Il se nourrit non seulement des écrits théoriques de ses prédécesseurs, mais aussi de tout ce qui a été produit sur le « style culturel » de la région où il va enquêter. Ces connaissances l'orientent vers tel ou tel objet d'étude. Puis il part sur le terrain et se forge les outils qui lui seront nécessaires pour résoudre les problèmes singuliers qui se posent à lui. Les seuls conseils que son directeur de thèse puisse lui donner c'est de cultiver sa curiosité intellectuelle, rester humble face à autrui et à ce qu'il croit avoir compris, savoir écouter ou plutôt entendre ce que ses interlocuteurs lui disent, et tirer profit de ses erreurs.

PLS

N'est-ce pas difficile de travailler ainsi ?

➜ **EMMANUEL DE VIENNE :** Si, pour un étudiant, cela ressemble à un piège initiatique. Philippe Descola m'avait dit que le savoir ethnographique ne se transmet pas, qu'il est de l'ordre de l'indicible. C'était passionnant..., mais un peu inquiétant ! Heureusement, il avait su organiser autour de son séminaire à l'EHESS un environnement formateur et chaleureux, où l'on pouvait apprendre comment faire de l'anthropologie par l'exemple et la contradiction de sa propre pensée. Ensuite, sur le terrain, on apprend à relire les textes à sa façon. On mesure la somme de travail ardu, quasi obsessionnel, que ces textes ont nécessité, et l'on s'en inspire. Lors de notre première rencontre, Philippe Descola me conseilla aussi de me « laisser porter » par le terrain, ce qui implique, bien souvent, d'abandonner tout ce que l'on avait imaginé avant de partir.
Quand je suis arrivé chez les Trumai, au cœur de la forêt amazonienne, je voulais étudier les bases cognitives de l'animisme, le fait qu'ils reconnaissent aux non-humains des capacités psychologiques humaines. Je comptais élaborer des questionnaires, mais j'ai vite mesuré le décalage entre ces questions imaginées *a priori* et ce qui intéressait mes hôtes. J'arrivais dans une société qui est au bord de la disparition depuis deux siècles, où les thèmes de la maladie, de l'infortune, de la perte de la « culture » traditionnelle sont omniprésents, une société fort différente de l'idée que je m'en étais forgée à travers mes lectures.

➜ **PHILIPPE DESCOLA :** Ce point est important : l'anthropologue ne cesse de réfléchir aux biais d'observation et d'interprétation de ceux qui l'ont précédé. Et ce que le professeur cherche à transmettre c'est précisément une façon de reformuler les problèmes abordés par l'anthropologie à l'aune des découvertes les plus récentes. Les anthropologues de ma génération ont plutôt fait de l'ethnologie générale, globalisante. Dès lors, ceux de la génération d'Emmanuel sont parfois surpris de se trouver confrontés, sur le terrain, à des situations qui ne sont pas conformes à ce qu'ils avaient imaginé. Ils doivent donc faire un pas de côté par rapport à ce que proposaient les descriptions antérieures.
Chaque génération fait ce pas de côté. Quand j'étais étudiant, la littérature ethnographique sur l'Amazonie portait surtout sur les systèmes de parenté très particuliers du Brésil central et sur les systèmes rituels et mythiques du Nord-Ouest amazonien. Mais ce qui m'intéressait en Amazonie c'était plutôt les rapports à l'environnement tels qu'ils transparaissent notamment dans ce que Lévi-Strauss appelait la mythologie implicite, c'est-à-dire ce que les petits rites qui ponctuent la vie quotidienne nous apprennent sur la façon dont les gens voient leurs rapports aux êtres qui les entourent. Des choses en apparence insignifiantes, comme le chant d'un chasseur ou l'interprétation d'un présage onirique, m'ont ainsi permis de mettre en évidence chez les Achuar une « ontologie animiste », c'est-à-dire un ensemble d'inférences quant à la nature des choses, dont j'ai montré ensuite qu'elle s'appliquait à toute l'Amazonie.

PLS

Quel a été votre pas de côté ?

➜ **EMMANUEL DE VIENNE :** Cette approche a donné des résultats impressionnants. Mais, parce qu'elle procède en proposant des généralisations à partir de données disparates, elle a pu omettre certains aspects. La théorie de l'animisme de Philippe Descola, par exemple, m'a laissé face à un double désarroi. Quand je suis arrivé chez les Trumai, j'ai d'abord réalisé que ce que j'avais lu et pris pour une description immédiate de la réalité était le fruit d'une interprétation incroyablement complexe. Je m'attendais à « voir » que les Trumai sont animistes, alors que tout dépend des situations et des personnes. Ensuite, il m'a semblé que cette théorie laissait dans l'ombre la dimension affective de l'animisme. C'est pourquoi j'ai voulu me pencher non pas sur ce qui est transmis, mais sur la façon dont c'est transmis et ressenti. Ce fut mon pas de côté.

PLS

S'intégrer, étudier, déconstruire le savoir consigné, élaborer sa théorie... L'échelle de temps de l'anthropologie est longue !

➜ **PHILIPPE DESCOLA :** Bronislaw Malinowski (1884-1942), le père de la méthode ethnographique, en a bien décrit les étapes. Selon ce que m'a rapporté l'un de ses étudiants à Cambridge, il lui avait évoqué avant son départ pour le terrain, dans les années 1930, ce à quoi il devait s'attendre : durant la première phase qui dure deux à trois mois, l'ethnographe voit tout en noir (le climat est insupportable, la nourriture détestable, les gens désagréables). Au cours de la deuxième phase, il commence à trouver ses repères. Au bout de six mois, le chercheur reconnaît qu'il a obtenu quelques résultats. Au bout d'un an, il écrit à son directeur de thèse qu'il peut rentrer, car il a compris comment fonctionne la société étudiée. Un mois plus tard, il avoue qu'il n'a rien compris et doit rester encore au moins six mois... Eh bien, c'est toujours comme cela aujourd'hui !

➜ **EMMANUEL DE VIENNE :** Rester longtemps est indispensable. Ainsi, j'avais fait un premier séjour d'un an et demi chez les Trumai, puis j'y suis retourné deux ans plus tard. J'ai recueilli en une semaine des informations qui ont représenté la moitié du matériel utilisé pour ma thèse ! Il a fallu le premier séjour long, une absence et le retour pour qu'ils me fassent certaines confidences et me révèlent des savoirs normalement tenus secrets. Le temps de l'anthropologue est long, et l'écriture souvent trop lente pour accompagner les transformations brutales et inquiétantes qui touchent aujourd'hui les sociétés amazoniennes. ■

En partenariat avec le CNRS
14-15-16 novembre 2013, à la Sorbonne, Paris

L'explosion mémorielle change la donne

On peut aujourd'hui facilement stocker d'énormes quantités d'information. Cette capacité nouvelle n'est pas qu'un changement quantitatif : elle annonce des transformations profondes dans la société et ses activités.

Gilles Dowek

Au cours des dernières années, les capacités de production et de stockage de l'information ont énormément augmenté, menant à un « déluge de données » *(voir l'article de S. Abiteboul et P. Senellart dans ce numéro)*. Cette croissance s'accompagne de nombreux défis scientifiques et techniques : il faut notamment améliorer les performances des matériels et des logiciels, garantir la sécurité et la pérennité des données, concevoir des méthodes pour les exploiter, etc. Ces questions scientifiques et techniques sont intimement liées à celle de l'impact de cette augmentation sur les individus et les sociétés.

Le déluge de données est d'abord un changement de nature quantitative. Nous pouvions naguère collecter, stocker, transmettre et transformer de petits volumes d'information ; nous pouvons désormais faire de même avec des volumes d'information plus grands. Toutefois, comme souvent, un changement quantitatif annonce un changement qualitatif. La lunette de Galilée n'était qu'un œil qui permettait de voir des objets plus petits et plus lointains ; mais la découverte de ces objets a révolutionné l'astronomie et, au-delà, la physique. De façon comparable, l'augmentation des capacités des mémoires numériques transforme le champ des problèmes scientifiques, les méthodes de la recherche et la forme même de nos connaissances.

La portée de ces changements ne concerne pas uniquement la science : notre rapport au temps, nos institutions politiques, nos économies, nos systèmes éducatifs, etc. se trouvent bouleversés et confrontés à de nouveaux défis.

Pour mieux comprendre ces transformations, commençons par donner quelques ordres de grandeur. Les quantités d'information s'expriment, comme toute autre grandeur, dans une certaine unité. L'unité fondamentale est le bit, qui correspond à un choix binaire (0 ou 1), mais on utilise souvent une unité dérivée, l'octet, égal à huit bits et qui correspond donc à un choix parmi $2^8 = 256$ possibilités.

Un octet est, *grosso modo*, la quantité d'information contenue dans une lettre de notre alphabet qui, en comptant les

L'ESSENTIEL

- Un disque de quelques centimètres, qui coûte moins de 100 euros, peut contenir aujourd'hui l'équivalent d'une bibliothèque nationale.
- De telles capacités de stockage, par les possibilités qu'elles offrent, bouleversent l'activité scientifique et nécessitent des développements informatiques.
- Elles modifient aussi notre rapport au passé et ouvrent des perspectives dans de nombreux domaines tels que la politique, l'économie ou l'éducation.

© Roman Gorielov ; Zastolskiy Victor/shutterstock.com

© Pour la Science - n° 433 - Novembre 2013

1. DE PETITS DISQUES DE SAUVEGARDE ayant une capacité de un téraoctet sont désormais courants et sont vendus pour quelques dizaines d'euros. Or un téraoctet correspond à peu près à l'information contenue dans une bibliothèque d'un million de livres de 1 000 pages chacun !

majuscules, les minuscules, les lettres accentuées, les chiffres, les ligatures, les signes de ponctuation, etc., en contient un peu plus d'une centaine.

