Du dépôt au partage de données

Antoine Blanchard, consultant en information scientifique chez Deuxième labo
antoine@deuxieme-labo.fr

Journée Thématique du GdR 2647 STIC-Santé CNRS/Inserm
“Avantages et limites d’un dépôt national de données biologiques”
16 mars 2011
Campus des Cordeliers, Paris

Pourquoi déposer des données ?

Pour faire avancer la science

- trop de résultats ne sont pas publiés
- ou ils sont sous-utilisés (“hypopublication”)

http://figshare.com

“the current scientific literature (…) contains huge amounts of undiscovered science”

Peter Murray-Rust (2007), “Data-Driven Science - A Scientist's View”, http://www.sis.pitt.edu/~repwkshop/papers/murray.html

Une responsabilité déontologique

- Committee on Responsibilities of Authorship in the Biological Sciences, National Research Council. *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences* (National Academy of Sciences, 2003)

Un mandat politique et stratégique

Un engagement institutionnel

Janvier
2011

Inserm

Sharing research data to improve public health: full joint statement by funders of health research

Introduction

Recent advances in information technology have revolutionised science - providing new opportunities for researchers to share data and build on one another's work. Informatics and the ability to mine large datasets and combine them with information from many other sources present a huge potential to advance developments in public health. The importance of data sharing in advancing health is becoming increasingly widely recognised, and has been strongly endorsed by the H8 group of global health organisations.

In some research fields - such as genetics and physics - data sharing is well-established and has accelerated the progress of research and its application for the public good. In public health research, however, while research collaborations are growing more common, the sharing of data is not yet the norm, even within the scientific community.

Much of the data collection that could improve public health research is expensive and time-consuming. As public and charitable funders of this research, we believe that making research data sets available to investigators beyond the original research team in a timely and responsible manner, subject to appropriate safeguards, will generate three key benefits:

- faster progress in improving health
- better value for money
- higher quality science.

Each funding institution will work within its own legal and operational framework, and we are committed to working towards these goals together. We intend to establish joint working groups where appropriate. We call on governments and other actors that generate routine health service statistics and other types of public health data to adopt a similar approach.

This Statement establishes guiding principles and desired goals. It recognises that flexibility and a variety of approaches will be needed in order to balance the rights of the individuals and communities that contribute data, the investigators that design research and collect and analyse data, and the wider scientific community that might productively use data for further research.

The joint statement of purpose

Vision

We, as funders of health research, intend to work together to increase the availability to the scientific community of the research data we fund that is collected from populations for the purpose of health research (1), and to promote the efficient use of those data to accelerate improvements in public health.

Principles

Funders agree to promote greater access to and use of data in ways that are:

- Equitable: Any approach to the sharing of data should recognise and balance the needs of researchers who generate and use data, other analysts who might want to reuse those data, and communities and funders who expect health benefits to arise from research

Quelles données déposer ?

EXAMPLES OF PREPUBLICATION DATA-RELEASE GUIDELINES

Project type	Prepublication data release recommended	Prepublication data release optional
Genome sequencing	Whole-genome or mRNA sequence(s) of a reference organism or tissue	Sequences from a few loci for cross-species comparisons in a limited number of samples
Polymorphism discovery	Catalogue of variants from genomic and/or transcriptomic samples in one or more populations	Variants in a gene, a gene family or a genomic region in selected pedigrees or populations
Genetic association studies	Genomewide association analysis of thousands of samples	Genotyping of selected gene candidates
Somatic mutation discovery	Catalogue of somatic mutations in exomes or genomes of tumour and non-tumour samples	Somatic mutations of a specific locus or limited set of genomic regions
Microbiome studies	Whole-genome sequence of microbial communities in different environments	Sequencing of target locus in a limited number of microbiome samples
RNA profiling	Whole-genome expression profiles from a large panel of reference samples	Whole-genome expression profiles of a perturbed biological system(s)
Proteomic studies	Mass spectrometry data sets from large panels of normal and disease tissues	Mass spectrometry data sets from a well-defined and limited set of tissues
Metabolomic studies	Catalogue of metabolites in one or more tissues of an organism	Analyses of metabolites induced in a perturbed biological system(s)
RNAi or chemical library screen	Large-scale screen of a cell line or organism analysed for standard phenotypes	Focused screens used to validate a hypothetical gene network
3D-structure elucidation	Large-scale cataloguing of 3D structures of proteins or compounds	3D structure of a synthetic protein or compound elucidated in the context of a focused project

Toronto International Data Release Workshop Authors (2009),
"Prepublication data sharing", *Nature* 461, 168-170

Quand déposer des données ?

- … : lors de la publication d'un article
  - 1992 (Etats-Unis) : dans les 6 mois
  - 1996 (Human Genome Project) : dans les 24 h
  - ≠ embargo de 18 mois pour les données des brevets !

Les obstacles au partage et à la réutilisation

Le “vol” de données

- “To date, many papers have been published by third parties reporting research findings enabled by data sets released before publication. The experiences shared in Toronto suggest that these have rarely affected subsequent publications authored by the data producers.”
- embargo : cf. *Principes des Bermudes* (Human Genome Project), 1996
- “the Toronto meeting participants recognized that this is an ongoing concern that is best addressed by fostering a scientific culture that encourages transparent and explicit cooperation on the part of data producers, data analysts, reviewers and journal editors”

Toronto International Data Release Workshop Authors (2009), “Prepublication data sharing”, *Nature* 461: 168-170

Les licences fermées

- En 1998, Celera annonce que les résultats de son initiative de séquençage seront "en accès libre"… mais est-ce suffisant ?

