Paul de Brem

► Le 11 septembre 2001, quatre avions s'écrasaient aux États-Unis. Certains ont considéré que ce drame aurait pu être évité. Les services de renseignements disposaient en effet des informations relatives à cet attentat. Ils n'ont cependant pas été capables de repérer ces données au sein de l'ensemble des informations recueillies. L'existence d'un déluge d'informations a peut-être été à l'origine de cette situation. Cet événement peut-il être comparé au travail des biologistes, qui seraient également submergés par une abondance de données ? <

Le déluge informationnel

La nouvelle production du savoir dans les sciences de la vie

Bruno J. Strasser

Université de Genève, IUFE,
Pavillon Mail, 40, boulevard du
Pont-d'Arve, CH-1211 Geneva,
Suisse.
bruno.strasser@yale.edu

Selon de nombreux commentateurs, un véritable « déluge informationnel » (*data deluge*) s'abattrait aujourd'hui sur le monde, et la condition moderne se caractériserait par une surabondance de données (*information overload*), représentant à la fois une menace et une opportunité (*Figures 1 et 2*). *The Economist*, par exemple, a avancé l'idée selon laquelle notre incapacité à donner sens aux abondantes données économiques serait à l'origine des crises financières et le magazine *Wired* est allé jusqu'à annoncer la fin de la science basée sur des hypothèses, des théories, et des expériences, remplacée aujourd'hui par une science fondée sur l'analyse de données (*data-driven science*). L'émergence de cette nouvelle notion a enthousiasmé l'industrie, en particulier l'industrie informatique et l'industrie du logiciel. Microsoft a ainsi popularisé l'idée de l'avènement d'une nouvelle science correspondant à un « quatrième paradigme » dans la production du savoir scientifique.

C'est dans le domaine de la génomique que les conséquences de ce déluge informationnel ont suscité le plus de discussions. Dès la fin des années 1990, le biologiste moléculaire David Botstein annonçait la naissance d'une nouvelle science dont l'objet était

Figure 1. *The Economist*, 25 février 2010 (© Brett Ryder).

la collection et l'analyse des données¹. Cette science nous permettrait de découvrir dans les données expérimentales des choses que nous ne suspicions même pas. La science n'aurait alors plus pour fondement de tester des théories et des modèles, mais de produire des connaissances à partir de données (*data-driven science*). Néanmoins, des débats existent sur la valeur et l'intérêt de cette nouvelle

¹ Botstein D, Cherry J.M. Molecular linguistics: Extracting information from gene and protein sequences. *Proc Natl Acad Sci USA* 1997 ; 94 : 5506-7.

Figure 2. *Wired*, 16 juillet 2008 (© Wired).

approche. Elle suscite les critiques, voire le scepticisme de certains scientifiques, et des philosophes avancent souvent l'idée qu'il est impossible de produire de nouvelles connaissances sans définir d'hypothèse préalable.

Tradition naturaliste versus tradition expérimentaliste

Les partisans comme les critiques de cette nouvelle vision de la science partagent une idée commune de l'histoire des sciences du vivant. Tous considèrent que celles-ci vivent une période sans précédent. Pour eux, depuis le début de l'époque moderne, l'étude du vivant était dominée par la tradition naturaliste qui a produit les grandes collections que l'on trouve dans les musées d'histoire naturelle, les galeries d'anatomie comparée et de paléontologie (Figures 3 et 4). Cette tradition a décliné dès la fin du XIX^e siècle sous les assauts de la recherche de laboratoire, un mouvement qui va s'accroître au XX^e siècle avec l'essor de la génétique et de la biologie expérimentale plus généralement. La biologie moléculaire illustrerait le triomphe de la biologie expérimentale et le déclin irrémédiable de la biologie naturaliste. Le Prix Nobel Frederick Sanger a résumé cette vision mieux que quiconque : « *“Doing”, for a scientist, implies doing experiments* ».

Ce récit est certes séduisant, mais il confond pratiques et disciplines. L'essor historique des sciences expérimentales n'est pas dû uniquement à la puissance des pratiques expérimentales mises en œuvre au laboratoire. Les pratiques issues de la tradition naturaliste, comme la collection, la comparaison et la classification, ont aussi joué un rôle important.

Dans la tradition des sciences expérimentales, les « organismes modèles » comme la bactérie *E. coli*, la mouche *Drosophila melanogaster* et la souris ont joué un rôle crucial. Jacques Monod et François Jacob disaient : « *Ce qui est vrai pour Escherichia coli doit l'être aussi pour les éléphants* ». Cette phrase illustre le fait que les chercheurs pensaient pouvoir produire un savoir général à partir d'observations sur des organismes particuliers. Cette vision contraste avec celle des naturalistes qui collectionnent de grandes quantités d'organismes différents pour

Figure 3. *Cabinet de Alfred Russel Wallace* (© R. Heggestad).