Un kilooctet, soit 1 000 octets, est à peu près la quantité d'information contenue dans une page de livre. Un mégaoctet, soit 1 000 kilooctets, correspond à l'information contenue dans un livre d'environ 1 000 pages. Un gigaoctet, égal à 1 000 mégaoctets, est *grosso modo* la quantité d'information contenue dans une bibliothèque personnelle de 1 000 volumes. Un téraoctet, soit 1 000 gigaoctets, est approximativement la quantité d'information contenue dans la collection de livres et imprimés de la Bibliothèque nationale de France, qui contient quelques millions de volumes. Le multiple suivant est le pétaoctet, qui vaut 1 000 téraoctets. Certains suggèrent qu'un pétaoctet correspond à la quantité d'information contenue dans l'ensemble des cultures humaines, même si cette quantité est difficile à évaluer.

Bien entendu, dans cette échelle – lettre (octet), page (kilooctet), livre (mégaoctet), bibliothèque personnelle (gigaoctet), bibliothèque nationale (téraoctet), ensemble des cultures (pétaoctet) –, ce sont les ordres de grandeur et non les détails qui comptent. Par exemple, les images et les vidéos sont plus denses en information que les textes – une heure de vidéo contient ainsi une dizaine de gigaoctets, alors qu'une heure de lecture (quelques dizaines de pages) contient quelques dizaines de kilooctets. Et la collection de livres et imprimés de la Bibliothèque nationale de France comporte en réalité 14 millions de volumes, et non un million.

Cependant, cette échelle aide à comprendre les transformations que nous sommes en train de vivre, depuis qu'un disque permettant de stocker un téraoctet ne coûte que quelques dizaines d'euros. Ces changements concernent toutes nos activités qui consistent à collecter, stocker, transmettre et transformer de l'information… c'est-à-dire une grande partie de nos activités !

Commençons par les sciences, qui ont toujours comporté une part de collecte d'information. Par exemple, l'observation des satellites de Jupiter par Galilée ou les observations recueillies par Darwin lors de son voyage à bord du *Beagle* ont joué un rôle essentiel dans la construction de leurs théories. Mais la quantité d'information collectée par les scientifiques d'aujourd'hui est d'un tout autre ordre de grandeur.

Un impact sur les sciences

Le résultat du séquençage d'un génome humain, qui comporte 3,4 milliards de paires de bases, équivaut ainsi à un texte dont la taille est de l'ordre d'un gigaoctet, soit la taille d'une bibliothèque de 1 000 volumes *(voir l'article de G. Perrière dans ce numéro)*. Quant au LHC, le grand collisionneur de particules du CERN, il produit *via* ses détecteurs de l'ordre de 15 pétaoctets de données chaque année, soit plusieurs milliers de fois la taille de la collection de livres et imprimés de la Bibliothèque nationale de France *(voir l'article de F. Malek dans ce numéro)*. Et les astronomes collectent des volumes encore

plus astronomiques de données *(voir l'article de C. Reylé dans ce numéro)*.

Parce qu'ils demandent de brasser de telles quantités d'information, des projets tels que le séquençage du génome ou la recherche du boson de Higgs n'étaient pas à la portée des scientifiques avant la fin du XX^{e} siècle. Ce sont donc des domaines de recherche entiers qui s'ouvrent, depuis peu, devant nous.

Si cette transformation touche la nature des problèmes que les scientifiques peuvent se poser et tenter de résoudre, elle a également un impact sur la démarche scientifique et la forme même de nos connaissances.

La démarche d'abord : naguère, les scientifiques notaient leurs observations dans des « cahiers de manip » *(voir la figure 2)* et s'inspiraient de ces données pour imaginer des théories – c'est ce que l'on appelle l'induction – ou pour confronter les théories imaginées aux observations, afin de réfuter celles qui n'étaient pas en adéquation avec ces dernières. L'importance relative de l'induction et de la réfutation est, aujourd'hui encore, une question débattue. Quoi qu'il en soit, avec des « cahiers de manip » dont la taille se mesure en gigaoctets ou en téraoctets, il est devenu impossible tant de confronter « à la main » les théories aux observations que de s'inspirer de ces observations pour imaginer des théories.

La confrontation des théories aux observations demande un traitement informatique aussi bien des données observées que de la théorie elle-même, qui doit être exprimée sous une forme prédictive, c'est-à-dire algorithmique, pour qu'elle puisse être automatiquement confrontée aux données observées. Et l'utilisation des observations pour imaginer des théories demande un traitement informatique plus difficile encore, appelé « apprentissage automatique ».

Une question typique en apprentissage automatique est de savoir si les lois de Kepler (qui décrivent les trajectoires des planètes autour du Soleil) peuvent être découvertes par un algorithme à partir des observations de Tycho Brahe, que Kepler a utilisées pour concevoir ses lois. Une autre est de savoir s'il est possible, à partir de grands corpus de textes, d'inférer automatiquement des règles de grammaire ou des règles de traduction d'une expression d'une langue en une autre.

En extrapolant – peut-être dangereusement – les évolutions actuelles de la méthode scientifique, nous allons peut-être vers un temps où les observations seront tellement volumineuses que la construction de théories à partir des observations sera une tâche laissée exclusivement aux ordinateurs, le rôle des scientifiques devenant alors de concevoir des algorithmes permettant de construire des théories à partir des observations.

Une transformation de la connaissance ?

Cette transformation progressive de la démarche scientifique en induit une autre sur la connaissance scientifique elle-même. Jusqu'à la fin du XX^{e} siècle, les théories scientifiques étaient exprimées par des propositions de petite taille. Par exemple, la deuxième loi de Newton (selon laquelle la force s'exerçant sur un objet est égale au produit de sa masse par son accélération) s'exprime par une proposition d'une dizaine de caractères : $\vec{F} = m\, d^2\vec{x} / dt^2$. Mais rien ne garantit que les théories automatiquement inférées à partir de gigantesques volumes d'observations soient petites. De fait, nous commençons à prendre conscience qu'il n'y a aucune raison *a priori* pour que les lois de la nature ou les règles de la grammaire d'une langue s'expriment par de courts énoncés.

■ **L'AUTEUR**

Gilles DOWEK est chercheur en informatique à l'Inria (anciennement nommé Institut national de recherche en informatique et automatique), au sein de l'équipe *Deducteam*.

L'échelle des octets

Octet : unité égale à huit bits, pouvant coder $2^8 = 256$ caractères différents.

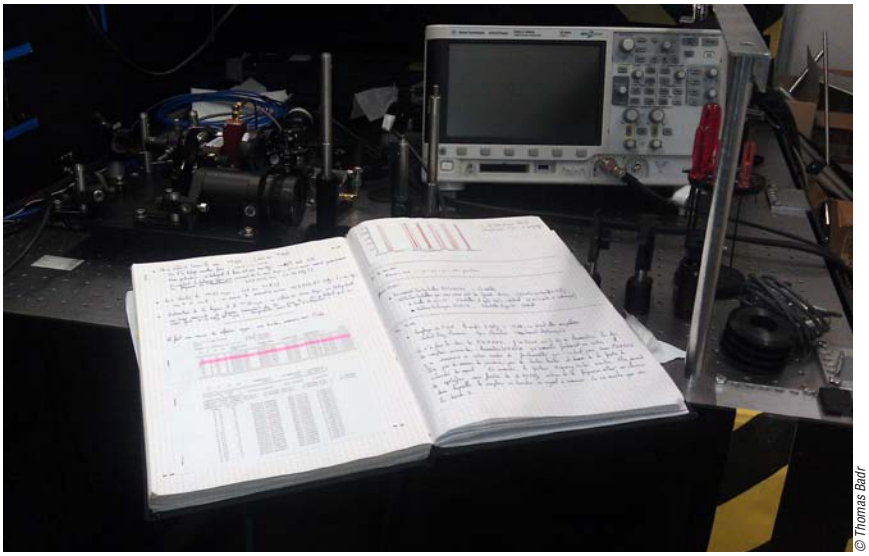
Kilooctet (ko) : 1 000 octets (environ une page d'un livre).

Mégaoctet (Mo) : Un million d'octets (un livre de 1 000 pages environ).

Gigaoctet (Go) : Un milliard d'octets (une bibliothèque de un millier de livres).

Téraoctet (To) : 10^{12} octets (une grande bibliothèque nationale).

Pétaoctet (Po) : 10^{15} octets (l'ensemble des cultures humaines ?).



© Thomas Badr

2. UN CAHIER DE LABORATOIRE, où l'on consigne les résultats des expériences. De plus en plus, ces cahiers prennent une forme électronique. Certains d'entre eux contiennent des téraoctets de données, voire plus ; leur exploitation ne peut se faire qu'avec des moyens informatiques.

© Pour la Science - n° 433 - Novembre 2013

La grammaire d'une langue, par exemple, contient sans doute des milliers de règles dont certaines ne portent que sur des expressions très particulières. En français, on peut dire « je viendrai lundi 2 septembre » ou « je viendrai le lundi 2 septembre » ; on peut aussi dire « je viendrai demain », mais pas « je viendrai le demain ». Il semble ainsi y avoir des règles de grammaire locales propres aux dates, qui diffèrent des règles de grammaire relatives aux adverbes de temps. Et si nous mettons bout à bout toutes ces règles de grammaire locales propres aux dates, aux noms de pays et autres toponymes (ainsi, on va « en Russie », mais « au Brésil » et « à Madagascar », et on prend le bateau pour aller « à Noirmoutier », mais « à l'Île de Ré » et non « à Ré »), on obtiendra probablement des connaissances qui ne peuvent pas s'exprimer, comme la deuxième loi de Newton ou les équations de Maxwell régissant les champs électromagnétiques, en quelques dizaines de caractères *(voir l'encadré ci-contre)*.

Le rôle des scientifiques est appelé à changer

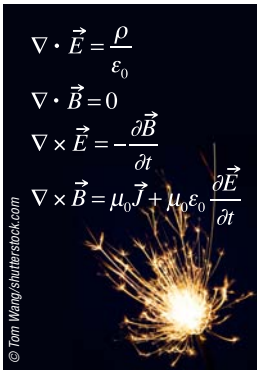
En fait, cette idée est légèrement antérieure au déluge de données. Elle est sans doute apparue quand nous avons pris conscience que les lois qui régissent l'évolution de l'atmosphère ou le fonctionnement d'une cellule, par exemple, ne pouvaient pas s'exprimer en quelques lignes.