- "any restrictions on use should be strongly resisted and we endorse explicit encouragement of open sharing, for example under the newly available CC0 public domain waiver of Creative Commons"

Paul N. Schofield *et al.* (2009), "Post-publication sharing of data and tools", *Nature* 461: 171-173

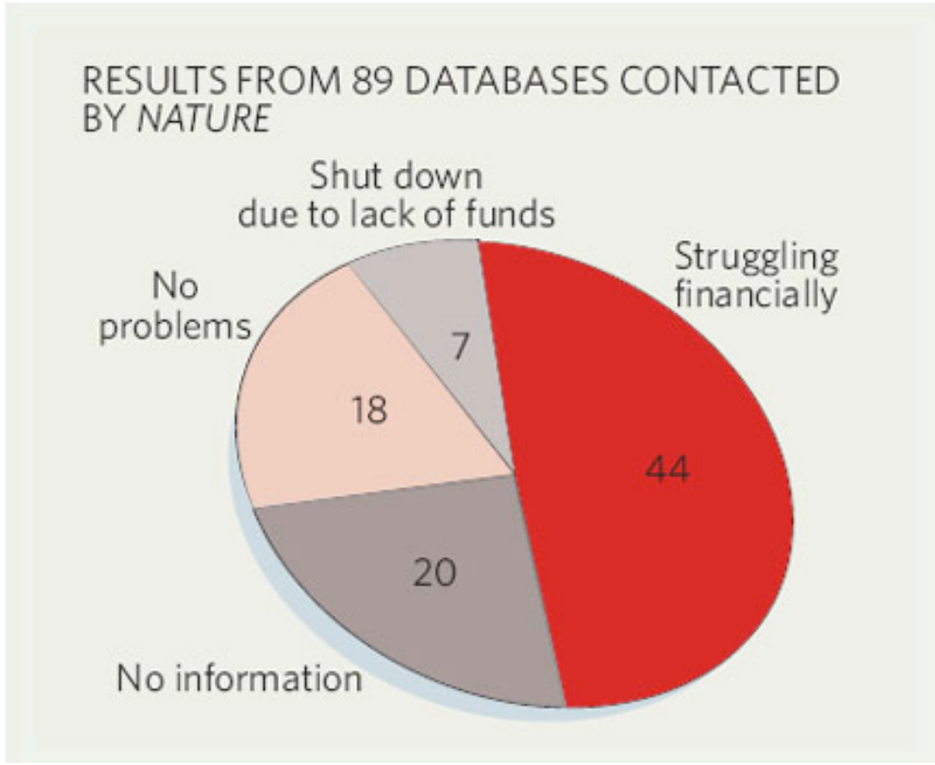
Les bases de données laissées en friche

- 1330 bases de données en biologie moléculaire (Nucleic Acids Research online Database Collection)

- il est plus facile d'obtenir un financement pour créer une nouvelle base de données que pour la maintenir en fonctionnement

- quand la taille des bases de données génomiques double (tous les 18 mois), les coûts de maintenance augmentent de 10%

OECD Megascience Forum Working Group on Biological Informatics (1999), *Final Report*, http://www.gbif.org



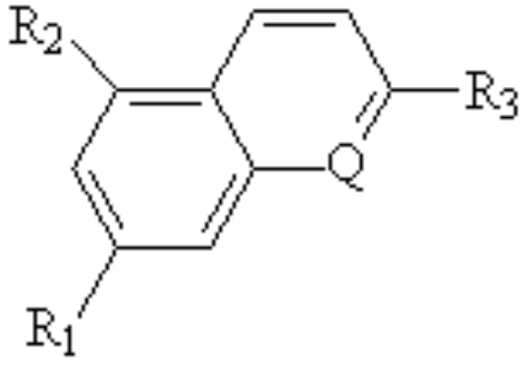
Merali Z. & J. Giles (2005), "Databases in peril", *Nature*, **435**: 1010-1

La difficulté à trouver l'information

- Un problème connu, cf. Altschul, S.F. et al. (1994), "Issues in searching molecular sequences databases", *Nature Genetics* **6**(2): 119-129

- mais accentué dans les bases de données brevets où les inventions sont décrites génériquement

Exemple :"A nucleic acid sequence **at least 90% identical** to the nucleotide sequence of any of the polynucleotides of (a)–(n)…"



Quelques exemples de réutilisation des données

Dans la recherche translationnelle

The New York Times

Health / Research

August 12, 2010

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-

No one would own the data. No one could submit patent applications, though private companies would ultimately profit from any drugs or imaging tests developed as a result of the effort.

on Thursday. "What was needed was to find some way of seeing what was happening in the brain as Alzheimer's progressed and asking if experimental drugs could alter that progression."

Dans l'éducation

Cell Biology Education
Vol. 2, 98–111, Summer 2003

Essays

Public Access for Teaching Genomics, Proteomics, and Bioinformatics

A. Malcolm Campbell

Biology Department, Davidson College, Box 7118, Davidson, North Carolina 28035-7118

Quelques bonnes pratiques

Accompagner la mise en ligne de ses données d'un "marker paper" citable

Ne pas se reposer sur les “supplementary data”

- une pratique courante…
- …alors que ces données ne sont pas standardisées, pas cherchables, et pas pérennes (17% des SD n’étaient plus en ligne 1 an après — surtout celles stockées ailleurs que sur le site de la revue).

N. Anderson *et al.* (2006), “On the persistence of supplementary resources in biomedical publications”, *BMC Bioinformatics* **7**: 260

Avoir une politique de dépôt et d'archivage

- soutenue institutionnellement
- avec des moyens financiers
- et des objectifs et des incitations clairs

© *Science*

Etc.

- Voir les autres interventions de la journée !