Samuel Goëta, doctorant à Télécom ParisTech, tient, dans le cadre de son projet de thèse sur « Les coulisses de l'*open data* : sociologie de la production et de la libération de données publiques », un carnet de recherche en ligne[1]. Nous reprenons ici un extrait d'un billet publié le 11 janvier 2013 à la suite d'une intervention de Bruno J. Strasser, professeur de biologie à l'université de Genève, venu présenter ses travaux sur l'histoire des données dans les sciences dans le cadre du projet Sacred[2].

Collectionner des données ou expérimenter : *une querelle des Anciens et des Modernes ?*

L'essor des sciences du vivant a été accompagné par l'apparition à la Renaissance des cabinets de curiosité où étaient entreposées, classifiées et exposées des espèces hétéroclites. Amasser des plantes et des espèces naturelles était alors un divertissement commun pour la haute société de l'époque. Collectionner et montrer sa collection était un marqueur social d'érudition. Cette tradition de collection relevait principalement d'une science amateur et d'une tradition naturaliste qui aboutit au XIXe siècle aux muséums de sciences naturelles et leurs immenses collections d'espèces. Au début du XXe siècle, cette longue tradition déclina sous l'assaut de la science expérimentaliste qui consacre le laboratoire comme le seul lieu de la recherche scientifique.

Deux méthodes scientifiques et deux traditions épistémologiques divisent au milieu du XXe siècle les sciences du vivant :

- les méthodes comparatives : collectionner, classifier, comparer, corréler ;
- les méthodes expérimentales : observer, analyser, généraliser à partir du cas particulier.

Pour Bruno J. Strasser, la *datadriven science* trouve ses sources dans la tradition comparative dont les pratiques et les métiers sont similaires à celle de cette « nouvelle » manière de faire de la science.

CODIFIER LE GÉNOME : UNE HISTOIRE DE BASE DE DONNÉES ET D'INDIVIDUS

Dans une période où l'expérimentation triomphe comme la seule manière de faire de la « vraie » science, le projet de codifier et de numériser l'ADN dans les années 60 marque le retour à la tradition comparative. Enregistrer une base de données, classifier et comparer des séquences de protéines ne diffère pas des pratiques de collection et de comparaison des espèces dans la science comparative. Pour Strasser, le musée et le serveur sont deux objets standardisés qui servent à produire du savoir.

Le premier projet de constitution d'une base de données massive en génétique, l'*Atlas of Protein Sequence* (1965), dirigé par Margaret Dayhoff, fut un échec du fait de la difficulté à collecter les données venant de chaque laboratoire. Dayhoff ne parvenait pas à convaincre ses collègues de diffuser les données du génome dans sa base de données en raison d'un régime de propriété intellectuelle qui, malgré un système d'accès par modem, ne permet pas la redistribution des données. Les données expérimentales sont alors un objet privé qui appartient à celui qui les a produites. Appliquant des techniques de cristallographie issues de la chimie, une discipline proche de l'industrie qui n'a pas pour habitude de diffuser ses données, le projet *Protein DataBank* lancé en 1969 ne parvient pas non plus à obtenir suffisamment de données et menace de fermer. Ce n'est finalement qu'à la fin des années 70 dans le Nouveau Mexique à l'université de Los Alamos qu'un projet de base de données génétiques parvient à décoller. Il s'agit du projet *GenBank* conduit par Walter Goad, un scientifique au parcours tumultueux qui a participé aux recherches sur la bombe H avant de concevoir ce projet, qui comporte aujourd'hui les séquences de nucléotides de près de 300 000 espèces. Quelles ont été les raisons de son succès ?

LA RECETTE DE L'*OPEN SCIENCE* : ÉCHANGE DE CAPITAUX SYMBOLIQUES ET APPARENCE D'OUVERTURE

Dès son lancement, *GenBank* est présenté comme un projet dans lequel l'usager est aussi contributeur. Dans les années 80, ce projet réussit le tour de force de l'*open access* à une époque où le partage des données des recherches n'a rien d'une évidence. Walter Goad met en place un système vertueux dans lequel il est indispensable de partager des données pour accéder aux publications. Selon Bruno J. Strasser, le succès de *GenBank* vient de son inspiration de la philosophie des économies morales, un système dans lequel les contributions s'équilibrent pour éviter le problème du passager clandestin (*free rider*)[3].