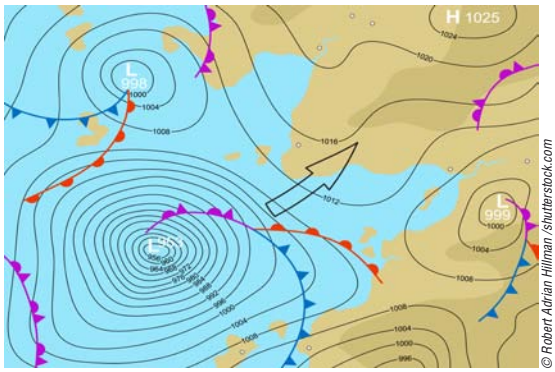
Lorsqu'une théorie s'exprime par des énoncés ou des algorithmes qui font des centaines de pages, le lien entre prévision et explication se distend. Ainsi, nous avons pris l'habitude d'accepter que les prévisions météorologiques indiquent le temps du lendemain avec une probabilité de succès assez élevée, mais qu'elles ne nous expliquent pas pour quelles raisons ce lendemain sera ensoleillé ou pluvieux. Les lois de Newton, au contraire, nous permettent non seulement de prédire que les trajectoires des planètes sont des ellipses, mais aussi de comprendre pourquoi.

Si nous ne pouvons pas expliquer le temps du lendemain, c'est précisément parce que la situation météorologique n'est pas le résultat d'une cause unique, comme le mouvement d'une planète est le résultat de l'attraction d'une étoile, mais de milliers de causes particulières desquelles il est difficile de faire émerger une explication globale.

DES THÉORIES SCIENTIFIQUES QUI S'ALLONGENT ?

Les phénomènes électriques et magnétiques, telles des étincelles *(ci-dessous à gauche)*, sont bien décrits à partir de la théorie électromagnétique élaborée par James Clerk Maxwell vers 1870. Cette théorie s'exprime sous forme compacte par quatre équations vectorielles (les « équations de Maxwell ») mettant en jeu le champ électrique $\vec{E}$ et le champ magnétique $\vec{B}$. En revanche, les prévisions météorologiques *(ci-dessous à droite)* résultent de modélisations théoriques et de calculs qu'il est impossible d'exprimer sous forme courte. Avec les données massives, les situations de ce type seront probablement de plus en plus fréquentes.





Comme les données, ces théories sont trop complexes pour être utilisées « à la main » : seul un ordinateur peut les exploiter pour faire des prédictions. En extrapolant – à nouveau, peut-être dangereusement – les évolutions de la forme des théories scientifiques, nous pouvons imaginer que l'élaboration des prédictions sera un jour entièrement déléguée aux ordinateurs, et que les scientifiques auront alors pour rôle de concevoir des algorithmes permettant aux ordinateurs de calculer des prédictions à partir d'une théorie.

Il est cependant essentiel de garder en tête que, comme les autres théories scientifiques, ces théories restent des hypothèses, toujours à la merci d'une réfutation par les observations, et que leur pertinence ne tient qu'à leur capacité à résister à ces réfutations potentielles. Le risque est grand de prendre pour des oracles des prédictions produites par des ordinateurs à partir de théories trop complexes pour être comprises. Par exemple, contrairement à ce que l'on entend parfois, ce n'est pas le fait que tous les modèles climatiques s'accordent dans leurs prédictions qui importe, mais le fait que les prédictions passées aient été en accord, ou non, avec les observations.

Comme souvent, ces transformations de la technique et de la science en annoncent d'autres, dans divers autres domaines de notre culture. Un premier changement, conséquence de cette capacité nouvelle de collecter, stocker, transformer et transmettre de grandes quantités d'information, est celui de notre rapport à la mémoire et au passé.

La plupart des quelque 100 milliards de femmes et d'hommes qui ont vécu avant nous n'ont laissé aucune trace de leur passage sur Terre. Par exemple, malgré l'écriture et l'imprimerie, peu d'entre nous peuvent citer le nom et le prénom de leurs 16 arrière-arrière-grands-parents. Et les plus anciennes personnes dont nous connaissons le nom, sans doute quelques souverains de Mésopotamie ou d'Égypte, remontent à un passé relativement récent. De l'histoire de l'humanité durant les centaines de milliers d'années qui les ont précédés, depuis l'apparition des premiers *Homo sapiens*, nous n'avons presque aucune trace.

L'humanité devient hypermnésique

Les femmes et les hommes des siècles prochains auront une vision bien différente de leur passé, car les capacités actuelles de stockage de l'information rendent l'humanité hypermnésique. Ainsi, nos descendants connaîtront mieux que nous leur arbre généalogique ; et au lieu d'un nom laissé, pour toute trace, dans un registre d'état civil, ils pourront conserver sur chacun de leurs lointains ancêtres un grand nombre d'informations.

Et les futurs historiens rendent jaloux ceux d'aujourd'hui, car ils disposeront de documents bien plus nombreux et complets...

Cette médaille a son revers : il n'est pas si facile de vivre en sachant que chacun de ses gestes laisse potentiellement une trace à la postérité – c'est vivre avec un *Big Brother* hypermnésique. Aussi est-il probable que se développera un véritable droit à l'oubli, afin de permettre à chacun de décider des traces qu'il souhaite effacer ou transmettre. Ces préoccupations sont à l'origine des recherches des informaticiens qui développent des techniques pour protéger la vie privée, par exemple celle des personnes enregistrées dans des bases de données *(voir l'article de T. Allard, B. Nguyen et Ph. Pucheral dans ce numéro)*.

Vers de nouvelles institutions politiques ?

D'autres éléments de la culture humaine transformés par cette capacité nouvelle à collecter, stocker, transmettre et transformer de grandes quantités d'information sont les institutions politiques. Notre démocratie représentative a été inventée aux XVII^{e} et XVIII^{e} siècles, à une époque ou le débit de l'information que chaque citoyen pouvait communiquer à la collectivité – et, de façon symétrique, la collectivité à chaque citoyen – était très faible.

De fait, en participant tous les cinq ans à l'élection d'un chef d'État parmi 16 candidats, chaque citoyen communique à la collectivité une information de quatre bits. Le débit d'information est donc de 0,8 bit par an – soit 2×10^{-8} bit par seconde. On peut ajouter l'information communiquée lors des élections législatives, locales, etc., l'ordre de grandeur reste le même : quelques bits par an.

Aussi, jusqu'au XX^{e} siècle, de nombreuses stratégies – tracts, pétitions, manifestations, etc. – ont été inventées pour accroître cette quantité. Ces formes de militantisme ont vieilli et sont aujourd'hui remplacées par d'autres : sites Web, pétitions en ligne, etc. La faible quantité d'information échangée a également pour conséquence que, lors d'une unique élection, nous décidons à la fois du devenir de l'armée, de la police, de la justice, de la diplomatie, de l'éducation, de la recherche, de la santé, des arts, du logement, de l'économie, de l'agriculture, du travail, etc. Un chef d'État, comme un roi de jadis, est en effet en charge de toutes ces questions.

© JM-Design ; Dimedro68/shutterstock.com

3. LORSQU'UN CITOYEN VOTE, il communique à la collectivité, *via* son bulletin, quelques bits d'information à peine. La technologie actuelle permet pourtant de transmettre beaucoup plus d'information par des sites Web, des blogs, etc. À quand des institutions politiques adaptées, qui profiteraient de ces capacités techniques ?

■ **BIBLIOGRAPHIE**

Dossier « Universités en ligne », *Pour la Science* n° 431, pp. 33-49, septembre 2013.

M. Serres, **Petite Poucette**, Le Pommier, 2012.

E. Hoog, **Mémoire année zéro**, Seuil, 2009.

L'augmentation du débit de l'information que chaque citoyen peut communiquer à la collectivité et que la collectivité peut communiquer à chaque citoyen permet d'imaginer d'autres formes de démocraties, par exemple des systèmes électifs où plusieurs représentants, élus indépendamment les uns des autres, gèrent chacune de ces questions. La séparation des pouvoirs, jusqu'ici limitée aux trois pouvoirs législatif, exécutif et judiciaire, sera peut-être bientôt rendue possible à une toute autre échelle. Cette augmentation de la quantité d'information que chaque citoyen peut échanger avec la collectivité est sans doute l'une des causes du vieillissement rapide de nos institutions, qui ne sont plus en phase avec les techniques.

De l'information pour nourrir l'économie

Une autre transformation est celle de nos économies, dont l'information est, selon une métaphore déjà un peu usée, le nouveau pétrole. Par exemple, chacun de nos achats en ligne, ainsi que dans un magasin dont nous détenons une carte de fidélité, permet aux commerçants de construire notre profil de consommateur et de nous proposer des produits adaptés à nos besoins et nos désirs. Mais ces constructions de profils individuels ne sont qu'une partie de l'énorme iceberg que représentent les données en économie.

Il est ainsi devenu possible de suivre en temps réel la progression d'une épidémie de grippe simplement en comptant le nombre de requêtes dans un moteur de recherche qui portent sur un mot lié à la grippe – alors que les signalements par les médecins aux autorités de santé ne permettent de suivre la progression d'une épidémie qu'avec plusieurs jours de retard. Ce type d'information est beaucoup plus utile à l'industrie pharmaceutique, qui essaie d'anticiper nos besoins en médicaments, que de construire le profil de tel ou tel consommateur.

De même, les achats de pétrole ou de gaz sont conditionnés à la fois par les prévisions météorologiques à moyen terme et par l'analyse du comportement collectif des consommateurs en réaction à une vague de froid ou à un hiver clément. Sans de telles analyses, il deviendra rapi-

© Pour la Science - n° 433 - Novembre 2013

dement impossible pour une entreprise de se maintenir sur le marché du gaz ou des médicaments. Et cela explique, parmi d'autres raisons, que certains pays, telle la Chine, aient décidé de développer leurs propres moteurs de recherche et de limiter l'accès aux moteurs de recherche étrangers, pour ne pas laisser fuir cette information, comme nous le faisons en Europe.

