de la santé et de l'Organisation des Nations Unies pour l'alimentation et l'agriculture.

On nomme cette intégration automatisée de données complémentaires en provenance de plusieurs sources le "mash-up", sur lequel parient aussi les écologues avec leur projet **iSpecies** qui intégre, pour chaque espèce recherchée, des données bibliographiques issues de Google Scholar, des données iconographiques issues de Yahoo Images et des données taxinomiques du National Center for Biotechnology Information. Le mash-up autorise aussi l'intégration des disciplines et échelles biologiques pour mieux comprendre des systèmes biologiques complexes comme le cerveau ou la biodiversité ; ces deux thématiques font l'objet d'une politique d'incitation dans ce sens, sous l'impulsion d'un rapport d'un groupe de travail de l'OCDE. Dans le second cas, cela se concrétise notamment à travers l'initiative **Global Biodiversity Information Facility**.



L'ouverture des bases de données et la standardisation des normes de stockage et d'annotation permet d'envisager la fouille de données, utile pour faire parler des données volumineuses et en tirer des informations complètement nouvelles, par exemple sur les relations entre gènes et maladie, facteurs environnementaux et maladie ou autres. Sans compter que ces pratiques émergentes vont se développer en même temps qu'elles se banaliseront et que les outils existants s'y adapteront.

Enfin, l'existence de bases de données ouvertes est aussi un espoir pour les chercheurs des pays en voie de développement qui ne peuvent se payer l'accès aux bases de données privées et payantes. Ainsi, l'apparition du PubChem comme possible alternative gratuite à l'incontournable base Chemical Abstracts a été vécue comme un soulagement dans ces pays, et combattue avec vigueur par l'American Chemical Society qui gère Chemical Abstracts.

Société impliquée

De manière plus inattendue, l'ouverture des données issues de la recherche peut contribuer à impliquer de nouveau la société civile dans la science. D'abord parce que c'est une manière de lui retourner les résultats de la recherche qu'elle finance indirectement par ses impôts. Ensuite parce que c'est un moyen d'intéresser le public à la science, comme le suggère l'engouement incroyable autour du logiciel Google Earth et son appropriation par le jeu des mash-ups (voir des exemples sur le **Google Earth Blog** ou **Ogle Earth**. À ce titre, la **transposition par la France** (avec plus d'un an de retard) de la **directive européenne Inspire** qui prévoit l'ouverture des données géospatiales est révélatrice d'un changement de mentalités.

On peut y voir aussi un facteur d'implication des citoyens dans les processus de décision de la recherche et de l'avènement d'une démocratie scientifique. Dans un modèle de science entièrement ouverte comme l'expérimente le forum **Synaptic Leap**, des chercheurs du privé et du public travaillent ensemble sur des médicaments contre certaines maladies tropicales négligées, notamment le paludisme et les schistosomoses ; tout le projet est mené en ligne, les données y sont publiées et les résultats sont sous licence libre. Et les citoyens de proposer des thèmes de recherche, **comme la chikungunya**.

Ou encore, s'ils s'engagent dans des débats de plus en plus complexes scientifiquement, ils gagnent à avoir accès aux mêmes données que les chercheurs — à condition qu'ils aient la culture scientifique nécessaire à une bonne interprétation de ces données. **L'affaire du "climategate"** marquée par le piratage des serveurs d'un laboratoire de recherche sur le climat en Grande-Bretagne, a montré que la science avait plus à perdre — notamment en terme de confiance — à cacher ses données qu'à les rendre publiques. Après avoir refusé à plusieurs reprises d'obtempérer à des **demandes relevant du Freedom of Information Act**, ce laboratoire **a lancé une réflexion** sur l'ouverture de ses données, en partenariat avec le British Atmospheric Data Centre, habitué de la publication des données sur le climat.

Finalement, après l'accès libre aux articles scientifiques qui se dessine depuis le début des années 2000, l'accès libre aux données de la recherche est la deuxième étape de cette ouverture de la science en marche. Il ne fait aucun doute que ce mouvement va s'amplifier et s'imposer dans le futur, et il appartient aux chercheurs mais aussi aux citoyens de le soutenir. L'étape suivante sera l'accès libre aux technologies — à commencer par les biotechnologies —, mise en commun nécessaire pour sortir du piège du "tout-brevet". Cette étape s'ébauche déjà à travers quelques projets et devrait avoir *in fine* un impact considérable sur la science mais aussi l'économie et la société.