Figure 4. *Galerie de paléontologie et d'anatomie comparée, Paris* (© Mathyld/under the pyramids).

les comparer. C'est en comparant des données issues d'organismes différents qu'ils produisaient du savoir général. Cette approche, basée sur les notions de collection, de comparaison et de classification, a été au cœur de l'approche naturaliste au cours de ces quatre derniers siècles (Figures 3 et 4). Mais cette approche est-elle propre à l'histoire naturelle ? Ne voit-on pas apparaître l'équivalent des collections des naturalistes avec le développement des grandes banques de données de séquences, de gènes, du transcriptome, ou de structures de protéines ? Celles-ci ne constitueraient-elles pas des « musées électroniques » permettant à nouveau, dans la

tradition naturaliste, la production de savoir basée sur la collection, la comparaison, et la classification de données ? (Figures 5 et 6).

Cette ressemblance entre pratiques naturalistes et pratiques du laboratoire nous aide à comprendre les continuités entre pratiques contemporaines et pratiques plus anciennes. La banque de données de structures de protéines, la *Protein Data Bank*, comprend des données relatives à plus de 1 000 espèces. Il ne s'agit donc pas uniquement de quelques organismes modèles. De même, *GenBank*, la banque de séquences d'acides nucléiques, comprend des données relatives à plus de 300 000 espèces et ce chiffre est en croissance rapide. À titre de comparaison, le plus grand musée d'histoire naturelle américain comprend environ 500 000 espèces. Les banques de données de séquences vont donc rapidement dépasser en biodiversité les plus grands musées et ce retour à des pratiques de comparaisons entre espèces montre que la biodiversité revient au premier plan de l'activité des laboratoires.

De nouveaux acteurs dans la production du savoir

Le développement des bases de données s'est très vite heurté à des défis majeurs. À l'instar des naturalistes Linné et Cuvier, les chercheurs ont voulu collectionner des données auprès d'autres chercheurs et les rassembler dans des collections. Cette pratique, en vigueur depuis des siècles chez les naturalistes, s'est trouvée confrontée à de nombreuses difficultés auprès des expérimentalistes. Ces derniers, contrairement aux naturalistes, n'étaient pas prêts à partager leurs informations et à les placer dans des banques de données publiques. Ils critiquaient ces « magiciens de l'ordinateur » et de la bio-informatique, qu'ils considéraient comme « des parasites » exploitant des données qu'ils avaient produites. L'apparition, dès les années 1980, de ce nouveau type de scientifiques, comparant et analysant les données produites par d'autres, s'est accompagnée de la naissance de tensions avec les producteurs de données, les expérimentalistes. En effet, dans la tradition expérimentale, l'auteur est un auteur individuel qui produit et analyse ses propres données. Dans la tradition naturaliste, les grands naturalistes écrivaient communément des monographies à partir des données d'autres chercheurs. Cette pratique, acceptée parmi les naturalistes, est souvent rejetée par les expérimentalistes. Les cristallographes qui produisent des interprétations de la structure des protéines, par exemple, se sont montrés très réticents à partager leurs données. Il a fallu que les journaux scientifiques s'allient avec les banques de données et rendent obligatoire la soumission des informations dans les banques de données pour que les cristallographes rendent leurs données publiques. D'autres parallèles sont éclairants. Les amateurs ont joué un rôle important dans le développement de l'histoire naturelle, par exemple en botanique, en ornithologie et en paléontologie. Leur contribution au développement des sciences expérimentales, en revanche, a

Figure 5. Collection ornithologique du Smithsonian National Museum of Natural History (© Chip Clark, SI Photo Services).

été extrêmement limitée. Mais aujourd'hui, grâce aux banques de données expérimentales, leur participation est possible. Les sites de génomique personnelle incitent les internautes à envoyer des échantillons pour obtenir l'analyse de leur ADN, les invitant également à partager leurs données médicales. De telles initiatives autorisent un grand nombre d'individus à participer à l'effort scientifique et à la production de

Figure 6. Modèles des protéines de la Protein Data Bank (© Protein Data Bank).

connaissances. Ces démarches sont nouvelles dans la tradition expérimentale et pourraient transformer les modes de production du savoir. Il serait toutefois erroné de considérer que les sciences expérimentales ont simplement adopté les pratiques anciennes de l'histoire naturelle. Aujourd'hui, on assiste plutôt à l'émergence d'un nouveau régime de production du savoir dans les sciences du vivant, un régime hybride entre la tradition de l'histoire naturelle et celle des sciences expérimentales. Une mise en perspective historique permet de comprendre les défis et les problèmes qui se posent aujourd'hui aux sciences du vivant, et se demander dans quelle mesure les solutions trouvées par les naturalistes à leur propre « déluge informationnel », aux problèmes de gestion des collections et de coordination des collectionneurs, pourraient s'appliquer aux sciences expérimentales actuelles. ♦