Des formations supérieures pour tous ?

Un autre domaine qui subit l'influence du déluge de données est celui de l'éducation. Depuis quelque temps déjà, les enseignants savent que toutes les connaissances qu'ils transmettent à leurs élèves sont déjà disponibles sur le Web. Le rôle des enseignants n'est plus d'être une source d'information, la bouche d'où sourd la connaissance, mais de donner un sens à cette information, d'aider leurs élèves à se l'approprier et de leur en montrer la cohérence et l'utilité.

Le déluge de données bouleverse le secteur de l'éducation plus radicalement encore, en transformant l'enseignement à distance. Nous prenons conscience rétrospectivement du caractère élitiste des universités du XX^{e} siècle, qui s'adressaient uniquement à ceux qui pouvaient être présents quand la connaissance sortait de la bouche de l'enseignant. Pour les autres, qui travaillent en plus d'étudier, ou qui vivent dans un village ou un bidonville loin des centres universitaires, rien n'était possible. L'augmentation des capacités de stockage de l'information, en permettant l'enregistrement des cours en vidéo et leur accès de manière souple et depuis le monde entier, contribuera sans doute à réaliser ce vieux rêve d'une éducation pour tous.

Nous comprenons plus ou moins l'impact de ces possibilités nouvelles de collecter, stocker, transformer et transmettre de l'information à grande échelle sur tel ou tel élément de notre culture – la science, la politique, l'économie, l'enseignement, etc. Mais nous avons encore du mal à penser globalement ces transformations. C'est le défi qui s'offre aujourd'hui à nous. ■

LE DÉFI EST DE PENSER GLOBALEMENT
les transformations induites par ces possibilités nouvelles de collecter, stocker, transformer et transmettre de l'information.

Un déluge de données

Serge Abiteboul et Pierre Senellart

Dans de nombreux domaines, scientifiques ou non, les données s'accumulent en masse. Les gérer et les exploiter est le défi posé à l'informatique du *Big Data*.

Au CERN, près de Genève, le collisionneur LHC *(Large Hadron Collider)* est équipé d'énormes détecteurs capables d'enregistrer les traces de dizaines à centaines de millions de collisions proton-proton par seconde. Évaluons grossièrement le volume de données que cela représente, en faisant des hypothèses basses. L'information relative aux produits de chaque collision est représentée à l'aide de quatre octets (soit $(2^8)^4 = 256^4 = 4{,}3$ milliards de possibilités), l'accélérateur fonctionne dix heures par jour en moyenne, et 100 millions de collisions sont enregistrées chaque seconde. On calcule facilement qu'au bout d'un an, le LHC produit ainsi 5×10^{15} octets, soit cinq pétaoctets d'information (une évaluation plus précise fournit des valeurs supérieures, *voir l'article de F. Malek dans ce numéro*). Il faut 5 000 disques durs du commerce pour stocker une telle masse de données.

Cependant, ce n'est pas le stockage qui est la difficulté principale, mais l'exploitation : extraire de l'information utile – de la connaissance – de cette masse de données dépasse largement les capacités humaines. Pour relever de tels défis, une nouvelle science est apparue, la science des données. Il s'agit d'extraire des contenus intéressants à partir de masses de données gigantesques, changeantes, hétérogènes, incertaines, et ce à l'aide d'algorithmes efficaces. L'objet de cet article est de décrire les principaux enjeux de cette science des données et les difficultés qu'elle doit surmonter.

L'exemple du LHC par lequel nous avons commencé est extrême en termes de volumes d'octets, mais de nombreux autres domaines scientifiques et industriels doivent faire face à une avalanche similaire de données. En recherche médicale, par exemple, l'un des problèmes a longtemps été l'absence de données en nombre suffisant ; aujourd'hui, c'est plutôt l'inverse : les chercheurs sont confrontés à la difficulté de produire des connaissances utiles à partir des myriades de données accumulées par les différents essais cliniques, les services hospitaliers, etc. Citons la base de données MIMIC-II *(Multiparameter Intelligent Monitoring in Intensive Care)*, constituée aux États-Unis par une équipe de chercheurs du

© Pour la Science - n° 433 - Novembre 2013

ANALYSE DES DONNÉES : LES TÂCHES CLASSIQUES

– Rechercher des données et les acquérir.

– Homogénéiser les données provenant de plusieurs sources.

– Détecter et éliminer les doublons et les erreurs.

– Interagir avec des humains pour obtenir plus de données, résoudre les contradictions et combler les manques (« crowdsourcing »).

– Faciliter le travail des analystes en leur fournissant des outils de visualisation

– Réaliser des analyses statistiques automatiques des données.

– Développer des applications et de nouveaux services.

© D'après Matvienko Vladimir/shutterstock.com

MIT, de *Philips* et du Centre médical *Beth Israel Deaconess*. Elle regroupe des données médicales portant sur 32 000 personnes hospitalisées en unités de soins intensifs, soit plus de 40 000 séjours . Le volume de données est nettement plus modeste que celui produit par le LHC, mais son analyse n'en nécessite pas moins des compétences spécifiques et des algorithmes complexes.

Les exemples de domaines scientifiques où l'on doit recueillir et gérer des données massives ne manquent pas. On peut citer la génomique *(voir l'article de G. Perrière)*, les relevés astronomiques *(voir l'article de C. Reylé)*, les recensements de la flore et de la faune, la recherche pharmacologique, les études démographiques, etc.

On le comprend, des algorithmes d'analyse de données massives sont indispensables à la recherche scientifique d'aujourd'hui. Mais il en faut également pour des applications plus quotidiennes. En voici un exemple.

Vous voulez choisir un film pour votre soirée ; vous vous connectez à Internet et allez sur un site détaillant les possibilités. Son système de recommandation vous fera des propositions. Il peut procéder de manière très simpliste, par exemple en vous suggérant le film récent le plus regardé au cours de la semaine, ou celui ayant reçu les meilleures critiques. Mais la recommandation peut aussi résulter d'un algorithme complexe qui prend en compte les films que vous avez aimés, votre humeur, peut-être les goûts de celui ou celle qui partagera votre soirée. Pour vous aider à répondre à la simple question « Quel film pourrais-je regarder ? », l'algorithme tiendra peut-être compte d'un océan de données : des avis de centaines de millions de personnes sur des milliers de films.

Or les programmes qui permettent de réaliser des analyses statistiques, même très simples, sur de telles quantités de données sont d'une grande complexité. Dans cette masse, le logiciel de recommandation de films trouvera des personnes qui ont des goûts voisins des vôtres, révélera des proximités avec des internautes que vous ne connaissez pas. Il pourra découvrir vos goûts cinématographiques et vous proposera des films en moins d'une seconde !

Dans la jungle des chaînes de télévision de plus en plus nombreuses, des VoD (*Video on Demand*, vidéo à la demande) et SVoD (*Subscription Video on Demand*, vidéo à la demande avec abonnement), des innombrables films disponibles légalement ou pas sur la Toile, l'usager est perdu ; le but des systèmes de recommandation est de l'aider à s'y retrouver.

L'ingrédient principal qui alimente de tels systèmes est bien sûr constitué par les données numériques, lesquelles tiennent une place de plus en plus importante dans le monde d'aujourd'hui. Depuis les années 1960, les logiciels de bases de données se sont imposés pour partager des données au sein d'une entreprise ou d'une organisation. D'abord isolées dans des centres de calcul, ces données sont devenues accessibles partout dans le monde avec l'arrivée d'Internet, le réseau des réseaux de machines, puis du Web, le réseau des contenus, et finalement du Web 2.0 avec la participation de chacun, les réseaux d'individus.

Et il n'y a pas que les données stockées, il y a aussi les données échangées. Nous sommes entourés de milliards d'objets communicants. En 2008, le Web comptait déjà plus de 1 000 milliards de pages et, chaque mois, les internautes y réalisaient des dizaines de milliards de recherches. On estime que le monde numérique double en volume tous les 18 mois et le trafic sur Internet est déjà, chaque année, supérieur à tout ce que l'on pourrait stocker en utilisant tous les disques et autres supports disponibles.

Analyser les données pour les valoriser

L'ensemble de ces données disponibles sur le réseau mondial constitue un énorme gisement de connaissances à découvrir et à valoriser. L'analyse de données a été un domaine très actif presque depuis les débuts de l'informatique et a revêtu divers noms, tels que *fouille de données* ou *business intelligence*. En raison de l'accroissement des capacités des disques et des mémoires, ainsi que des puissances de calcul avec des grappes d'ordinateurs pouvant réunir jusqu'à des milliers de machines, en raison aussi de l'explosion du volume de données disponibles, l'analyse de données pour en extraire de la valeur est devenue une industrie florissante. Et c'est sous le nom de *Big Data* (données massives) qu'elle se développe aujourd'hui.

Le point de départ est de valoriser les gisements massifs de données. Le *Big Data* inclut en général deux aspects. D'une part, il sous-entend l'idée de croiser des données très structurées, par exemple celles d'une entreprise, avec des masses de données moins structurées et plus défectueuses disponibles sur le Web. D'autre part, il nécessite la mise en œuvre de calculs massivement parallèles en utilisant des techniques logicielles telles que *Hadoop*, issues des moteurs de recherche du Web *(voir l'encadré page ci-contre)*.

Des difficultés multiples

L'objectif étant de faire émerger de nouvelles connaissances à partir des données, les tâches sont celles, classiques, de l'analyse de données, qui vont de leur acquisition à leur exploitation *(voir l'encadré page 33)*. Les difficultés sont nombreuses et tiennent en premier lieu des quatre « V » du *Big Data* : le volume des données (il se compte typiquement en téraoctets ou pétaoctets, soit 10^{12} ou 10^{15} octets), leur variété ou hétérogénéité (sur le plan de la structure, de la langue, du format, etc.), leur vélocité (le rythme auquel elles sont modifiées), leur véracité (erreurs, incomplétude, confiance, provenance, fraîcheur, etc.).