Pour en savoir plus

Brooksbank, Cath et Quackenbush, John. **Data standards: a call to action**. *OMICS: A Journal of Integrative Biology*. 10(2): 94-99, juin 2006

Butler, Declan. Mashups mix data into global service. *Nature*. 439:6-7, 5 janvier 2006

Commission européenne, ***Riding the wave – How Europe can gain from the rising tide of scientific data***, octobre 2010

Cragin, Melissa. **Toward Integrative Science: Organizing Biodiversity and Neuroscience Data**. *Bulletin of the American Society for Information Science and Technology* . 30 (1), octobre/novembre 2003

OECD Megascience Forum Working Group on Biological Informatics. **Final Report of the OECD Megascience Forum Working Group on Biological Informatics**, janvier 1999

Rumble Jr, John et al. **Developing and Using Standards for Data and Information in Science and Technology**. Congrès Ensuring Long-term Preservation and Adding Value to Scientific and Technical Data, The Royal Society (Edinburgh), 21-23 novembre 2005

>> **Article initialement publié sur le blog "Mon bouillon"**

>> Illustration CC **Unhindered by Talent**, **Idodds** et **Jerrycharlotte**

>> Retrouvez tous les articles d'**OWNIsciences**

La partie “que sont exactement ces données ?” évite une question primordiale à mon sens : quelles données peuvent être partagées ? Je ne parle pas de confidentialité, mais bel et bien de technique. La réponse est en filigrane de cet article: génome, transcriptome… tout simplement parce que ces données sont dans un langage commun (le code génétique, etc.). Or ce ne sont des “données” que pour un petit nombre de personnes : bio-informaticiens, qui d’autre ?
Pour tous les autres chercheurs, les “données” sont intimement liées à l’expérience qui a permis de les obtenir, et sont de nature variée (gels, images de microscopie, données fMRI, etc.). Elles sont par conséquent difficile à intégrer dans une base de données, et surtout inutilisables pour celui qui n’a pas conçu l’expérience. Quelle est donc la solution ?
Dans les neurosciences, un domaine que je connais, il serait plus juste de parler de mise à disposition des résultats plutôt que des données : par exemple faire en sorte que les chercheurs qui font de l’imagerie fonctionnelle se mettent d’accord sur un référentiel à utiliser pour présenter leurs résultats (les régions du cortex ne sont pas définies de manière certaine).

VOUS AIMEZ 0 VOUS N'AIMEZ PAS 0 LUI RÉPONDRE

ARNAUD
le 8 décembre 2010 - 16:38 • SIGNALER UN ABUS - PERMALINK

@Piero:
L’intégration de données sous diverses formes comme vous en listez des exemples, est un “détail” technique. Elle est toujours possible d’une manière ou d’une autre.

Mais vous avez raison la mise à disposition des données doit être reformulé dans certains cas en une mise à disposition de résultats. Je pense que l’article en parlant de “données” ne tient pas en compte la manière manière dont elles sont persistées ou ce qu’elles représentent. Une donnée peut correspondre à une simple chaine de caractères ou à tout un ensemble de “sous-données” formant un résultat.

Vous avez parfaitement raison, il manque notamment en Biologie, une standardisation des résultats par domaine. Mais ce qu’il manque encore plus est une standardisation unifiée des résultats entre tous les domaines de la biologie, de manière à ce que ces résultats sont réutilisable par des chercheurs dans un domaine différent de celui ayant générés ces résultats. Ceci pourrait permettre une recherche holistique en Biologie par exemple si l’ont veut étudier l’effet d’un principe actif sur le corps humain, on ne peut plus se permettre d’étudier l’effet sur le système nerveux central séparemment de l’effet sur le système cardiovasculaire. Or généralement un chercheur en neurobiologie générera une serie de résultats de type “neurobiologique” et le chercheur en cardiologie générera des résultats de types “cardiologie” que le chercheur étudiant l’effet du principe actif sera incapable de coupler entre eux par manque de standardisation globale.
Et si l’on cherche la petite bête encore plus loin, il manque une standardisation des résultats entre toutes les sciences. Mais je ne pense pas que l’on en soit encore à ce stade là. Un jour peut-être.

VOUS AIMEZ 0 VOUS N'AIMEZ PAS 0 LUI RÉPONDRE

1 ping

Données libérées, chercheurs débridés, société impliquée « La Boussole le 10 décembre 2010 - 12:57

[...] Lire l’article : owni.fr [...]