RÉFÉRENCES

1. Strasser BJ. Data-driven sciences: from wonder cabinets to electronic databases. *Stud Hist Philos Biol Biomed Sci* 2012 ; 43 : 85-7.
2. Strasser BJ. The experimenter's museum: GenBank, natural history, and the moral economies of biomedicine. *Isis* 2011 ; 102 : 60-96.
3. Strasser BJ. Collecting, comparing, and computing sequences: The making of Margaret O. Dayhoff's *Atlas of Protein Sequences and Structure*, 1954-1965. *J Hist Biol* 2010 ; 43 : 623-60.
4. Strasser BJ. Laboratories, museums, and the comparative perspective: Alan A. Boyden's serological taxonomy, 1925-1962. *Hist Stud Nat Sci* 2010 ; 40 : 149-82.

DISCUSSION

Paul de Brem

Au niveau des problématiques, il semble exister une continuité entre le XIX^e siècle et aujourd'hui.

De la salle

Vous avez soulevé l'idée que les données issues des expérimentalistes étaient privées. Or ces informations sont bel et bien publiques. La recherche académique a pour obligation d'inscrire les données dans un cahier de laboratoire auquel tout le monde peut avoir accès.

Bruno J. Strasser

Ces cahiers peuvent en effet être utilisés en cas d'enquête. Dans la pratique, les données ne sont pas réellement publiques et sont conçues dans un cadre strictement privé.

De la salle

Il semble exister une ambiguïté en ce qui concerne le domaine de la recherche publique. L'observation qui était au cœur des sciences naturalistes semble disparaître avec l'avènement des sciences expérimentales. Il serait nécessaire de redonner une place prépondérante à l'observation afin de faire avancer le progrès.

De la salle

Votre exposé était très intéressant. En ce qui concerne le déluge de données, le problème ne relève-t-il pas du fait qu'il n'existe plus de modèles et d'hypothèses qui permettent d'analyser les données ?

Bruno J. Strasser

La notion de gène est en train de se développer à travers toutes sortes de conceptions différentes. Nous espérons qu'un renouveau conceptuel permettra à l'avenir de pouvoir mieux organiser les données.

Pierre Le Ber

Les données qui sont aujourd'hui disponibles amènent à se poser de nombreuses questions, qui, à leur tour, feront évoluer les concepts et permettront à l'avenir d'analyser les mêmes données sous des angles renouvelés. La façon dont les données seront traitées évoluera constamment en fonction des nouvelles problématiques qui apparaîtront.

De la salle

Votre exposé donne l'impression que les chercheurs n'appartiennent à aucun organisme et travaillent de manière indépendante. Quel est alors le rôle des organismes ? De quelle manière le travail et la production des scientifiques sont-ils évalués ?

Bruno J. Strasser

L'effort réalisé au niveau de la coordination institutionnelle a permis aux grandes banques de données de devenir publiques. Le projet de génome humain a joué un rôle majeur en la matière. Il s'agissait de coordonner les centres et de s'assurer que tous les acteurs prennent part au partage. Cette démarche a rendu possible une plus grande ouverture des données.

De la salle

Ce sentiment de déluge de données est-il synonyme d'une certaine désillusion vis-à-vis de la science elle-même et de l'infini de la connaissance ?

Bruno J. Strasser

Pas réellement. Les personnes qui parlent du déluge sont très enthousiastes et considèrent celui-ci comme une opportunité de refonder la science.

De la salle

Certaines revues ont joué un rôle majeur. L'arrivée des revues payantes en ligne a un impact sur le système économique de publication. Quelles étaient à l'époque les motivations des éditeurs de revues ? La situation va-t-elle changer à l'avenir ?

Bruno J. Strasser

Certaines revues majeures ont pris la décision de mettre en place une politique de soumission obligatoire des données. La revue *Nature* a résisté à cette démarche

pendant longtemps, mais a fini par céder. En effet, les responsables de ce magazine craignaient la concurrence d'autres journaux. Par ailleurs, l'essor des journaux en ligne change la donne. L'accès à la publication est différent. Les personnes qui ont fortement préconisé l'accès aux banques de données soutiennent ces journaux ouverts².

De la salle

Pourrait-il être intéressant de traiter la question des données à travers le prisme du dualisme entre le quantitatif et le qualitatif ? S'agit-il d'une approche trop simpliste ?

² NDLR : journaux ouverts : il s'agit des publications en accès libre sur internet (*open access*) par exemple la série des *PLoS*.

Bruno J. Strasser

Cette question a souvent été posée et suppose d'opposer la science quantitative du laboratoire à la science qualitative du musée. Cependant, les naturalistes sont férus de méthodes quantitatives. L'usage des nombres n'est pas réservé aux seuls laboratoires. ♦

TIRÉS À PART

B.J. Strasser