D'autres difficultés proviennent de la répartition des données dans l'espace, des protections éventuelles (droit d'accès, restriction sur leur usage), etc. Par ailleurs, la nature des traitements auxquels ces données sont soumises a une importance considérable. Par exemple, un algorithme dont le temps d'exécution est proportionnel au cube du nombre de données reste inutilisable sur une base contenant un milliard d'enregistrements, même avec des milliers de machines qui fonctionneraient pendant des centaines d'années. Quant aux logiciels tels que *Hadoop*, développés spécifiquement pour traiter des données massives, ils sont encore relativement jeunes et compliqués à utiliser, en particulier parce qu'ils font travailler ensemble un grand nombre de machines.

L'analyse de grands volumes de données intervient dans d'innombrables domaines, afin de faire des prédictions de plus en plus fines, d'anticiper des épidémies, de mieux comprendre l'évolution du climat, d'aider à soigner les cancers, etc. Au niveau du

LES AUTEURS

Serge ABITEBOUL est directeur de recherche à l'INRIA-Saclay et à l'École normale supérieure de Cachan. Il est membre du Conseil national du numérique.

Pierre SENELLART est maître de conférences en informatique à *Télécom ParisTech*.

BIBLIOGRAPHIE

S. Abiteboul, **Sciences des données : de la logique du premier ordre à la Toile**, Leçon inaugurale au Collège de France, Fayard, 2012 : www.college-de-france.fr/site/serge-abiteboul/

S. Abiteboul *et al.*, **Web Data Management**, Cambridge University Press, 2011 : http://webdam.inria.fr/Jorge

D. Agrawal *et al.*, **Big data and cloud computing : current state and future opportunities**, EDBT/ICDT 2011, www.edbt.org/Proceedings/2011-Uppsala/papers/edbt/a50-agrawal.pdf

C. Lynch, **Big data : How do your data grow ?**, *Nature*, vol. 455, pp. 28-29, 2008.

G. Linden *et al.*, **Amazon.com recommendations : Item-to-item collaborative filtering**, *IEEE Internet Computing*, vol. 7(1), pp. 76-80, 2003.

© Pour la Science - n° 433 - Novembre 2013

grand public et pour l'instant, ce qui est le plus apparent du *Big Data*, c'est l'utilisation de données personnelles par de grandes entreprises commerciales, principalement pour des publicités ciblées. C'est le cas de *Google*, qui analyse les requêtes et les courriels des internautes pour mieux cibler sa publicité, ou d'*Amazon*, qui suggère aux usagers des livres à acheter.

Des armes nouvelles aux mains des dictatures... et des citoyens

Récemment, en particulier dans le cadre de l'affaire Edward Snowden aux États-Unis, la presse a souligné que divers gouvernements utilisent l'analyse de données privées à des niveaux surprenants. La principale raison invoquée pour ce type d'utilisations est la lutte contre le terrorisme. Mais le système PRISM de la NSA (*National Security Agency*, l'agence américaine de sécurité) et les systèmes équivalents des autres États sont aussi utilisés pour l'espionnage, notamment industriel. Ils peuvent l'être pour surveiller les opposants politiques. Le contrôle et l'analyse des données sont certainement les nouvelles armes parmi les plus inquiétantes des dictatures et des gouvernements totalitaires.

A contrario, la généralisation dans les pays démocratiques de l'*open data* (l'accès libre aux données du secteur public) devrait permettre aux « journalistes de données », et plus généralement aux citoyens concernés, de contrôler les actions de leurs gouvernants ainsi que celles des grandes entreprises. Les mêmes technologies rendent possibles la surveillance des personnes par l'État et celle de l'État par les citoyens. Les gouvernements de pays tels que les États-Unis, le Royaume-Uni, ou plus récemment la France se sont engagés dans le mouvement de l'ouverture des données. Cela ouvre la voie à de nouveaux services et permet aussi de mieux contrôler les actions des gouvernements et des grandes entreprises. On peut espérer que cela conduise à donner plus de responsabilité au citoyen et à refonder la démocratie.

Les technologies du *Big Data* semblent particulièrement adaptées pour prévoir des crises sanitaires, des problèmes environnementaux, des catastrophes naturelles, et pour y réagir. Elles devraient aider à résoudre les problèmes de santé, de transport, d'écologie, à lutter contre la pauvreté. Dans nombre de ces situations, il faut combiner des analyses de grandes masses de données stockées et des analyses en ligne sur un flux de données obtenues en temps réel. De telles combinaisons se retrouvent ainsi, par exemple, dans le suivi personnalisé de personnes en grande difficulté, de personnes très âgées, d'élèves en échec scolaire, etc. Il faut être capable d'analyser des évolutions sur de longues périodes, de prévenir si possible les problèmes, mais aussi savoir détecter l'urgence et réagir en temps réel à des crises.

Prenons l'exemple de la santé. Les données en rapport avec la santé d'un individu croissent sans cesse. Le cœur en est constitué par des informations telles que les examens médicaux, les diagnostics, les soins et les prises de médicaments. Mais on voit aussi se généraliser les données génomiques – la société californienne de biotechnologie *23andMe* propose ainsi aux individus le séquençage d'une partie importante de leur génome pour 99 dollars (moins de 80 euros). Il faut aussi prendre en compte les données sur la vie quotidienne de l'individu, son alimentation, son activité physique, son exposition à des pollutions particulières, etc. De plus en plus, de telles informations deviennent disponibles *via* des équipements de type téléphone intelligent ou *via* les réseaux sociaux.

Parallèlement, les chercheurs dans le domaine médical font un emploi croissant de l'analyse de données. Par exemple, ils espèrent découvrir des corrélations entre la prise de certaines combinaisons de médicaments et des pathologies particulières. Les patients, surtout, pourraient en profiter. Il s'agit d'abord de personnaliser les soins en adaptant les médicaments à chacun, en contrôlant les quantités administrées. On peut imaginer agir de manière préventive en proposant à chacun une hygiène de vie adaptée à ses risques, en accompagnant chaque personne dans sa nutrition, ses activités physiques, sa vie quotidienne. Mais on peut aussi espérer détecter les crises et aider à les gérer, par exemple en régulant la prise de médicaments.

Un problème : l'accès aux données personnelles

Le fait de disposer de toutes les données relatives à une personne ouvre ainsi considérablement le champ des possibles. Mais cela soulève le problème de l'accès aux données personnelles. Un assureur ou un employeur doit-il avoir accès à tout ou partie de l'information constituée par les données médicales et génomiques d'un client, ses données fiscales, ses achats, sa géolocalisation, ses courriels, ses échanges dans les réseaux sociaux ?

Ce sont les données personnelles du client. À ce titre, elles lui appartiennent, et il devrait être seul à pouvoir décider qui y a accès et comment elles sont utilisées. Mais ce n'est pas aussi simple. Par exemple, pour faire progresser la médecine, on devrait pouvoir faire des analyses sur les données médicales de tous. Il semble raisonnable que les résultats de ces statistiques soient publics. Mais il semble tout aussi clair que les données brutes, par exemple celles d'un hôpital, ne puissent pas elles-mêmes faire partie des données ouvertes, même anonymisées, puisqu'il serait impossible de garantir leur confidentialité. Il reste beaucoup à faire pour concilier ces différents aspects. ■

Hadoop pour gérer les données massives

Hadoop, logiciel libre de la fondation *Apache*, a été conçu en 2004 par l'Américain Doug Cutting. Ce logiciel est adapté à l'analyse de grandes masses de données. Il est fondé sur la technique *MapReduce*, que *Google* a utilisée pour son moteur de recherche.

Dans un calcul *MapReduce*, on commence par découper le problème en de nombreux sous-problèmes indépendants (étape *Map*) que l'on confie à des ordinateurs distincts. Ces machines résolvent les sous-problèmes et envoient leurs résultats à d'autres machines qui ont pour tâche de combiner les résultats (étape *Reduce*). L'objectif est de pouvoir travailler sur d'énormes volumes de données en parallélisant les calculs sur des grappes d'ordinateurs. *MapReduce* ne permet de traiter que des problèmes qui peuvent se décomposer en de multiples tâches parallèles.

Le logiciel de *MapReduce* le plus populaire est *Hadoop*, à la base des centres de données de géants du Web tels que *Amazon* et *Facebook*, et on le retrouve de plus en plus dans des offres de « Cloud computing ». Cependant, malgré ses atouts, *Hadoop* est un logiciel encore jeune et imparfait : il n'utilise pas un langage standardisé, ce qui rend sa gestion complexe, et ne réalise qu'un traitement différé, et non en temps réel, des données.

Les pétaoctets du LHC

L'accélérateur de particules LHC, au CERN, produit chaque seconde des milliards de milliards de collisions entre protons. Chacune engendre de l'ordre de un mégaoctet de données. Comment cette avalanche est-elle traitée ?

Faïrouz Malek

Installé au CERN (le Laboratoire européen de physique des particules), près de Genève, le LHC est l'accélérateur de particules le plus puissant au monde. À partir de novembre 2009, il a livré des données issues de collisions frontales entre protons à une énergie de 7 téraélectronvolts (7×10^{12} électronvolts, ou 7 TeV), puis de 8 TeV, jusqu'à son arrêt en mars 2013. Cet instrument géant et international redémarrera dans deux ans, le temps de le préparer à des collisions encore plus violentes, à environ 13 TeV.

Le LHC (*Large Hadron Collider*, ou Grand collisionneur de hadrons) s'accompagne de quatre gigantesques complexes de détection, nommés ALICE, ATLAS, CMS et LHCb, chacun ayant ses spécificités, qui enregistrent les traces des collisions. Lorsque le collisionneur est en fonctionnement, leurs détecteurs transmettent d'énormes volumes et flux de données mises sous forme numérique. Comme nous allons le voir, le stockage de ces données, leur traitement et leur analyse nécessitent une impressionnante infrastructure informatique, l'une des plus puissantes au monde.

L'enjeu du LHC et des équipements du même type est de tenter de percer quelques-uns des mystères de la matière et de l'origine de l'Univers. La traque du boson de Higgs, une particule attendue par les théoriciens depuis une quarantaine d'années et dont la découverte au LHC a été annoncée en juillet 2012, est l'objectif principal des deux expériences ATLAS et CMS. Voyons d'abord pourquoi cette traque s'apparente à la recherche d'une aiguille dans une botte de foin.