[1] http://coulisses-opendata.com

[2] www.iscc.cnrs.fr/spip.php?article1708

[3] En théorie économique, celui qui profite d'un système sans y contribuer et le mettant ainsi en péril.

Tablettes rennaises/ Bibliothèque de Rennes Métropole

***Cabinet de curiosité de Ferrante Imperato*, 1672.**
Entreposer et classifier : des cabinets de curiosité aux entrepôts de données du XXIe siècle…

L'autre aspect du succès de *GenBank* sur lequel insiste Strasser, c'est l'apparence d'ouverture du système. « *Une force importante de votre projet est son ouverture* », écrit un ami de Goad dans une lettre. Pour obtenir le contrat qui a financé le lancement du projet en 1982, son concepteur ne cesse de donner des signes d'ouverture y compris en insistant sur la connexion du service au réseau Arpanet qui commence à relier les universités américaines.
Pour Strasser, le succès de *GenBank* réside finalement dans le registre symbolique et la communication plutôt que dans la technologie du service.

NOUVELLES PRATIQUES, NOUVEAUX MÉTIERS

Avec la disponibilité de données génétiques de plus en plus importantes, de nouveaux métiers émergent, certains parlent même d'une « nouvelle espèce de scientifiques » (*a new bride of scientists*). Les *computationals scientists* font partie de cette nouvelle manière de faire de la science, ni vraiment expérimentale ni vraiment comparative, qui s'emploie à analyser les données que produisent d'autres. Ils revendiquent rapidement leur statut d'auteur scientifique en proposant des publications aux revues scientifiques, qui voient d'un mauvais œil ces scientifiques qui abandonnent le microscope pour l'ordinateur en réutilisant les données mises à disposition. En 1987, le journal *American Statistics* réduit leur travail à cette expression « *Have computer, give me data* », signe d'un malaise de la communauté scientifique devant ces chercheurs qui publient en leur nom avec les données des autres.

Autre métier déconsidéré : celui de « *database curator* », en charge d'enrichir les métadonnées et de nettoyer les données pour les rendre réutilisables. Strasser raconte le témoignage d'un *database curator* qui se plaignait que personne ne comprenait son travail à un cocktail lors d'une conférence et laissait entendre qu'il n'était pas perçu comme un collègue par ses pairs. On retrouve là une réaction commune devant le travail souvent déconsidéré des « petites mains de la société de l'information », souvent jugées comme des gratte-papier ainsi que l'expliquent Jérôme Denis et David Pontille dans leur article « Travailleurs de l'écrit, matières de l'information »[4].
Aujourd'hui, l'*open access* est la norme pour les publications scientifiques bien que les régimes de licence et les coûts de publication dans les principales revues forment un méli-mélo incompréhensible. Les pratiques de réutilisation de données scientifiques sont désormais courantes dans la recherche ; selon Strasser, un des prochains prix Nobel de médecine pourrait même ne « jamais avoir tenu une pipette de sa vie ».
Enfin, l'*open science* questionne le rôle du chercheur : son monopole remis en cause, le modèle qui émerge rappelle celui des cabinets de curiosité à la Renaissance. En rompant avec l'emprise de la science expérimentale, il est possible d'envisager des formes de science ouvertes à tous. Par exemple, le projet Foldit[5] se présente sous la forme d'un jeu qui permet à chacun de contribuer à l'étude de la structure des protéines en résolvant des puzzles.

Samuel Goëta
Doctorant à Télécom ParisTech
samuel.goeta@telecom-paristech.fr

[4] www.cairn.info/resume.php?ID_ARTICLE=RAC_015_0001

[5] Foldit est un jeu vidéo expérimental développé en collaboration entre le département d'informatique et de biochimie de l'université de Washington : http://fold.it/portal