Dans le tunnel du LHC circulent en sens opposés deux faisceaux de protons accélérés à une vitesse proche de celle de la lumière. Chaque seconde, près de 40 millions de paquets regroupant chacun quelque 10^{11} protons entrent en collision. L'énergie de la collision proton-proton est si grande qu'elle peut se matérialiser en toutes sortes de particules, dont des bosons de Higgs.

La théorie indique qu'*a priori*, un seul boson de Higgs est produit en moyenne toutes les 10^{10} collisions. Par ailleurs, le boson de Higgs est instable et se désintègre très vite en deux bosons Z, lesquels se désintègrent à leur tour en leptons détectables (électrons ou muons), ou en deux photons, ou en un boson W^+ et un W^-, etc. Si les détecteurs étaient parfaits, ATLAS ou CMS enregistreraient un boson de Higgs toutes les 10^{13} collisions *via* sa désintégration en deux photons ou en deux Z. En réalité, le taux de détection est très faible. Ainsi, sur les deux

L'AUTEUR

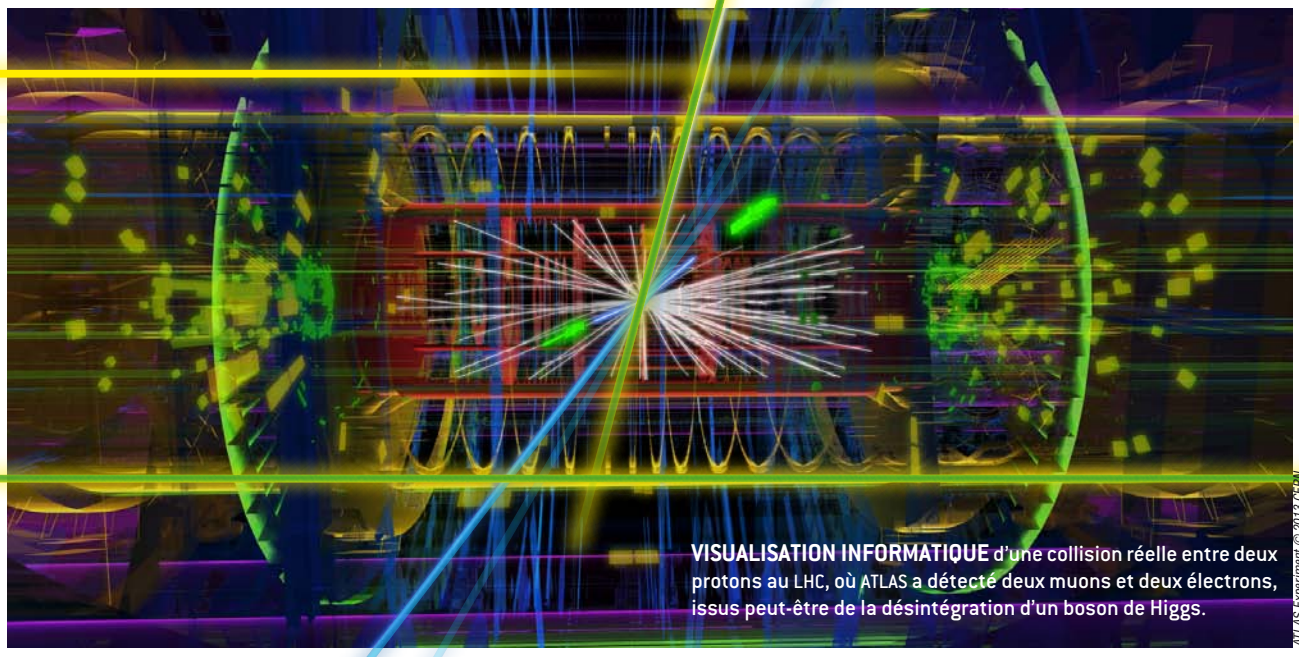
Faïrouz MALEK est physicienne au Laboratoire de physique subatomique et de cosmologie (LPSC) du CNRS et de l'Université Joseph Fourier, à Grenoble. Elle participe à l'expérience ATLAS du LHC et est responsable scientifique du projet LCG-France, la partie française de la grille de calcul du LHC.

BIBLIOGRAPHIE

Dossier « La saga du boson de Higgs », *Pour la Science*, n° 419, pp. 22-35, septembre 2012.

F. Malek, **Le calcul scientifique des expériences LHC - Une grille de production mondiale**, *Reflets de la Physique*, n° 20, pp. 11-15, juillet-août 2010.

© Pour la Science - n° 433 - Novembre 2013



VISUALISATION INFORMATIQUE d'une collision réelle entre deux protons au LHC, où ATLAS a détecté deux muons et deux électrons, issus peut-être de la désintégration d'un boson de Higgs.

ATLAS Experiment © 2013 CERN

ans de prise de données (2011 et 2012) où des millions de particules ont été produites (dont environ 500 000 bosons de Higgs, en théorie), ATLAS n'a détecté que quelques centaines de bosons de Higgs se désintégrant en deux photons (contre quelques milliers si le détecteur avait été parfait).

Les flux de données que transmettent les détecteurs sont considérables. Les collisions engendrent en effet des milliers de particules secondaires, le passage d'une particule dans un détecteur se traduisant par des impulsions électroniques à partir desquelles les physiciens peuvent déterminer sa trajectoire et son identité. Les millions d'impulsions électroniques créées par chaque collision sont alors numérisées par des circuits intégrés aux détecteurs, avant d'être transmises par fibres optiques aux ordinateurs d'acquisition de données. Elles y sont regroupées sous formes de fichiers binaires de « données brutes ».

Chaque collision constitue, au niveau de chacune des quatre expériences, un événement dont la taille en données brutes est de l'ordre du mégaoctet. Un tri intelligent est effectué en moins d'une seconde pour ne garder que les événements potentiellement intéressants. Réalisé grâce aux circuits électroniques en amont puis à des grappes d'ordinateurs et des algorithmes spécialisés, ce tri réduit le taux d'acquisition à environ 200 événements intéressants par seconde. Ces données brutes sont enfin enregistrées à un taux équivalent à deux CD par seconde pour chacune des quatre expériences, sachant que la capacité d'un CD est de 700 mégaoctets.

Un gigaoctet par seconde

Les données brutes sont traitées afin d'extraire les informations « physiques » (les caractéristiques de chaque trajectoire, l'impulsion et l'identité de chaque particule détectée), puis de les analyser. Ces traitements réduisent la taille de l'information d'un facteur deux à dix. Ils sont en général répétés plusieurs fois par an sur l'ensemble des données acquises depuis le démarrage, pour tenir compte du développement continu des algorithmes de traitement et d'étalonnage du détecteur.

Chacun des quatre complexes de détection enregistre en un an 10^{10} collisions, chiffre comparable au nombre d'étoiles de la Voie lactée. Le traitement des données ainsi accumulées est un défi, en raison tant des flux (de l'ordre du gigaoctet par seconde) que des volumes (plusieurs dizaines de pétaoctets chaque année, un pétaoctet étant égal à 1 000 téraoctets, soit 10^{15} octets).

À tout instant, quelques milliers de chercheurs dans le monde sollicitent des ressources de calcul et de stockage afin d'analyser ces données. Pour faire face à ce défi, une « grille de calcul » a été mise en œuvre.

Cette grille met en commun les ressources de calcul et de stockage de nombreuses machines indépendantes, chaque processeur traitant un seul événement. Les événements étant indépendants les uns des autres, peu de données sont à transférer entre les processeurs. Le principal intérêt de ce système est qu'il suffit d'ajouter de nouvelles entités pour augmenter d'autant les capacités de calcul de la grille, laquelle peut traiter des pétaoctets de données. La grille du LHC s'appuie ainsi sur près de 200 centres de calcul répartis dans le monde entier. Ce réseau est hiérarchisé en quatre niveaux, selon la nature et la qualité des services fournis, connectés par des liaisons à haut débit.

Pour traiter l'ensemble des données enregistrées par les expériences du LHC, les besoins en capacité de calcul atteints en 2012 équivalaient à 500 000 ordinateurs de bureau et plus de 200 pétaoctets d'espace de stockage sur disques. Un gigantisme informatique à la mesure du gigantisme (27 kilomètres de diamètre...) du LHC lui-même ! ■

Génomes, protéomes et transcriptomes à foison

Guy Perrière

Quelques grandes banques de données centralisent les résultats du séquençage de l'ADN, des ARN et des protéines des organismes vivants. Mais l'avalanche de séquences fait émerger un autre type d'organisation.

L'étude moderne du monde vivant et de son histoire repose en bonne partie sur l'analyse bio-informatique de données moléculaires provenant des différents organismes. Il s'agit en particulier du matériel génétique – le génome, dont le support est l'ADN –, des ARN transcrits par le génome, ou « transcriptome », et des protéines exprimées – le « protéome ».

La possibilité d'accéder rapidement, par le biais de banques de données, aux séquences de nucléotides (pour les génomes et les transcriptomes) ou d'acides aminés (pour les protéomes) est l'une des raisons des succès qu'a rencontrés la bio-informatique au cours des 30 dernières années. Or l'arrivée à partir de 2005 de nouvelles techniques de séquençage à très haut débit, puis leur démocratisation vers 2010, ont révolutionné ce domaine.

En effet, si les instituts en charge de la maintenance des banques ont pu gérer pendant trois décennies un flux sans cesse croissant de données, ce n'est désormais plus le cas. Confrontés à un déluge sans précédent de séquences, ces organismes doivent désormais s'adapter et, ici comme dans bien d'autres domaines des sciences et de la technologie, l'avenir appartient aux structures en réseau et non plus à des grands systèmes centralisés.

L'histoire des banques de données de séquences prend ses racines en 1965 avec la publication par la physico-chimiste américaine Margaret Dayhoff du premier volume de son *Atlas of Protein Sequence and Structure*. Entre 1965 et 1978, ces livres, qui compilent les séquences de quelques dizaines de protéines, constitueront une source d'inspiration pour toutes les futures collections informatisées. C'est ainsi que *GenBank*, la toute première banque de séquences nucléotidiques d'ambition internationale, a été mise en place en 1979 au Laboratoire américain de Los Alamos. Elle sera rapidement suivie par la création de celle de l'EMBL (le Laboratoire européen de biologie moléculaire) en 1981, puis de la *DNA Data Bank of Japan* (DDBJ) en 1984.

Si la DDBJ est gérée depuis sa création par le *National Institute of Genetics* à Mishima, *GenBank* et la banque de l'EMBL ont connu plusieurs localisations différentes. Aujourd'hui, *GenBank* est maintenue par le NCBI *(National Center for Biotechnology Information)* à Bethesda. La base de l'EMBL l'est par l'EBI *(European Bioinformatics Institute)* à Hinxton, en Grande-Bretagne, qui dispose d'une capacité de stockage d'environ 18 pétaoctets (18×10^{15} octets).

Bien qu'à leurs débuts, le contenu et la taille de ces trois banques étaient relativement différents, une collaboration internationale s'est peu à peu établie et, depuis une vingtaine d'années, leur contenu est virtuellement identique : une séquence soumise à n'importe lequel des trois centres est transmise aux deux autres dans un délai de 24 heures au plus.

Les grandes banques ne sont plus exhaustives

La caractéristique principale de ces banques est qu'elles permettent d'accéder librement à la quasi-totalité des séquences biologiques obtenues par des laboratoires publics ainsi que par un grand nombre de laboratoires privés. C'est cet accès non limité qui a contribué aux avancées importantes obtenues par la bio-informatique au cours de son existence en tant que discipline scientifique. En effet, cette disponibilité immédiate et exhaustive des données a énormément facilité la reproductibilité des résultats, un point essentiel en recherche scientifique.

Pendant près de 30 ans, le volume des données soumises à ces trois collections a crû de façon exponentielle, avec un temps de doublement moyen de l'ordre de 18 mois. La période 2000-2010 a même vu ce temps de doublement diminuer, avec

© isak55/shutterstock.com

le séquençage de nombreux génomes ou transcriptomes complets, dont le premier génome humain en 2001.

Cependant, depuis 2010, un changement de tendance assez surprenant s'est produit : ce temps de doublement s'allonge, alors même que les méthodes de séquençage à très haut débit se sont généralisées et qu'il n'y a jamais eu autant de séquences produites dans les laboratoires. Une première explication à ce phénomène inattendu tient au fait que les centres responsables de la maintenance des banques de séquences sont de moins en moins à même de supporter la charge financière que représentent l'achat continuel de capacités de stockage supplémentaires ainsi que le développement et la maintenance des infrastructures associées.

Un autre problème est que les volumes de données sont désormais tels qu'il n'est plus possible de les transmettre en un temps raisonnable aux centres *via* le réseau. Une solution de plus en plus adoptée, pour qui veut voir ses séquences figurer dans les banques, est d'expédier un disque dur sur lequel sont sauvegardées les séquences en question, le transfert s'effectuant ensuite sur place ! Il s'agit là d'une sorte de retour aux pratiques antérieures à Internet : jusque vers la fin des années 1980, c'est par envoi postal de supports physiques (à l'époque, des bandes magnétiques ou des disquettes) que l'on transférait les séquences.

Enfin, la question de l'accès aux séquences « brutes », non assemblées et non annotées (c'est-à-dire telles qu'elles sortent des machines de séquençage), est également un problème d'importance. Du fait de la quantité de séquences de ce type, les centres ne proposent plus un accès direct aux entrées individuelles, mais à des archives compressées pouvant en contenir un très grand nombre. La survie de ces archives est fréquemment remise en question, ce service ayant déjà été supprimé une fois au NCBI, puis rétabli.

Dans ce contexte, de plus en plus de séquences ne sont tout simplement pas envoyées aux centres de saisie et leur mise à disposition à la communauté se fait par l'intermédiaire de banques de données locales, mises en place dans le cadre de projets limités. Une conséquence de cela est qu'il existe désormais, outre les trois collections généralistes mentionnées, une multitude de banques spécialisées, que celles-ci soient dédiées à un organisme spécifique ou à une problématique biologique particulière. La revue *Nucleic Acids Research* publie chaque année un numéro spécial consacré aux principales banques disponibles dans le monde. Dans son édition de janvier 2013, elle en recensait pas moins de 178. Cependant, ce catalogue est très loin d'être exhaustif et le nombre de banques spécialisées existantes est bien plus élevé.

Structuration en réseau et moyens mutualisés

Du fait qu'une quantité croissante de séquences ne sont plus envoyées aux centres de saisie, les trois collections généralistes ne peuvent plus être considérées comme complètes. Or cette perte d'exhaustivité a d'ores et déjà des répercussions sur la facilité de la reproduction des résultats qui était l'apanage de la bio-informatique. Dans ce contexte, une tendance lourde est à la structuration en réseau et à la mutualisation des moyens.

En France, depuis 2005, la bio-informatique s'est structurée sous la forme d'un réseau national (le ReNaBi) regroupant six centres régionaux, chaque centre fédérant un nombre variable de plates-formes et d'équipes de recherche. Suite à l'obtention en 2011 d'un financement dans le cadre des « Investissements d'avenir » du gouvernement, le ReNaBi est en train d'évoluer vers une nouvelle structure qui prendra le nom d'IFB (Institut français de bio-informatique). L'IFB sera constitué d'une part d'une plate-forme centrale et, d'autre part, des six centres ReNaBi originaux. Ce financement à long terme permettra d'accroître sensiblement les capacités de calcul et de stockage des données de séquençage mises à la disposition de la communauté des chercheurs français.

Au niveau européen, l'EBI a lancé, début 2007, l'initiative ELIXIR. Elle vise à fédérer les grands centres de bio-informatique (nationaux ou régionaux) dans un réseau européen, chaque nœud ayant une spécificité thématique. Tout comme dans le cas français, une telle structuration permettrait effectivement de répartir la charge, aussi bien en quantités de données à gérer qu'en termes d'infrastructures, pour ce qui est du stockage ou de l'archivage des séquences. Dans le cas d'une participation de la France à cette initiative, c'est l'IFB qui deviendrait le nœud national ELIXIR. ■

Quelques chiffres

6 887 génomes complets publiquement accessibles (en septembre 2013).

670 milliards de nucléotides (324 millions de séquences) dans la version de septembre 2013 de la banque de l'EMBL, soit environ 223 fois le génome humain.

54 milliards de nucléotides supplémentaires entre juin et septembre 2013, soit plus que tout ce qui a été introduit dans la banque de l'EMBL entre juin 1982 et septembre 2004.

3 milliards de séquences de 100 nucléotides de long produites au cours d'une seule expérimentation sur une machine *Illumina HiSeq 2500*.

L'AUTEUR

Guy PERRIÈRE est directeur de recherche du CNRS à l'UMR 5558, à Lyon. Il dirige le Pôle Rhône-Alpes de bio-informatique (PRABI) et est représentant français dans le réseau européen EMBnet.

BIBLIOGRAPHIE

M. A. Quail *et al.*, **A tale of three next generation sequencing platforms : comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers**, *BMC Genomics*, vol. 13, article 341, 2012.

Y. Kodama *et al.*, **The Sequence Read Archive : explosive growth of sequencing data**, *Nucleic Acids Res.*, vol. 40, pp. D54-D56, 2012.

G. Cochrane *et al.*, **The future of DNA sequence archiving**, *GigaScience*, vol. 1, article 2, 2012.

SDSS, un relevé... astronomique !

Une carte tridimensionnelle couvrant le tiers de la voûte céleste et recensant plus de 470 millions d'astres : c'est notamment ce qu'a produit SDSS, un projet emblématique des données massives en astronomie.

Céline Reylé

Chaque nuit depuis l'année 2000, le télescope de l'Observatoire d'Apache Point, dans les monts Sacramento, au Nouveau-Mexique (États-Unis), scrute le ciel. Grâce à cet instrument de 2,5 mètres de diamètre, les astronomes établissent un gigantesque relevé du ciel, nommé SDSS *(Sloan Digital Sky Survey)*. Il aura fallu plus d'une décennie d'observations avec ce télescope dédié pour construire l'image du ciel la plus grande et la plus détaillée jamais obtenue *(voir l'illustration)*.

Cette image est unique non seulement par l'étendue couverte, puisqu'elle balaie le tiers du ciel, mais aussi par sa profondeur de champ : elle recense tous les objets célestes situés dans la ligne de visée de la caméra jusqu'à des distances cosmologiques. L'accumulation de plusieurs millions d'images prises avec cette caméra a produit une image d'un térapixel (10^{15} pixels), en cinq couleurs. L'image est tellement grande que 500 000 écrans de télévision à haute définition seraient nécessaires pour la visualiser entièrement à sa résolution maximale. Elle est de surcroît complétée par des analyses spectroscopiques d'un grand nombre des astres détectés.

Les objectifs scientifiques du SDSS sont multiples et le programme comporte, dans sa phase actuelle, quatre champs d'étude bien définis. Le volet SEGUE *(Sloan Extension for Galactic Understanding and Exploration)* étudie les étoiles de la Voie lactée pour comprendre le scénario de sa formation. Les vieilles étoiles de notre galaxie sont observées avec un spectrographe, qui décompose les rayonnements électromagnétiques reçus de l'astre visé. L'étude spectroscopique révèle ainsi sa température, sa composition chimique, sa masse, son âge, sa vitesse.

Une image de 10^{15} pixels

Ces informations individuelles, mises en lien d'une étoile à une autre, permettent de révéler en détail la structure spirale de la Voie lactée, de découvrir de nouvelles classes d'étoiles, d'étudier les déplacements des astres à l'intérieur de la Galaxie, etc. L'ensemble des renseignements ainsi obtenus aide à élaborer des scénarios expliquant la formation et l'évolution de la Galaxie.

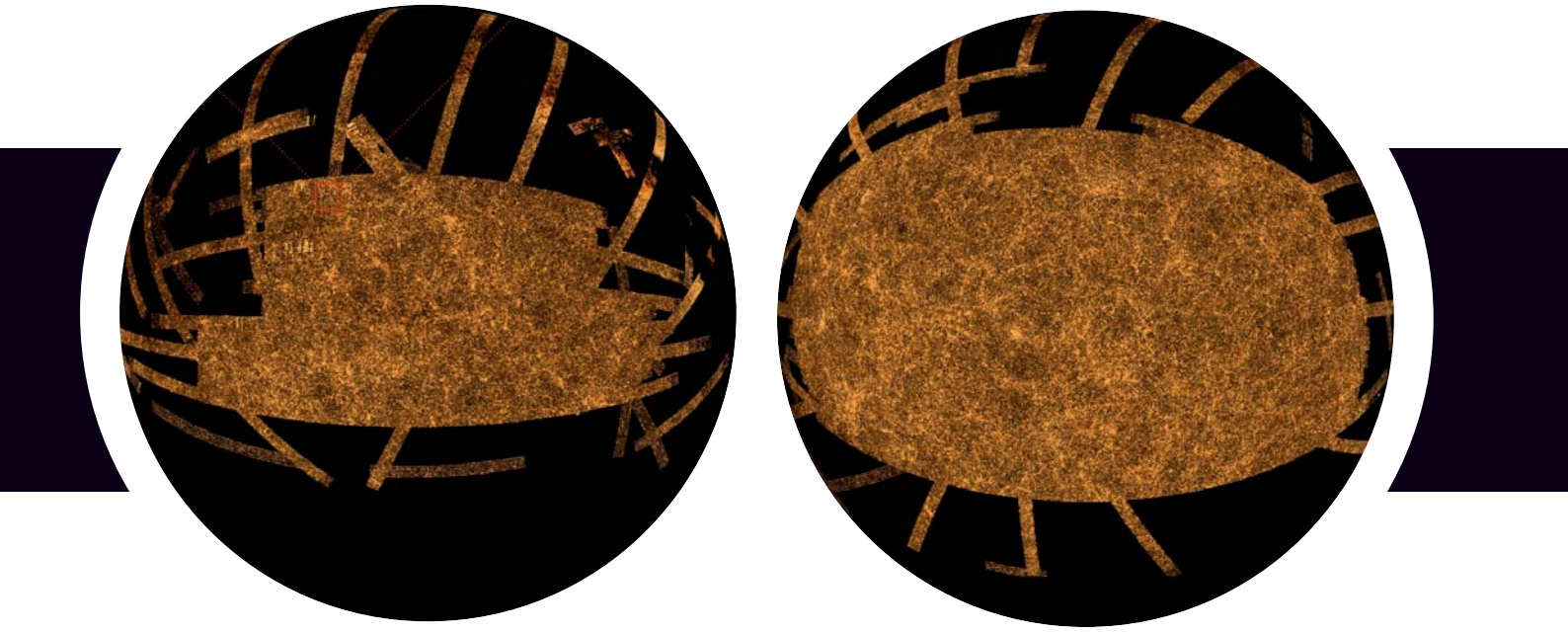
Un deuxième volet de SDSS est le relevé BOSS *(Baryon Oscillation Spectroscopic Survey)*, qui concerne l'Univers lointain. Son objectif est d'éclairer le mystère de l'« énergie sombre », entité de nature inconnue qui serait responsable de l'accélération de l'expansion de l'Univers, et d'aider à améliorer les modèles cosmologiques. Il utilise un spectrographe capable de décomposer la lumière d'un millier d'astres à la fois. BOSS est chargé de mesurer les distances de plus d'un million de galaxies et quasars (noyaux de galaxies actives). Il dresse ainsi la plus grande carte tridimensionnelle des galaxies dans l'Univers. En analysant la répartition spatiale des galaxies lointaines et des quasars, on espère détecter les empreintes laissées par les ondes acoustiques (ondes de pression) qui se sont propagées dans l'Univers peu après le Big Bang.

Ces informations seront très utiles aux scientifiques qui tentent de comprendre l'évolution de l'Univers. La structure à grande échelle de l'Univers, la façon dont les galaxies sont regroupées, sont autant d'indices pouvant jouer en faveur ou en défaveur de telle ou telle théorie cosmologique.

Le relevé MARVELS *(Multi-object APO Radial Velocity Exoplanet Large-area Survey)*, quant à lui, est voué à la détection d'exoplanètes. Pour ce faire, les astronomes de ce projet mesurent très précisément et de manière répétée le mouvement de va-et-vient d'environ 8 500 étoiles de type solaire et proches, mouvement oscillant dû à la présence d'un compagnon du type de Jupiter en orbite autour de l'étoile. On s'attend avec MARVELS à la découverte

M. Blanton et collaboration SDSS-III

© Pour la Science - n° 433 - Novembre 2013



CES DEUX IMAGES DE LA VOÛTE CÉLESTE *(au-dessous du plan de notre galaxie, à gauche ; au-dessus du plan galactique, à droite)* ont été construites avec les données du relevé astronomique SDSS. Chaque point lumineux est une source détectée. Les régions du ciel qui n'ont pas été observées sont en noir.

d'une centaine de nouvelles exoplanètes géantes, ainsi qu'un nombre similaire de naines brunes, astres intermédiaires entre les plus petites étoiles et les planètes gazeuses.

Enfin, le quatrième volet de SDSS, l'expérience APOGEE *(Apache Point Observatory Galactic Evolution Experiment)*, consiste à observer la Voie lactée dans l'infrarouge proche. Ce domaine du spectre électromagnétique permet de pénétrer à travers la poussière qui obscurcit la plupart des étoiles situées dans le disque et le bulbe galactiques.

APOGEE utilise un spectrographe à haute résolution très performant. Les astronomes mesurent grâce à lui la composition chimique et la vitesse (le long de la ligne de visée) de 100 000 étoiles géantes, ce qui fournit un aperçu sans précédent sur la structure dynamique et l'histoire chimique de la Voie lactée.

Ainsi, SDSS constitue un projet astronomique de grande ampleur, qui a déjà apporté de nombreuses découvertes. Cette immense observation du ciel, qui couvre 14 500 degrés carrés sur la voûte céleste, a répertorié 470 millions d'astres. Et parmi eux, près de deux millions de galaxies et 500 000 étoiles sont la cible d'études spectroscopiques.

La quantité d'informations produites par SDSS est énorme. Le volume total des données actuellement accessibles en ligne est de 71 téraoctets, ce qui équivaut à l'information contenue dans une bibliothèque comptant 71 millions de livres de 1 000 pages, soit bien plus qu'il n'y en a à la Bibliothèque nationale de France...

Un travail colossal est nécessaire pour traiter les données brutes provenant des instruments d'observation et en extraire les informations sur les paramètres physiques des objets astronomiques. Bien sûr, une grande partie de ce travail est effectuée par des ordinateurs. Pour BOSS, par exemple, le traitement est réalisé au Laboratoire américain Lawrence Berkeley par une grappe de 50 machines, chacune comportant huit cœurs (huit processeurs). Les capacités de stockage doivent être également à la hauteur du projet : les différents nœuds de stockage des données brutes et traitées totalisent une capacité de 250 téraoctets...

Il reste qu'une centaine de scientifiques doivent travailler des mois entiers pour préparer les données et les rendre publiques. En astronomie, on a pour tradition de rendre les données d'observation accessibles à tous. Celles du SDSS renferment sans aucun doute de nombreuses découvertes pour les années à venir. ■

L'AUTEUR

Céline REYLÉ est astronome à l'Institut UTINAM, à Besançon, et directrice de l'Observatoire des sciences de l'Univers THETA Franche-Comté/Bourgogne. Elle participe au projet SDSS.

SUR LE WEB

Site du projet SDSS : www.sdss.org

Francis Daumas et Marion Massol

Au cœur d'un centre d'archivage

© Shutterstock/sentiicus

Pour stocker le déluge récent de données numériques, notamment scientifiques, de grands centres d'archivage ont été mis en place. Comment fonctionnent-ils ? Exemple avec le CINES, un centre national à Montpellier.

La plupart du temps, quand nous souhaitons conserver un document numérique, nous le copions sur un support externe, que nous stockons dans un placard. Mais sa récupération longtemps après peut se révéler compliquée. D'abord, on note rarement, ou de façon imprécise, la localisation du support et les informations pertinentes sur le contenu. Si l'on parvient à retrouver le support, encore faut-il pouvoir le lire : imaginez une photo de mariage stockée sur une vieille disquette des années 1980 ! Les équipements de lecture de ce support ont parfois disparu ou sont difficiles à trouver après tant d'années. Supposons que vous ayez trouvé le bon équipement : vous devez aussi disposer du logiciel capable de reconnaître le format du fichier et de restituer le contenu sous une forme intelligible. Or les formats évoluent et les logiciels peuvent disparaître ou créer des fichiers qui ne seront pas lisibles par leurs futures versions. Enfin, le fichier initial ne doit pas comporter trop d'erreurs et son support doit avoir été bien conservé.

Cet exemple illustre les difficultés posées par la préservation à long terme des données : l'obsolescence et la détérioration du support, la disparition de l'outil logiciel de lecture, les évolutions du format du fichier et la perte de la localisation et de la signification de son contenu. Pour répondre à ces difficultés, des techniques d'archivage ont été développées. Nous les présenterons à travers l'exemple d'un des principaux sites d'archivage français : celui du CINES (Centre informatique national de l'enseignement supérieur), à Montpellier, un établissement public placé sous la tutelle du ministère de l'Enseignement supérieur et de la Recherche.

Des données variées

Aujourd'hui, l'information numérique est omniprésente et protéiforme. Les données à conserver sur le long terme sont de natures variées : administratives, culturelles, scientifiques (issues d'observations ou de calculs), etc. Les livres des bibliothèques, les thèses, les articles scientifiques, les encyclopédies sont numérisés *(voir le dossier « Le savoir numérisé » dans ce numéro)*. En résultent une multitude de documents à préserver, sous divers formats : textes, sons, images, vidéos…

© Pour la Science - n° 433 - Novembre 2013